JPhonABX: Japanese Phonological Discrimination Benchmark for Moraic and Accentual Minimal Pairs

Anonymous ACL submission

Abstract

Self-supervised speech models are known to encode phonological information, yet it remains underexplored whether they capture phonological properties characteristic of Japanese, such as mora and pitch accent. This paper introduces JPhonABX, a Japanese phonological discrimination benchmark targeting moraic and accentual contrasts. It consists of 192,000 moraic and 98,000 accentual ABX test items, covering various types of contrasts such as those involving moraic nasal, geminate consonant, vowel length, and pitch accent. Our human evaluation by native Japanese speakers shows high discrimination accuracy and inter-rater agreement, validating the quality of the benchmark. We further evaluate several self-supervised speech models and find that while they perform comparably to humans in mora discrimination, they perform substantially worse in accent discrimination, and even models trained on Japanese fail to reach human-level performance.

1 Introduction

Self-supervised speech models (S3Ms) have become the dominant approach to speech representation learning, due to their strong performance in downstream tasks (Yang et al., 2021; Mohamed et al., 2022; Tsai et al., 2022). Recent work has shown that S3M representations capture various types of information, from non-linguistic one such as acoustic (Pasad et al., 2023) and speaker (Fan et al., 2021) information to linguistic one such as phonological (Martin et al., 2023; Bentum et al., 2025), lexical (Pasad et al., 2024), syntactic (He et al., 2025), and semantic (Yang et al., 2021) information.

One way to study phonological information in S3M representations is the ABX discrimination test. In this test, a model is asked to determine whether a target stimulus X is more similar to stimulus A or B in the representational space. Because the evaluation directly compares speech representations without additional supervised training, ABX provides a lightweight framework for evaluating phonological information in S3Ms. This test can also be applied to humans, enabling direct comparison between S3Ms and humans. Phonological ABX benchmarks have been developed for several languages including English, French, German, and Dutch (Dunbar et al., 2017; Nguyen et al., 2020; de Heer Kloots et al., 2025). However, the systems of these languages do not involve mora or pitch accent contrasts, which are characteristic of Japanese. While prior work has created small-scale benchmarks targeting specific Japanese moraic and pitch-accent contrasts (Ren, 2022; Mugitani et al., 2009; Sun et al., 2026), there is still no large-scale and comprehensive Japanese benchmark.

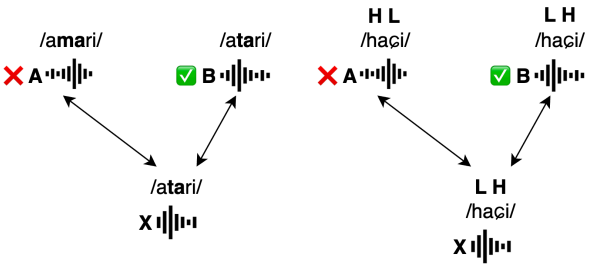


Figure 1: Overview of the mora discrimination (left) and pitch accent discrimination (right) ABX task used in the JPhonABX benchmark. The task is to select the closer stimuli to X from A and B. For mora discrimination, the morae "ma" and "ta" are discriminated. For pitch accent discrimination, the accent patterns "HL" and "LH" are discriminated.

To fill this gap, we introduce JPhonABX, a large-scale phonological discrimination ABX benchmark targeting various types of Japanese moraic and accentual minimal pairs, including those involving moraic nasal, geminate consonant, and vowel length. This ABX test evaluates the discriminability of sound stimuli differing in mora sequence or pitch accent pattern (see Figure 1 for examples). Our benchmark consists of 192,000 ABX

test items for mora discrimination and 98,000 items for accent discrimination, both extracted from a Japanese read speech corpus. To our knowledge, this is the largest-scale and the most comprehensive phonological ABX benchmark that fully reflects the phonological properties of Japanese.

To assess the quality and reliability of the benchmark, we conduct a human evaluation on a subset of the items, recruiting native Japanese speakers as participants. We find that humans achieve high discrimination accuracy with moderate inter-rater agreement, indicating that the benchmark items are well-formed. In addition, we report the performance of several speech foundation models on the whole dataset for further analysis. Regarding mora discrimination, we find that S3Ms perform comparably to native Japanese speakers even when trained on other languages than Japanese. For accent discrimination, however, S3M accuracy is considerably lower than in mora discrimination, particularly for models not trained on Japanese, and all models perform worse than humans regardless of their training language. These results suggest that while moraic contrasts are well encoded by S3Ms regardless of their training language, pitch accent contrasts are difficult for S3Ms and moderately encoded only when trained on Japanese.

