Measuring Phoneme-Level Pronunciation Deviations in Japanese Learners of English Using Self-Supervised Speech Representations

Atharv Kulkarni : Lead Researcher | System Architect
India International School in Japan
Tokyo, Japan
{s2013236}@iisjapan.com

Abstract—**Research suggests a part of the reason for the limited English Proficiency, or bilingualism, in Japan, is tied to a societal 'fear' of mispronouncing certain words in English. Accurate pronunciation feedback is therefore essential for Japanese learners of English, but traditional Computer-Assisted Language Learning (CALL) software typically provides general feedback, lacking the necessary regional nuance required for greater effectiveness. This paper provides insight into the phoneme-level deviations of Japanese speakers of English, compared to a baseline of American English speakers, conducted on the UME-ERJ corpus by the NII-SRC (Japan's National Institute of Informatics - Speech Resources Consortium) using a self-supervised speech representation (SSSR) model, which, to the best of our knowledge, represents the first attempt to quantify pronunciation deviations in Japanese-English using this state-of-the-art method.**

Index Terms—**self-supervised learning, HuBERT, japanese-english, phoneme analysis.**

I. Introduction

Despite its global economic standing, Japan has consistently been ranked as 'low' or 'very low' in the context of English proficiency, holding the 96th position on the EF EPI index in 2025 [1]. Prior research suggests that this is not due to exposure, as English is a standard subject in the Japanese education system, but, in fact, due to a cultural 'shyness' or a 'lack of willingness to speak English', predominantly because of the fear of mispronunciation [2], [3]. This psychological barrier negatively impacts overall proficiency, where learners hesitate to speak English due to the fear of phonetic mistakes, or appearing non-proficient at the language. Computer-Assisted Language Learning systems attempt to bridge this, and can be further improved by reliable region-specific data.

Most modern CALL systems rely on supervised speech-to-text (STT) or Mel-Frequency Cepstral Coefficients (MFCC) to judge pronunciation against an internal metric [4]. One notable downside of this method is not being able to classify deep phoneme-level nuances, notably the distinction between liquid phonemes (/l/ and /r/) and fricatives (/θ/ and /ð/) [3], [5], [6], which are known to be difficult for Japanese native (L1) speakers. Supervised models need massive amounts of labeled non-native speech data, which is scarce and expensive to produce on such scales [7], [8]. Thus, there is a need for a self-supervised metric that can quantify 'phonetic distance' without manual transcription.

In this paper, we address the previously mentioned challenges by using Self-Supervised Learning (SSL) representations [9], [10]. By using layer 9 of HuBERT, we can extract high precision latent feature encodings which depict the underlying structure of speech. The dataset utilized is the UME-ERJ corpus, a robust set containing a total of 69,888 (35,277 Female | 34,611 Male) usable non-native speaker data files, spread across 102 Female and 100 male volunteers. A collection of 17,055 (10,126 Female | 6,929 Male) native samples were used to establish a baseline. Transcripts were obtained from the provided documentation, and were sliced into individual text files to ensure compatibility with Montreal-Forced-Aligner (MFA) for phoneme-level classification, which were then used for sorting individual phonemes post HuBERT encoding.

II. Literature Review

This section will summarize previous papers with similar methodologies and goals:
(Brief)
The evaluation of pronunciation has shifted from traditional phonetic analysis to the use of high-dimensional self-supervised speech representations (SSSR). [1], [2] For Japanese-English (JE) learners, quantifying deviations requires tools that can handle both the mapping complexity of non-native speech and the phonetic nuances of L2-specific errors. [2], [3] This section reviews recent research utilizing HuBERT, the Montreal Forced Aligner (MFA), and the UME-ERJ corpus for phonetic distance quantification. (The author used M. McAuliffe's [4] work on (MFA) to gain functional knowledge of the different ARPAbet phonemes used by (MFA))

A. Distance-Based Feedback for JE Learners

Kawamura et al. [12] developed the *DDSupport* system to visualize phonetic distance using self-supervised learning on the UME-ERJ corpus.

- **Methods:** The system calculates pronunciation scores based on the distance between learner and native centroids in HuBERT latent space, using attention layers for visualization. [12]
- **Findings:** The use of SSL effectively reduces the reliance on large, annotated non-native datasets. [12]

B. Probing Phonetic Content in SSL Models

Martin et al. [5] examined the internal representational distinctions of self-supervised models regarding phonetic and phonemic contrasts.

