Can Highly Proficient L2 Speech Serve as a Reference for Automatic Pronunciation Assessment?

Toru Shirai [1], *Bagus Tris Atmaja* [1], *Sakriani Sakti* [1]

[1] Nara Institute of Science and Technology, Ikoma, Japan

{toru.shirai.sy1,bagus.tris}@naist.ac.jp, ssakti@is.naist.jp

Abstract

Automatic pronunciation assessment (APA) generally uses native speech as a reference. However, native-like pronunciation may not be the only relevant reference for adult L2 learners, whose speech often retains the influence of L1 phonetic categories and language-specific production patterns. We therefore compare native English speech and highly proficient Japanese L2 English speech as reference sources for assessing Japanese learners of English. Phoneme distributions are modeled using Gaussian mixture models, and pronunciation is assessed from phoneme likelihoods under the same experimental settings. The highly proficient Japanese L2 reference condition produced correlation values numerically close to those of the native English reference condition, although the effect of reference-speaker selection varied across phonemes. These results indicate that highly proficient L2 speech is a relevant reference condition to consider in automatic pronunciation assessment.

Index Terms: automatic pronunciation assessment, self-supervised speech models, goodness of pronunciation

1. Introduction

Pronunciation is one of the essential components of second language (L2) speaking proficiency. Computer-Assisted Pronunciation Training (CAPT) aims to improve learners' pronunciation by analyzing their speech with speech processing techniques and providing automatic feedback on pronunciation quality [1, 2]. Automatic Pronunciation Assessment (APA) is a core technology in CAPT that automatically evaluates learners' speech from acoustic signals. APA systems assess various aspects of pronunciation, including phonetic accuracy, fluency, rhythm, and prosody.

Goodness of Pronunciation (GoP) was originally developed as a phone-level measure of the extent to which non-native pronunciation resembles native-speaker pronunciation, using posterior probabilities derived from acoustic models trained on native speech [3]. GoP estimates how likely an observed phoneme segment corresponds to its target phoneme using posterior probabilities derived from acoustic models trained on native English speech. Subsequent studies incorporated deep speech representations into GoP-based features to improve sentence-level pronunciation assessment [4, 5]. More recently, several methods have been proposed to jointly estimate pronunciation quality at multiple linguistic levels, including sentence, word, and phoneme levels [6, 7, 8, 9].

These methods require a large amount of annotated data with pronunciation scores for training. MixGoP has been proposed to assess pronunciation based on the phoneme distribution of typical speech rather than using pronunciation scores as direct supervision [10]. MixGoP models the distribution of each phoneme using a Gaussian mixture model (GMM) over speech representations extracted from typical speech. It then evaluates atypical pronunciation based on the likelihood of each phoneme segment. This approach enables pronunciation assessment without requiring a large amount of scored training data.

MixGoP was evaluated on speechocean762 [11], which consists of L2 English speech produced by native Mandarin speakers. Because the corpus does not include native English speech, high-scoring L2 speech was used to construct the phoneme distributions representing typical speech. MixGoP estimates these distributions directly from the selected speakers, and the resulting scores may depend on the reference speaker group. Reference-speaker selection is particularly important for adult L2 learners, whose speech retains the influence of established L1 phonetic categories and language-specific production patterns [12]. However, it remains unclear how reference-speech selection affects MixGoP and how the best assessment performance obtained from highly proficient L2 speech compares with that obtained from native speech.

Accordingly, the reference speaker group used for pronunciation assessment may need to be selected according to the purpose of the assessment and the intended interpretation of the scores [13]. In this study, we examine how the choice of reference speakers affects MixGoP-based pronunciation assessment at both the utterance and phoneme levels. We use J-AESOP, a Japanese-English bilingual speech corpus [14]. Because J-AESOP contains recordings of native Japanese and native English speakers reading the same sentences, we compare two reference speaker conditions: a native English speech condition and a highly proficient Japanese L2 English speech condition.

The main contributions of this study are threefold. First, we systematically compare two reference speaker conditions within the same MixGoP-based pronunciation assessment framework. Second, we examine how the choice of reference speakers affects pronunciation assessment at both the utterance and phoneme levels. Third, we find that the observed correlations obtained using highly proficient Japanese L2 speech were numerically close to those obtained using native English speech, and no statistically significant differences were detected after correction for multiple comparisons. The effect of the reference speaker group nevertheless varied across phonemes. These findings highlight the choice of reference speakers as an important consideration in automatic pronunciation assessment.