2 Related Work

2.1 Phonological Evaluation of Speech Representations

The phonological information encoded in S3M representations has been the focus of a large body of work. Probing studies using classifier probes have examined both phoneme-related and prosody-related tasks. Phoneme-related tasks are exemplified by phoneme recognition (Ma et al., 2021; Cormac English et al., 2022; Martin et al., 2023; Shen et al., 2024; de Heer Kloots et al., 2025). Prosody-related tasks include lexical tone and stress prediction (Yang et al., 2023; de la Fuente and Jurafsky, 2024; Shen et al., 2024; Bentum et al., 2025; Gogoi et al., 2025), phrasal accent prediction (de la Fuente and Jurafsky, 2024), and prosody-sensitive tasks such as sentiment analysis, sarcasm detection, and persuasiveness detection (Lin et al., 2023). These studies report high accuracy achieved by S3Ms, suggesting that phonemic and prosodic information is encoded in S3M representations.

While probing can be applied to a wide range of tasks, it requires substantial computational cost, and the probing performance can be sensitive to task design (Hewitt and Liang, 2019; Voita and Titov, 2020; Ravichander et al., 2021; Belinkov, 2022). An alternative is intrinsic evaluation, which assesses properties of representations without training additional classifiers. Canonical Correlation Analysis (CCA) has been used to investigate how well representations correlate with phoneme or word labels (Pasad et al., 2021, 2023, 2024). Word discrimination tests evaluate whether two representations correspond to the same word by comparing their similarity (Sanabria et al., 2023; Pasad et al., 2024). The ABX discrimination test (Schatz et al., 2013), derived from human psychophysics, has been used to evaluate whether representations can distinguish phonemic or prosodic minimal pairs by comparing distances in representational space (Dunbar et al., 2021; de Heer Kloots et al., 2025; Sun et al., 2026). In this work, we adopt the ABX discrimination test as our evaluation framework in order to compare the results with humans.

2.2 Phonological ABX Benchmarks

Several benchmarks have been proposed for phonological ABX evaluation. Regarding phone discrimination, the ZeroSpeech benchmark series (Versteegh et al., 2015; Dunbar et al., 2017; Nguyen et al., 2020) provides large-scale evaluation datasets across six languages: English, French, German, Mandarin, Wolof, and Xitsonga. More recently, a similarly large-scale Dutch dataset was introduced (de Heer Kloots et al., 2025).

For Japanese, a few large-scale ABX datasets (Schatz et al., 2021; Poli et al., 2024) have been constructed for phone discrimination from the GlobalPhone (Schultz, 2002) and CSJ (Maekawa, 2003) corpora. Schatz and Feldman (2018) additionally targeted vowel length contrasts from the same corpora. However, none of these datasets fully reflect the phonological properties of Japanese: the contrasts they use are not based on morae and they do not control for accentual context, which means that minimal pairs may differ not only in phonemic sequence but also in pitch accent pattern, making the evaluation less controlled.

Human perception studies have also employed phonemic or moraic minimal pairs in their experiments, mainly targeting contrasts characteristic of Japanese, such as vowel length contrasts (Mugitani et al., 2009; Fonseca, 2023) and presence or absence of geminate consonant (Ren, 2022). These

studies evaluate the performance of learners of Japanese as a second language or children, and involve only a small number of test items, with fewer than 1,000.

Regarding accent discrimination, several small-scale datasets exist for Japanese. Sun et al. (2026) proposed an ABX benchmark covering Japanese, English, and Mandarin accentual contrasts for representation evaluation, and human perception studies have also produced accentual minimal pair datasets in Japanese (Atria and Hazan, 2015; Fonseca, 2023). However, none of these are large enough for systematic evaluation of speech representations, with 3,000 or fewer test items.

We address these limitations by providing a large-scale ABX benchmark that controls for both phonemic content and pitch accent pattern, covering moraic and accentual minimal pairs.