- **Methods:** Using HuBERT Base, the researchers extracted frame-level representations and applied multinomial regression probes to decode phonetic information (e.g., stop-voicing contrasts). [4] Aggregation was performed by averaging frames over a phone's duration. [4]
- **Findings:** Middle layers, specifically Layer 9, were foundto represent a "Goldilocks zone" where both low-level acoustic cues and abstract phonological distinctions are most robustly encoded. [1], [5]

C. Reliability of Forced Alignment on L2 Speech

Williams et al. [7] investigated the performance of the Montreal Forced Aligner (MFA) across diverse accented varieties of English.

- **Methods:** The study utilized the MFA General American English acoustic model to align accented speech from 45 speakers. [7] Accuracy was measured via Onset Boundary Displacement and Overlap Rate. [7]
- **Findings:** While 80–93% of boundaries were within 20ms of human standards, accuracy was found to be comparable to conversational L1 English. [7]

D. Allophonic Modeling in Atypical Speech

Choi et al. [13] introduced *MixGoP*, a method for improving pronunciation assessment by modeling phonemic variations through Gaussian Mixture Models (GMMs).

- **Methods:** The approach integrates GMM subclusters with frozen SSSR features (WavLM and HuBERT) to capture how a phoneme is realized differently based on its environment. [13]
- **Findings:** By calculating log-likelihoods rather than categorical assignments, MixGoP achieved SOTA results on non-native and dysarthric datasets (e.g., L2-ARCTIC). [13]

E. Sub-Phonetic Representation of Hidden Units

Wells et al. [16] provided a granular analysis of how discrete HuBERT units map to broad phonetic classes.

- **Methods:** The study focused on k-means clustering of hidden units to observe sub-phonetic events. [16]
- **Findings:** Results indicated that fine temporal dynamics, such as the distinct closure and release portions of plosives, are represented by specific sequences of discrete units. [16]

F. Articulatory Ground Truth for Liquid Deviation

Nagamine [14] analyzed Japanese learners' production of English liquids using a combination of acoustic and articulatory tools.

- **Methods:** The study utilized MFA for segmentation and ultrasound imaging for tongue tracking, applying Principal Component Analysis (PCA) to reduce tongue spline data. [14]
- **Findings:** PCA identified tongue retraction (PC1) as the primary differentiator, showing that Japanese learners exhibit distinct magnitude and timing patterns for /r/ and /l/. [14]

III. Method

A. Classifying the data

The UME-ERJ dataset [11] is divided into three subsets, hence referred to as UME-ERJ 1, UME-ERJ 2, and UME-ERJ 3. UME-ERJ 1, UME-ERJ 2 consisted of speakers from 19 Japanese universities speaking from a set of English words which were both general sentences and words/sentences predicted to be 'hard' for L1 Japanese speakers. From each university, each speaker was identified with an ID number in the form M/F-XY, where XY are unique to the person. The content of the audio uttered by the speaker was from a common set of sentences labeled S(0-8)-(001-125) and words W(0-5)-(001-228).

UME-ERJ 3 consisted of 20 American speakers of English, 12 Female speakers and 8 Male speakers. The content was different from the sentences/words uttered by the Japanese volunteers.

We also created a unified, cleaned version of the provided transcripts for sentences uttered by both American English (AE) and Japanese English (JE) speakers. Using a Python script, we matched the `speaker_id` in the transcript files to the corresponding audio filename and inserted the transcript content into a new text file. This preprocessing step ensured compatibility between the `.wav` and `.txt` files required by the Montreal Forced Aligner (MFA). Finally, all files were systematically renamed to standardize processing, using the format `S(A)_(PQR)_(SQT)_F|M(XY).extension`, where A is sentence group number, PQR is sentence track, STQ is university ID, and XY is speaker ID. The extensions were either `.wav` or `.txt`.

B. Identifying Phonemes and Their Boundaries

Montreal Forced Alignment (MFA) was used to identify phonemes present in the `.wav` files and their region boundaries [5]. Taking the `.wav` file and a `.txt` file as inputs, it ran MFCC localization to output `.textgrid` files with multiple layers of phonemic information [16].