2. Method

2.1. Definition of GoP

Goodness of Pronunciation (GoP) is a phoneme-level pronunciation scoring method [3]. It evaluates how well an acoustic segment matches the target phoneme relative to competing

phonemes. Let p denote the target phoneme, X the aligned acoustic segment, N the number of frames, and $\mathcal{Q}$ the set of competing phonemes. We define the GoP score using the following likelihood-ratio formulation:

$$\mathrm{GoP}(p) = \frac{1}{N} \log \frac{P(X \mid p)}{\max_{q \in \mathcal{Q}} P(X \mid q)}. \tag{1}$$

where $P(X \mid p)$ denotes the likelihood of the acoustic segment X given phoneme p,

2.2. MixGoP

Posterior-based GoP relies on softmax posterior probabilities, which can become overconfident when the input speech differs from the training distribution. In addition, a standard phoneme classifier often represents the acoustic distribution of each phoneme as a single cluster, which may not fully capture multiple acoustic realizations within the same phoneme. To address these issues, we employ MixGoP [10]. For an utterance u consisting of N aligned phone segments, let X_i denote the i-th phone segment and p_i its target phoneme label. MixGoP uses the phoneme likelihood $P_\theta(X_i \mid p_i)$ instead of the posterior probability $P_\theta(p_i \mid X_i)$. First, an acoustic encoder extracts a feature representation from each phoneme segment X_i:

$$\mathbf{h}_i = \mathrm{Enc}(X_i), \tag{2}$$

where $\mathbf{h}_i \in \mathbb{R}^F$ is the F-dimensional representation of the i-th phoneme segment. For each phoneme p, MixGoP fits a Gaussian Mixture Model (GMM) to model the phoneme distribution with multiple subclusters:

$$P_\theta(X_i \mid p_i) = \sum_{c=1}^{C} \pi_{p_i}^c \mathcal{N}\left(\mathbf{h}_i \mid \boldsymbol{\mu}_{p_i}^c, \boldsymbol{\Sigma}_{p_i}^c\right), \tag{3}$$

where C is the number of mixture components, π_p^c is the mixture weight, and $\boldsymbol{\mu}_p^c$ and $\boldsymbol{\Sigma}_p^c$ are the mean vector and covariance matrix of the c-th component for phoneme p, respectively. This formulation allows one phoneme to be represented by several acoustic subclusters, which can capture allophonic variation and other within-phoneme differences. The MixGoP score for the i-th segment is defined as its log-likelihood under the GMM of the target phoneme p_i:

$$\mathrm{MixGoP}_{p_i}(X_i) = \log P_\theta(X_i \mid p_i). \tag{4}$$

The utterance-level MixGoP score is defined as the average of the N phoneme-level scores

$$\mathrm{MixGoP}(u) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{MixGoP}_{p_i}(X_i). \tag{5}$$

2.3. MixGoPAttn

The utterance-level MixGoP score in Eq. 5 treats all phonemes equally. MixGoPAttn [10] extends this formulation by assigning different weights to phoneme-level scores based on the observation that individual phonemes contribute differently to pronunciation assessment.

For each phoneme type p, a learnable scalar parameter w_p is introduced. For an utterance containing the phoneme sequence $\{p_i\}_{i=1}^N$, the attention weight assigned to the i-th phoneme segment is computed as

$$\alpha[p_i] = \frac{\exp(w_{p_i})}{\sum_{j=1}^{N} \exp(w_{p_j})}. \tag{6}$$

$$\mathrm{MixGoPAttn}(u) = \sum_{i=1}^{N} \alpha[p_i] \log P_\theta(X_i \mid p_i). \tag{7}$$

The attention module is trained by maximizing Spearman's rank correlation coefficient between the predicted pronunciation scores and the human segmental ratings:

$$\mathcal{L} = -\mathrm{spearman}\left(\mathrm{MixGoPAttn}(u),\, y\right), \tag{8}$$

where y is the human segmental rating. The sorting operation in Spearman's rank correlation was made differentiable using soft sorting [15]. During training, the MixGoP parameters were fixed, and only the attention module was optimized.