C. Using the Self-Supervised Speech Representation Model

Layer 9 of HuBERT [10] was utilized due to being ranked as having the best classification for phoneme-level analysis [5], [10], [16]. Using the raw encodings from HuBERT, a Python script utilized the previously created `.textgrid` files to locate specific phonemes, and using the mapped boundary times, sliced the HuBERT encodings from start to end in 20ms intervals. The mean of the collected zone was then utilized to be part of the final global μ_0, the token-weighted phoneme centroid, calculated by again averaging the parts, to arrive at a final phoneme value in the form of a 768-dimensional latent vector [10]. This process was repeated 6 times, for Jp male/female, Am male/female, Jp male+female, Am male+female (Jp = Japanese/Japanese English, Am = American/American English)

The data is stored in a pickle `.pkl` binary file.

For further analysis, the following terms were also calculated to ensure data quality: $\mu_{\text{speaker-norm}}$: the speaker-balanced centroid, σ_{global} and $\sigma_{\text{inter-speaker}}$.

D. Data Analysis

Data analysis was also conducted using the Python data analysis and graphing libraries of Seaborn, Matplotlib, SciPy and Numpy. The process involved depickling of the `.pkl` files and comparing the vector datapoints using cosine distance [19] between JE and the AE baseline across all metrics.

The cosine distance between two vectors **a** and **b** is defined as:

$$d_{\text{cos}}(\mathbf{a}, \mathbf{b}) = 1 - \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$$

The Euclidean distance is given by:

$$d_{\text{eucl}}(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

IV. Experimental Details

A. Experimental Setup

The primary computation device used was a personal computer, having an Intel i-9 13900HX processor, RTX 4070 mobile GPU for CUDA acceleration, and 32GB DDR5 5600MT/s RAM for `numpy` headroom.

Python was the primary programming language used, with the standard machine learning libraries being utilized.

MFA was installed in a separate Anaconda virtual environment to avoid conflicts with the main program. MFA computation took 730.143 seconds for 35,277 JE Female tracks, and 643.312 seconds for JE Male tracks.

HuBERT encoding took roughly 1 minute per 1,000 files with the described setup (16.67 items per second [10])

V. Results

The obtained data was analyzed across four major categories:

1) **Japanese Female vs. American Female:** To determine the average phonemic deviations between female speakers across both cohorts.
2) **Japanese Male vs. American Male:** To determine the average phonemic deviations between male speakers across both cohorts.
3) **Sex-Based Deviation Comparison:** A comparative analysis of [Japanese Male | American Male] deviations against [Japanese Female | American Female] deviations to observe how sex may correlate with phonetic deviations
4) **Aggregate Composite Analysis:** A comparison of the total baseline composite (Male + Female) against the total Japanese English composite (Male + Female) to quantify the global phonetic distance.

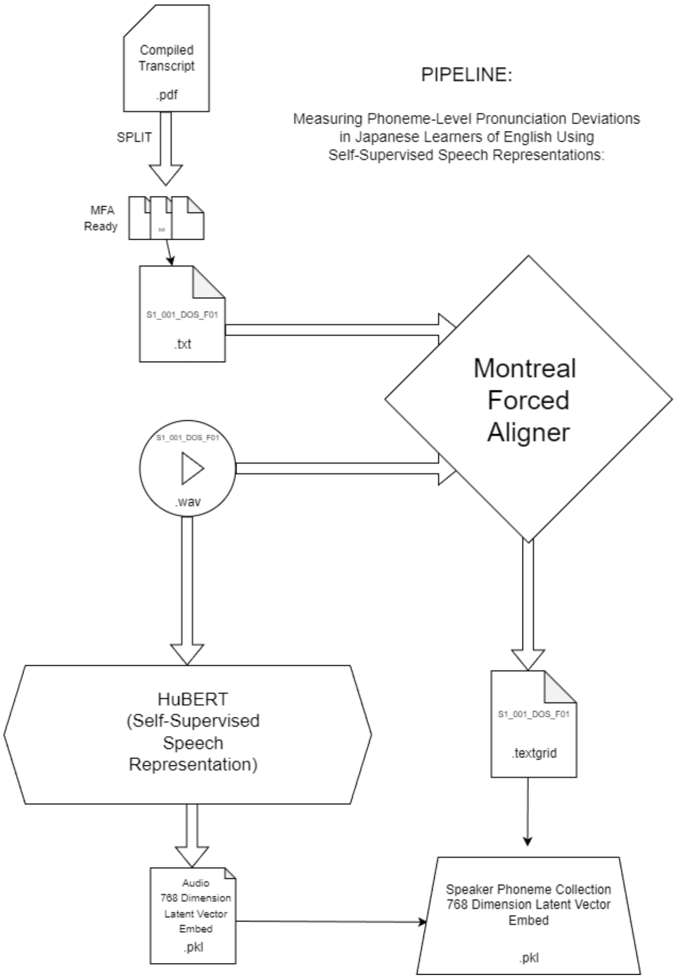
Compiled Transcript
.pdf
SPLIT
PIPELINE:
Measuring Phoneme-Level Pronunciation Deviations in Japanese Learners of English Using Self-Supervised Speech Representations:
MFA Ready
S1_001_DOS_F01
.txt
Montreal Forced Aligner
S1_001_DOS_F01
.wav
HuBERT (Self-Supervised Speech Representation)
S1_001_DOS_F01
.textgrid
Audio 768 Dimension Latent Vector Embed
.pkl
Speaker Phoneme Collection 768 Dimension Latent Vector Embed
.pkl