2.4. Sparse-MixGoPAttn

MixGoPAttn assigns phoneme-wise attention weights by applying the softmax function to learnable phoneme-level parameters. However, softmax assigns a non-zero weight to every phone segment in an utterance. As an exploratory extension, we replace softmax with Sparsemax [16], which can assign exactly zero weights to segments whose logits fall below the Sparsemax threshold. For an utterance containing the phoneme sequence $\{p_i\}_{i=1}^N$, we first construct the attention logit vector from the learnable phoneme-level parameters:

$$\mathbf{e} = [w_{p_1}, w_{p_2}, \ldots, w_{p_N}]^\top, \tag{9}$$

where $e_i = w_{p_i}$. The sparse attention weights are then obtained by applying Sparsemax. Following the original formulation [16], Sparsemax computes the attention weights by projecting the logit vector onto the probability simplex $\Delta^{N-1} = \{\boldsymbol{\alpha} \in \mathbb{R}^N \mid \alpha_i \geq 0, \sum_{i=1}^N \alpha_i = 1\}$:

$$\boldsymbol{\alpha} = \mathrm{Sparsemax}(\mathbf{e}) = \underset{\boldsymbol{\alpha}' \in \Delta^{N-1}}{\arg\min} \left\|\boldsymbol{\alpha}' - \mathbf{e}\right\|_2^2, \tag{10}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_N]^\top$. This projection has the closed-form solution

$$\alpha_i = \max\left(e_i - \tau(\mathbf{e}),\, 0\right), \qquad \tau(\mathbf{e}) = \frac{\sum_{j \in S(\mathbf{e})} e_j - 1}{|S(\mathbf{e})|}, \tag{11}$$

where $S(\mathbf{e}) = \{j \mid \alpha_j > 0\}$ denotes the support set of the solution. The support set is determined from the sorted logits according to the standard Sparsemax procedure. The threshold $\tau(\mathbf{e})$ is chosen so that the positive components sum to one, and the $\max(\cdot, 0)$ operation enforces non-negativity. Thus, as with softmax, $\boldsymbol{\alpha}$ forms a probability distribution over the phoneme segments in the utterance but permits exactly zero weights. The utterance-level Sparse-MixGoPAttn score is defined as

$$\mathrm{SparseMixGoPAttn}(u) = \sum_{i=1}^{N} \alpha_i \log P_\theta(X_i \mid p_i). \tag{12}$$

The phoneme-level attention parameters are trained using the same differentiable Spearman correlation objective and optimization setting as MixGoPAttn:

$$\mathcal{L} = -\text{spearman}\left(\text{SparseMixGoPAttn}(u),\, y\right). \quad (13)$$

This formulation allows Sparse-MixGoPAttn to exclude phoneme segments assigned zero attention weights while retaining MixGoP's phoneme-wise likelihood modeling.

3. Experiment

3.1. Dataset

We used the J-AESOP corpus in our experiments. This corpus contains speech from 180 native Japanese speakers (114 female and 66 male) and, as a reference group, 25 native English speakers (16 female and 9 male). The corpus is organized into eight tasks. In this study, we used Task 6.01, in which all speakers read the English version of "The North Wind and the Sun" [17]. The utterances in this task were rated by 16 phonetically trained raters. In the rating procedure, each speaker's read speech was divided into three segments and rated on a 1-to-10 scale for four criteria: *Segmental Accuracy*, *Prosody*, *Fluency*, and *Nativelikeness*.

The utterances were also force-aligned using the Hidden Markov Model Toolkit (HTK) with acoustic models trained on the TIMIT corpus, and the resulting phone alignments were mapped to phoneme labels corresponding to the CMU Pronouncing Dictionary. In addition, seven phonetics experts manually corrected the alignment annotations, such as misaligned phoneme boundaries, unpronounced phonemes, and inserted phonemes.

3.2. Experimental Details

3.2.1. Speech Encoders

Previous work on MixGoP has shown that representations extracted from self-supervised learning (SSL) speech models achieve better performance for L2 English pronunciation assessment than conventional acoustic features such as MFCCs and Mel-spectrograms [10]. In addition, SSL representations have been shown to encode detailed information about phonetic context and within-phoneme variation [18]. Based on these findings, we used representations extracted from a frozen SSL speech model as input acoustic features.

Specifically, we compared HuBERT-Large [19] and WavLM-Large [20] as speech encoders. HuBERT has been reported to outperform wav2vec 2.0 in APA, suggesting that its hidden representations contain pronunciation-relevant information [21]. WavLM extends the masked prediction framework of HuBERT and is designed to learn general-purpose speech representations for a wide range of speech processing tasks. Based on these previous studies, we treat HuBERT as a strong baseline for pronunciation assessment and compare it with WavLM. Both encoders consist of 24 Transformer layers. We extracted hidden representations from each layer and evaluated the layers independently.

3.2.2. Training Data Construction and Evaluation Metrics

In this study, we examined how the choice of reference speakers affects MixGoP-based pronunciation assessment. The estimation of phoneme distributions and the scoring procedure generally followed those described in the previous study [10]. For each phoneme segment, SSL representations were extracted, and a single feature vector was derived. Rather than averaging the representations over the entire segment, we used the hidden representation at the temporal midpoint of each segment as the representative feature.

The phoneme distribution for each reference condition was modeled independently with a Gaussian Mixture Model (GMM). Following the previous study, the number of mixture components and the maximum number of training samples were set to $C = 32$ and $N = 512$, respectively. When more than 512 feature vectors were available for a phoneme, 512 vectors were randomly subsampled; otherwise, all available vectors were used for training. Phoneme categories that were absent or occurred only once in the reading passage were excluded because the number of reference samples was insufficient for GMM estimation. Each GMM was initialized by k-means clustering and estimated with the Expectation-Maximization (EM) algorithm.

Two reference speaker conditions were examined: a native English speech condition and a highly proficient Japanese L2 English speech condition. The native English condition used speech from all 25 native English speakers in the J-AESOP corpus. For the highly proficient Japanese L2 condition, we computed a speaker-level Segmental Accuracy score by averaging the ratings from the 16 raters across the three annotated utterance segments. To match the number of speakers and gender ratio of the native English condition, the top 16 female and top 9 male speakers were selected as references for GoP distribution estimation. Their mean Segmental Accuracy score was 8.32, compared with 4.71 for the remaining Japanese speakers.

The remaining 155 Japanese speakers were used in the evaluation experiments. Since each speaker contributed three annotated utterance segments, the dataset contained 465 utterances. For MixGoPAttn and Sparse-MixGoPAttn, we performed five-fold speaker-independent cross-validation. In each fold, the speakers were divided into training, validation, and test sets without speaker overlap. The attention parameters were optimized on the training set, while the validation set was used for model selection. The held-out test speakers were not used for training or validation of the attention parameters.

Kendall's rank correlation coefficient was used to evaluate the predicted scores against Segmental Accuracy. For each reference condition and method, the layer with the highest Kendall's rank correlation was used to report the best performance. Higher values indicate stronger agreement between the predicted scores and the human ratings. Differences between the two reference conditions were evaluated with 10,000 paired bootstrap samples drawn at the speaker level. The resulting p–values for the six comparisons were adjusted according to the Holm method.

4. Results

4.1. Utterance-Level Performance Analysis

Table 1 presents the Kendall's rank correlation coefficients between the GoP scores and the human segmental ratings under different speech encoders, reference speaker conditions, and GoP aggregation methods.

We first compare the best performance obtained under the two reference speaker conditions. Across the six combinations of encoder and method, the absolute differences in Kendall's τ between the two reference conditions ranged from 0.0013 to 0.0162. Paired speaker-level bootstrap tests showed no statis-

Table 1: *Kendall's rank correlation coefficients at the highest-performing SSL layer for each reference condition and method.*

Model	Reference	Unsupervised	Supervised	
		MixGoP	MixGoPAttn	Sparse-MixGoPAttn
WavLM	English	0.7451	0.7517	**0.7545**
	Japanese	0.7464	0.7473	0.7383
HuBERT	English	0.7411	**0.7441**	0.7335
	Japanese	0.7389	0.7398	0.7349

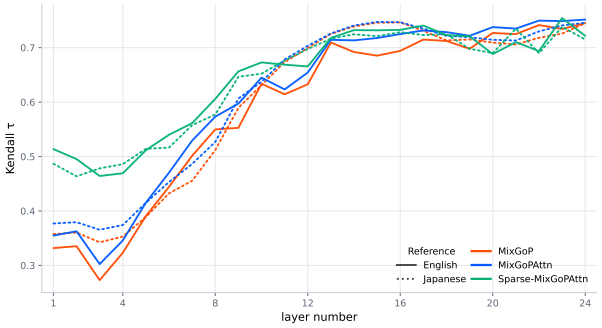


Figure 1: *Layer-wise Kendall's rank correlation coefficients of WavLM.*

tically significant differences after Holm correction. Five of the six 95% confidence intervals included zero. For WavLM with Sparse-MixGoPAttn, the difference was 0.0162 (95% CI: [0.002, 0.031]), but it was not significant after Holm correction ($p = 0.16$). This configuration showed the largest difference between the two reference conditions and was the only one whose 95% confidence interval did not include zero. Overall, the observed correlation values were numerically close, although the present analysis does not establish statistical equivalence between the two reference conditions.