Fig. 1. System Architecture of the Proposed Evaluation Pipeline. The process integrates a bifurcated workflow: (1) temporal phoneme boundary identification via the Montreal Forced Aligner (MFA) using compiled transcripts and raw audio, and (2) high-dimensional feature extraction through the HuBERT self-supervised model. The final stage performs a temporal-to-latent mapping to generate speaker-specific phoneme collections in a 768-dimensional vector space.

A. Japanese Female vs. American Female

In the first comparison, there are a few notable phonemes that are instantly discernible due to high phonetic drifts compared to the baseline. The compiled data can be seen below:

B. Japanese Male vs. American Male

We will perform the analysis similar to the one in section 1, with the results being classified below:

From the preliminary analysis of the graph and the table, it can be seen that while the males report an otherwise lower deviation from the American baseline, the deviation in the ɾ phoneme is significantly greater than that from the female cohort. The g^{w} deviation is also seen to be greater than that of the female cohort. The deviation from m is also not as significant.

TABLE I
PHONETIC DRIFT COMPARISON: JAPANESE FEMALE VS. AMERICAN FEMALE BASELINE

Phoneme	**Drift**	**Phoneme**	**Drift**	**Phoneme**	**Drift**
ɾ	6.8086	ow	2.0469	ʃ	1.6758
m	5.1133	t^{h}	2.0332	ŋ	1.6738
g^{w}	4.8633	ɒː	2.0078	ɛ	1.6689
ɫ	3.2676	k^{w}	2.0020	f	1.6387
ɟ^{w}	3.1191	v	1.9932	iː	1.6387
ɝ	3.0586	ɐ	1.9316	ɔj	1.6279
ð	2.9727	æ	1.9277	p^{h}	1.6182
ɾ^{j}	2.8867	ɪ	1.9229	d	1.5928
l	2.7500	d^{j}	1.8936	z	1.5732
ʎ	2.7422	aw	1.8701	i	1.5703
ɚ	2.5625	w	1.8633	h	1.5352
t	2.5410	ʉː	1.8262	k^{h}	1.5166
v^{j}	2.5117	dʒ	1.8262	ej	1.5166
t^{w}	2.4883	ə	1.8252	p^{j}	1.4893
ʒ	2.4844	tʃ	1.8154	c^{h}	1.4561
n	2.4512	ʊ	1.8135	b^{j}	1.3496
θ	2.4277	c^{w}	1.8008	ɲ	1.3320
t^{j}	2.3457	c	1.7920	p	1.3125
ɹ	2.0703	f^{j}	1.7910	k	1.2764
ɑː	2.0488	ʉ	1.7822	j	1.2656
Summary Statistics					
Mean Global Drift					2.0625

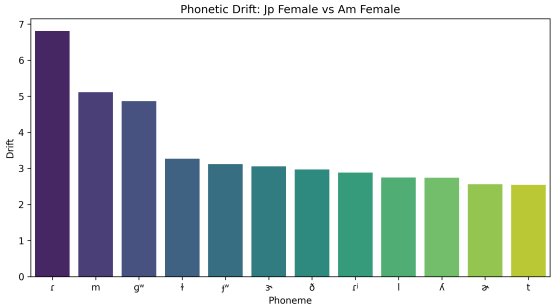

Fig. 2. Comparison of Phonetic Drift for Female Cohorts

TABLE II
PHONETIC DRIFT COMPARISON: JAPANESE MALE VS. AMERICAN MALE BASELINE

Phoneme	**Drift**	**Phoneme**	**Drift**	**Phoneme**	**Drift**
ɾ	8.9453	v	2.1973	ʉ	1.7783
g^{w}	5.3320	k^{w}	2.1797	i	1.7617
ɟ^{w}	3.9238	t^{h}	2.1563	ɒ	1.7305
m	3.8613	ɪ	2.0957	ɛ	1.7070
l	3.6367	ɐ	2.0508	ʃ	1.6514
ɫ	3.5059	æ	2.0352	p^{j}	1.6396
ʎ	3.4961	f^{j}	2.0195	k^{h}	1.6191
ɝ	3.4238	d^{j}	2.0098	ɔj	1.6182
ð	3.3066	ə	2.0098	h	1.6113
ɾ^{j}	3.2617	z	2.0039	p^{h}	1.6035
ɚ	3.0176	c	2.0000	iː	1.5947
v^{j}	2.8086	ow	1.9502	s	1.5020
t^{w}	2.8027	ɒː	1.9316	ej	1.5020
t	2.6309	w	1.9229	ɲ	1.4893
θ	2.6133	tʃ	1.8877	c^{h}	1.4297
n	2.5879	f	1.8545	k	1.3955
t^{j}	2.5859	ʊ	1.8359	b^{j}	1.3926
ʒ	2.4004	ʉː	1.8281	p	1.3584
ɑː	2.3086	ŋ	1.8203	g	1.3516
ɹ	2.2754	dʒ	1.8037	j	1.3467
c^{w}	2.2734	ɑ	1.7969	aj	1.3057
d	1.7920	aw	1.7900	ɟ	1.2959
Summary Statistics					
Mean Global Drift					2.2363

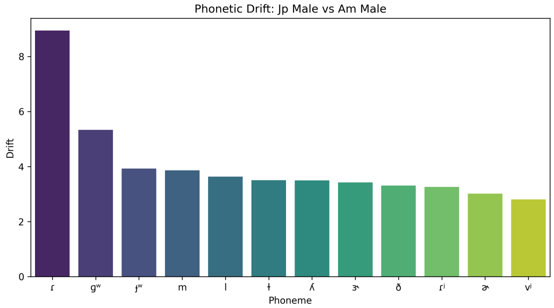

Fig. 3. Comparison of Phonetic Drift for Male Cohorts

C. [Japanese Male | American Male] vs [Japanese Female | American Female]

This general comparison can help identify which sex has more issues with pronunciation in general. As with the earlier comparison, it can be observed that while males in general have less deviation in most phonemes, the significantly greater variation in the few phonemes which show a higher rate of deviation skews the average drift of the males higher than that of the females.

TABLE III
SEX-BASED PHONETIC DEVIATION COMPARISON

Metric	**Value**
Japanese Male Average Drift	2.2363
Japanese Female Average Drift	2.0625
Difference (Male - Female)	**0.1738**

D. [Japanese Composite (Male + Female)] vs [American Composite (Male + Female)]

An aggregate analysis can help in identifying the common deviations for all speakers of L1 Japanese, providing a general overview of pronunciation deviations, and the most likely mispronunciations [5], [16].