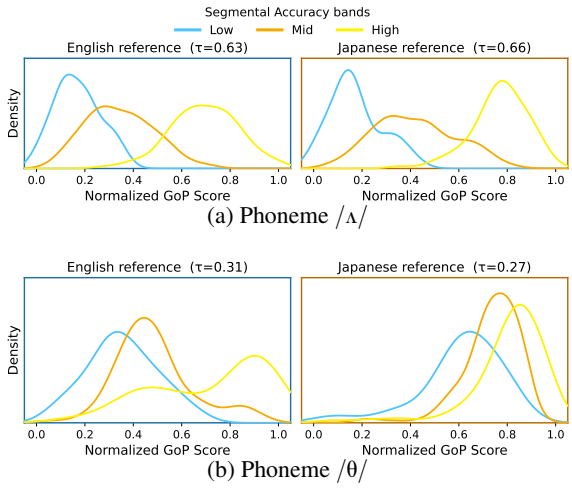
Next, we compare the effects of the speech encoder and GoP aggregation method. WavLM generally produced higher correlation coefficients than HuBERT, with maximum values of 0.7545 and 0.7441, respectively. MixGoPAttn produced slightly higher values than MixGoP under both reference conditions, while Sparse-MixGoPAttn showed less consistent performance across encoders and reference conditions.

Figure 1 shows the layer-wise Kendall's rank correlation coefficients for WavLM. Correlations were lower in the early layers and increased from the middle layers onward under both reference conditions. Sparse-MixGoPAttn showed higher values than the other methods around Layers 1–4, but this advantage diminished in the middle and upper layers. We therefore regard Sparse-MixGoPAttn as an exploratory variant.

The two reference conditions showed similar overall layer-wise trends, although their relative performance varied across individual layers. This suggests that the effect of reference-speaker selection may depend on the representation layer.

4.2. Phoneme-level analysis

In this analysis, we used features extracted from the last layer of WavLM, which showed high performance in the layer-wise analysis, and computed Kendall's rank correlations between the MixGoP score for each phoneme and utterance-level Segmental Accuracy. Thus, this analysis reflects the association of each phoneme score with utterance-level Segmental Accuracy, rather than direct phoneme-level assessment accuracy. Under the native English reference condition, the top five phonemes were /ʌ/, /ð/, /t/, /i/, and /ɹ/, while the bottom five were /aʊ/, /ɑ/, /g/, /h/, and /ʊ/. Under the highly proficient Japanese L2 reference condition, the top five were /ʌ/, /æ/, /ð/, /t/, and /ɔ/, while the bottom five were /h/, /b/, /ʃ/, /aʊ/, and /θ/. /ʌ/, /ð/, and /t/ appeared in the top five under both conditions, whereas /aʊ/ and /h/ appeared in the bottom five. We next compare the score distributions of /ʌ/ and /θ/ to examine the reference-condition effect at the phoneme level.



(a) Phoneme /ʌ/

(b) Phoneme /θ/

Figure 2: *Density distributions of GoP scores across Segmental Accuracy bands for two phonemes.*

Figure 2 compares the min–max-normalized GoP score distributions for /ʌ/, which showed the highest correlation under both reference conditions, and /θ/, which showed the lowest correlation under the Japanese reference condition. The Segmental Accuracy scores were divided into three ranges: Low = [1, 3), Mid = [3, 6), and High = [6, 10]. For /ʌ/, the GoP score distributions of the three bands were relatively well separated under both the native English and highly proficient Japanese L2 reference conditions. In contrast, the GoP score distributions for /θ/ were less clearly separated. Under the native English reference condition, the three bands were still partially separated, but under the highly proficient Japanese L2 reference condition, they largely overlapped.