TABLE IV
PHONETIC DRIFT COMPARISON: GLOBAL AGGREGATE COMPOSITE

Phoneme	Drift	Phoneme	Drift	Phoneme	Drift
ɾ	6.1016	ow	1.9785	ɛ	1.6406
g^{w}	4.9375	c^{w}	1.9746	ʃ	1.6172
m	4.2461	ɪ	1.9717	i	1.6113
ɫ	3.3535	ɐ	1.9541	p^{h}	1.5830
ɟ^{w}	3.3320	æ	1.9531	ɔj	1.5684
ɝ	3.1914	ɒː	1.9297	iː	1.5674
l	3.1445	d^{j}	1.9150	h	1.5371
ð	3.1309	ə	1.8887	k^{h}	1.5293
ʎ	3.0820	f^{j}	1.8633	p^{j}	1.5254
ɾ^{j}	2.9531	w	1.8623	ej	1.4717
ɚ	2.7422	c	1.8584	c^{h}	1.4082
v^{j}	2.6309	aw	1.8154	ɲ	1.3594
t^{w}	2.5801	ʊ	1.8096	b^{j}	1.3359
t	2.5234	tʃ	1.7822	p	1.2881
θ	2.5020	ʉː	1.7773	k	1.2822
t^{j}	2.4160	dʒ	1.7725	j	1.2725
n	2.3711	ʉ	1.7393	ɡ	1.2559
ʒ	2.3633	f	1.7207	ɟ	1.2227
ɹ	2.1406	ɑ	1.7051	s	1.2090
ɑː	2.0957	z	1.7021	b	1.2021
t^{h}	2.0625	ɒ	1.6865	aj	1.1875
v	2.0625	ŋ	1.6621	m^{j}	1.1641
k^{w}	2.0352	d	1.6562		
Summary Statistics					
Mean Global Drift					2.0723

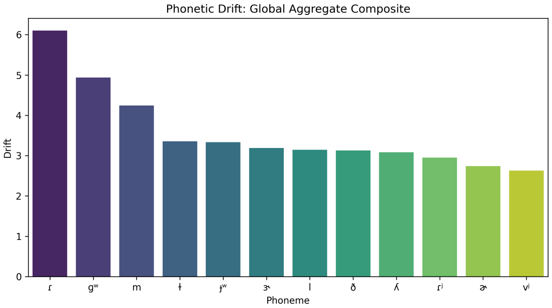

Fig. 4. Global Aggregate Comparison of Phonetic Drift

E. Analysis of Inferred Data

Analyzing the aggregate, we can classify the phonemes into three subcategories based on their calculated drift:

1) **High Deviation (Drift $\geq$ 3.0)**
   - ɾ (6.1016), g^{w} (4.9375), m (4.2461), ɫ (3.3535), ɟ^{w} (3.3320), ɝ (3.1914), l (3.1445), ð (3.1309), and ʎ (3.0820).
2) **Medium Deviation (1.7 $\leq$ Drift < 3.0)**
   - ɾ^{j} (2.9531), ɚ (2.7422), v^{j} (2.6309), t^{w} (2.5801), t (2.5234), θ (2.5020), t^{j} (2.4160), n (2.3711), ʒ (2.3633), ɹ (2.1406), ɑː (2.0957), t^{h} (2.0625), v (2.0625), k^{w} (2.0352), ow (1.9785), c^{w} (1.9746), ɪ (1.9717), ɐ (1.9541), æ (1.9531), ɒː (1.9297), d^{j} (1.9150), ə (1.8887), f^{j} (1.8633), w (1.8623), c (1.8584), aw (1.8154), ʊ (1.8096), tʃ (1.7822), ʉː (1.7773), dʒ (1.7725), ʉ (1.7393), f (1.7207), ɑ (1.7051), and z (1.7021).
3) **Low Deviation (Drift < 1.7)**
   - ɒ (1.6865), ŋ (1.6621), d (1.6563), ɛ (1.6406), ʃ (1.6172), i (1.6113), p^{h} (1.5830), ɔj (1.5684), iː (1.5674), h (1.5371), k^{h} (1.5293), p^{j} (1.5254), ej (1.4717), c^{h} (1.4082), ɲ (1.3594), b^{j} (1.3359), p (1.2881), k (1.2822), j (1.2725), ɡ (1.2559), ɟ (1.2227), s (1.2090), b (1.2021), aj (1.1875), and m^{j} (1.1641).

These findings are largely consistent with traditional phonetic studies [3], [4], [6], [10], [11], [14], [16]

A brief discussion on certain outlying phonemes:

1) **The High Deviation of /ɾ/:** The alveolar flap/tap /ɾ/ shows a high deviation, across both sexes and the composite. With the deviation of the composite being 6.1, the male cohort exhibiting 8.9, and the female cohort displaying 6.8, compared to the mean of 2.07. This may be attributed towards a difference in the use of the alveolar flap, with L1 Japanese speakers possibly substituting the native use of the flap, such as the sentence S8_036 ("Ralph prepared red snapper with fresh lemon sauce for dinner"), wherein the word 'prepared' was found to have a Japanese pronunciation of /puɾepeaðo/, when the standard pronunciation is /pɹIp^{h}ɛəɹd/. This contextual difference could contribute towards the high deviation score
2) **The Difference between Palatized and Unpalatized consonants:** Analyzing the specific deviations of /m^{j}/ and /m/, it is apparent that while /m/ has been placed into the 'high deviation' category, /m^{j}/ has the lowest observed deviation. As noted in Timothy J. Vance's 'The Sounds of Japanese' [21], non-palatized consonants are often substituted by palatized consonants which maintain a high degree of similarity with the native pronunciation.
3) **The Labialized Consonants:** Another observation is that Labialized consonants, such as /g^{w}/ show higher than the mean deviation. This is largely due to the adapted linguistic habits from Japanese, particularly prioritizing one consonant at a time, and an under-articulation of the lips (lip height/distance)

F. Known Limitations

As mentioned by Kawamura in the DDSupport paper utilizing the UME-ERJ corpus and HuBERT based phonemic analysis [12], Because the UME-ERJ corpus is unbalanced, with just 20 AE speakers for 202 JE speakers, distance metrics for less frequent phonemes may exhibit statistical bias. For example, this may result in the higher than expected deviations for /ɾ/ due to either limited native comparison or the assumption of /ɾ/ sounds by JE speakers as a substitute for other high-deviation consonants

The reliance on MFA for phonemic classification may also disrupt the temporal phonemic classification of Japanese speakers, having been trained almost entirely on American English data.

Lastly, there is the lack of tests with Mahalanobis diagonal mapping. The hardware, as mentioned in methodology, was deemed insufficiently powerful for the 768x768 images, and was not worth producing in the author's resource-strained state.

VI. Conclusion

This work presented a large-scale, phoneme-level analysis of pronunciation deviations in Japanese learners of English using SSSRs. By leveraging HuBERT embeddings in conjunction with forced alignment, we quantified phonetic drift between Japanese English (JE) speakers and an American English (AE) baseline without reliance on supervised pronunciation scoring or handcrafted acoustic features. The results demonstrate that SSL-based representations are sufficiently sensitive to capture fine-grained phonemic deviations that are well-documented in second language acquisition literature, including liquid and labialized consonants, rhotics, and (dental) fricatives. [8], [9], [19]

Across all experimental conditions, the alveolar flap /ɾ/ consistently exhibited the highest deviation, confirming its central role in Japanese–English phonological interference [6], [19]. Additional high-drift phonemes, such as /g^{w}/, /ɫ/, /ð/, and rhotic vowels, further reinforce known articulatory and perceptual mismatches between the two phonological systems. Sex-based analyses revealed that while male speakers exhibited lower average deviation across most phonemes, a small subset of highly divergent phonemes disproportionately increased their global drift metric, highlighting the importance of phoneme-wise rather than aggregate evaluation.

The global aggregate analysis demonstrated that pronunciation deviation in Japanese learners is not uniformly distributed across the phoneme inventory, but instead concentrated within a relatively small subset of sounds. This finding has direct implications for Computer-Assisted Language Learning systems, suggesting that targeted, phoneme-specific feedback may be substantially more effective than holistic pronunciation feedback [15], [19]. Importantly, the use of self-supervised representations eliminates the need for large-scale labeled non-native speech corpora, significantly lowering the barrier for extensible and language-agnostic pronunciation assessment [7], [8].

Overall, this study demonstrates that SSSR models provide a robust foundation for objective, data-driven pronunciation analysis in second language learning contexts.

All files utilized in the project, barring from UME-ERJ audio data and transcripts, are availible on the author's GitHub page, accessible through the OrCID link. This includes the final .pkl weights used in all the comparisons, held under a standard LGPL-2.1 license.

VII. Acknowledgements

The original idea of studying Japanese Learners of English came from working on an app to aid in pronunciation with a focus on Japan; a CALL of our own, for the Intel Digital Readiness Project, mentor-led program held weekly at our school. I would like to acknowledge our school, India International School in Japan for providing an environment where curiosity was encouraged, and thrived. My teachers also played a large part in motivating me to conduct independent research. Mrs. Charu Negi has been extremely helpful in reviewing my paper and suggesting improvements. I would also like to present my gratitude to my fellow teammates, Yatharth Jain, Zayaan Nanji and Shaunak Jadhav, for their help in the initial stages of the project. Zayaan Nanji in particular has been of great help in taking our findings and applying them within the final CALL application.