This difference for /θ/ may reflect which realizations of the phoneme are present in the reference speech. Under the native English reference condition, the score reflects how closely a learner's /θ/ production matches the distribution of native En-

glish speakers, and proficiency differences appear to be captured relatively well. In contrast, under the Japanese reference condition, proficiency differences in /θ/ were not clearly observed. The acoustic realization of English /θ/ produced by Japanese speakers has been shown to vary with proficiency [14]. Therefore, one possible interpretation is that such proficiency-related acoustic differences were not reflected in the same way in the SSL representation space used in this study.

5. Conclusion

This paper investigated the use of highly proficient L2 speech as reference speech for automatic pronunciation assessment. By comparing the best assessment performance obtained with native English and highly proficient Japanese L2 English reference speech, we found that the observed correlations were numerically close, with no statistically significant differences after Holm correction, while exhibiting phoneme-dependent differences in pronunciation likelihood estimation. These results show that reference-speaker selection had a limited effect on the observed utterance-level correlations in most of the examined configurations, although its effect varied across phonemes and methods. Since this study was limited to highly proficient L2 reference speakers and Japanese learners of English reading the same passage, further studies involving different reference proficiency levels, learners from other first-language backgrounds, and spontaneous speech are needed.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 24K02967 and JSPS Program for Forming Japan's Peak Research Universities (J-PEAKS) Grant Number JPJS00420240021. Part of this work was also supported by JSPS KAKENHI Grant Numbers JP23K21681 and JP25H01139.

7. Generative AI Use Disclosure

Generative AI was used solely to assist in the editing and polishing of this manuscript and was limited to minor grammatical corrections and stylistic improvements.

8. References

[1] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer assisted pronunciation training," *Computer Assisted Language Learning*, vol. 15, no. 5, pp. 441–467, 2003.

[2] D. M. Chun, "Signal analysis software for teaching discourse intonation," *Language Learning & Technology*, vol. 2, no. 1, pp. 74–93, 1998.

[3] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2–3, pp. 95–108, 2000.

[4] B. Lin and L. Wang, "Deep feature transfer learning for automatic pronunciation assessment," in *INTERSPEECH*, 2021, pp. 4438–4442.

[5] ——, "Gated fusion of handcrafted and deep features for robust automatic pronunciation assessment," in *APSIPA ASC*, 2022, pp. 1396–1401.

[6] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," in *ICASSP*, 2022, pp. 7262–7266.

[7] J. H. M. Wong, H. Zhang, and N. F. Chen, "Variations of multi-task learning for spoken language assessment," in *INTERSPEECH*, 2022, pp. 4456–4460.

[8] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "3M: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment," in *APSIPA ASC*, 2022, pp. 575–582.

[9] H. Do, Y. Kim, and G. G. Lee, "Hierarchical pronunciation assessment with multi-aspect attention," in *ICASSP*, 2023, pp. 1–5.

[10] K. Choi, E. Yeo, K. Chang, S. Watanabe, and D. R. Mortensen, "Leveraging allophony in self-supervised speech models for atypical pronunciation assessment," in *NAACL*, 2025, pp. 2613–2628.

[11] 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," in *INTERSPEECH*, 2021, pp. 3710–3714.

[12] J. E. Flege, "Second language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, 1995, pp. 233–277.

[13] A. V. Moere and M. Suzuki, "Using speech processing technology in assessing pronunciation," in *Assessment in Second Language Pronunciation*. Routledge, 2017, pp. 137–152.

[14] K. Yazawa, T. Konishi, and M. Kondo, "The J-AESOP corpus: Design, application, and future directions of a Japanese-English bilingual speech corpus," *Acoustical Science and Technology*, vol. 46, no. 5, pp. 553–563, 2025.

[15] M. Blondel, O. Teboul, Q. Berthet, and J. Djolonga, "Fast differentiable sorting and ranking," in *ICML*, 2020, pp. 950–959.

[16] A. F. T. Martins and R. F. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *ICML*, 2016, pp. 1614–1623.

[17] I. P. Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press, 1999.

[18] H. Ji, T. Patel, and O. Scharenborg, "Predicting within and across language phoneme recognition performance of self-supervised learning speech pre-trained models," *arXiv preprint*, vol. arXiv:2206.12489, 2022.

[19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[20] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[21] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," in *INTERSPEECH*, 2022, pp. 1411–1415.