Lastly, this project was possible due to the cooperation provided by the NII-SRC, in granting us access to the UME-ERJ dataset, which was an integral part of the research.

References

[1] D. A. Patankar, "The evolution of English language learning in Japan: A historical and statistical analysis," Int. J. Prog. Res. Eng. Mgmt. Sci., vol. 5, no. 1, pp. 1411–1420, 2025.

[2] P. Cutrone, "Overcoming Japanese EFL learners' fear of speaking," Language Studies Working Papers, vol. 1, pp. 55–63, 2009.

[3] G. E. Oh, S. Guion-Anderson, K. Aoyama, J. S. Flege, R. Akahane-Yamada, and T. Yamada, "The effect of age of acquisition on first- and second-language vowel production," J. Phonetics, vol. 39, no. 4, pp. 585–600, 2011.

[4] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in Proc. Interspeech 2017, pp. 498–502, 2017.

[5] K. Martin, J. Gauthier, C. Breiss, and R. Levy, "Probing self-supervised speech models for phonetic and phonemic information: a case study in aspiration," in Proc. Interspeech 2023, 2023, pp. 2355–2359.

[6] A. M. Diaz, "Phonetics applied: An acoustic analysis of Japanese L2 English," in P. Ferguson et al. (eds.), Learning from Students, Educating Teachers – Research and Practice (Proc. JALT 2022), JALT, 2023, pp. 67–75.

[7] S. Williams, P. Foulkes, and V. Hughes, "Analysis of forced aligner performance on L2 English speech," Speech Communication, vol. 158, pp. 1–14, 2024.

[8] M. Shahin, J. Epps, and B. Ahmed, "Phonological level Wav2Vec 2.0-based mispronunciation detection and diagnosis method," Speech Communication, vol. 173, pp. 1–15, 2025.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 12449–12460.

[10] 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 Trans. Audio, Speech, Lang. Process., vol. 29, pp. 3451–3460, 2021.

[11] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "English Speech Database Read by Japanese Students (UME-ERJ)," developed under the Grant-in-Aid for Scientific Research (MEXT, Japan), 2001–2003. Primary descriptions appear in: *Proc. ASJ Autumn Meeting*, pp.199–200, 2001; *Proc. COCOSDA Workshop*, pp.76–81, 2001; *Proc. LREC*, pp.896–903, 2002; *J. Acoust. Soc. Jpn.*, vol.59, no.6, pp.345–350, 2003; and *J. Jpn. Soc. Educ. Technol.*, vol.27, no.3, pp.259–272, 2003.

[12] K. Kawamura, J. Rekimoto, "DDSupport: Language Learning Support System that Displays Differences and Distances from Model Speech" arXiv:2212.04930

[13] K. Choi, E. Yeo, K. Chang, S. Watanabe, "Leveraging allophony in self-supervised speech models for atypical pronunciation assessment," in Proc. NAACL 2025.

[14] T. Nagamine, "Dynamic tongue movements in L1 Japanese and L2 English liquids," in Proc. 20th Int. Congress of Phonetic Sciences (ICPhS), 2023.

[15] P. Sofroniev and Ç. Çöltekin, ”Phonetic vector representations for sound sequence alignment,” in Proc. 15th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, Brussels, Belgium, 2018, pp. 111–116.

[16] D. Wells, H. Tang, and K. Richmond, ”Phonetic analysis of self-supervised representations of English speech,” in Proc. Interspeech 2022, pp. 3583–3587, 2022.

[17] J. Moore, J. Shaw, S. Kawahara, T. Arai, ”Articulation strategies for English liquids used by Japanese speakers,” Acoust. Sci. Technol., vol. 39, no. 2, pp. 1-10, 2018.

[18] A. Kitagawa, ”An acoustic analysis of English pronunciation systems of Japanese learners,” Waseda University, 2016.

[19] Q. Wang and K. A. Lee, ”Cosine scoring with uncertainty for neural speaker embedding,” arXiv:2403.06404 [cs.SD], 2024.

[20] P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages*. Oxford, UK: Blackwell, 1996.

[21] T. J. Vance, *The Sounds of Japanese*. Cambridge, UK: Cambridge University Press, 2008.

[22] V. W. Zue and M. Laferriere, ”Acoustic study of medial /t, d/ American English,” *J. Acoust. Soc. Am.*, vol. 66, no. 4, pp. 1039-1050, 1979.