3 JPhonABX

3.1 Task Description

The ABX discrimination test, derived from the discrimination paradigm in human psychophysics, presents three stimuli A, B, and X, and asks whether X belongs to the same category as A or B (Schatz et al., 2013). In our phonological discrimination tasks, A and B are speech stimuli that differ only in mora sequence (mora discrimination) or pitch accent pattern (accent discrimination). The stimuli A and B are produced by the same speaker.

The model performance is computed by comparing the representational distance between A and X to that between B and X in each triplet. In the following, the stimulus (A or B) that matches X is referred to as *target*, while the other is referred to as *other*. The discrimination score D is defined as follows (Millet and Dunbar, 2022):

$$D = \mathrm{DTW}(M_{other}, M_X) - \mathrm{DTW}(M_{target}, M_X),$$

where DTW denotes the distance computed using dynamic time warping with cosine similarity, and M_x represents the representation of stimulus x. Dynamic time warping computes the distance between two sequences of feature vectors by finding an optimal alignment that minimizes the cumulative frame-level distance under monotonicity and continuity constraints. This enables sequences of different lengths to be compared while allowing temporal variation. A positive D indicates that the model correctly identifies the target, while a negative D indicates failure.

Table 1: Mora contrast types in the JPhonABX dataset.

Subset	Contrast Type	Example
Normal	Initial mora	***boo**to* vs. ***too**to*
	Medial mora	*a**ma**ri* vs. *a**ta**ri*
	Final mora	*san**so*** vs. *san**do***
EN-hard	Moraic nasal	*i**n**shi* vs. *ishi*
	Geminate consonant	*mo**tt**o* vs. *moto*
	Vowel length	*f**ii**ru* vs. *firu*

3.2 Mora Discrimination

A mora is a rhythmic unit in Japanese that takes the form of either CV, CjV, V, /N/ (moraic nasal), or /cl/ (geminate consonant), and is characterized by a relatively uniform duration across all morae. The JPhonABX dataset covers six distinct types of mora contrasts shown in Table 1, and consists of two subsets: Normal and EN-hard. Contrasts in the Normal subset involve three-mora sequences, where A and B differ only in the initial, medial, or final mora. The EN-hard subset targets contrasts known to be particularly challenging for English-speaking learners of Japanese as a second language (Toda, 1998): the presence or absence of a moraic nasal, a geminate consonant, and the distinction between long and short vowels. For these contrasts, the *target* stimulus and X are three-mora sequences, while *other* is a two-mora sequence.

The pitch accent pattern is matched between A and B for each triplet, ensuring that models focus solely on differences in mora sequence when determining whether X is closer to A or B. Pitch accent patterns are defined by assigning either H (high) or L (low) to each mora, reflecting relative pitch height within an utterance. For example, *hashi* can be read with the "HL" accent to refer to chopsticks, or with the "LH" accent to refer to bridges. For the EN-hard subset, a full match of accent patterns between A and B is impossible, as A and B differ in mora count. Instead, we match the accent patterns of the first and last morae of A and B. We treat HH and LL as equivalent, since these patterns differ only in overall pitch height rather than in contour shape.

We further define two conditions based on the accent pattern of X: a within-accent condition, in which X shares the same accent pattern as A and B, and an across-accent condition, in which it does not. This design allows us to examine whether a model encodes stimuli of the same mora sequence closely in representational space, even when their pitch accent patterns differ. Similarly, the dataset includes within-speaker, across-speaker, within-gender, and across-gender conditions, to evaluate robustness to

speaker variation and acoustic differences between male and female voices.

3.3 Accent Discrimination

The accent discrimination test consists of contrasts between two-mora sequences and those between three-mora sequences. In each triplet, A and B share the same mora sequence and differ only in pitch accent pattern. Similar to the mora discrimination test, we define within-speaker, across-speaker, within-gender, and across-gender conditions.

3.4 Data Construction

All audio files were extracted from the parallel100 subset of the JVS corpus (Takamichi et al., 2019), which consists of clean recordings of 100 sentences, each read aloud by 100 speakers in Standard Japanese. Files containing truncations, misreadings, or other recording errors were excluded. We used speaker information distributed with JVS, as well as transcripts and text-speech alignments distributed by a third party.[1]

Mora-level accent labels were automatically assigned using a publicly available model from HuggingFace.[2] This model takes a spoken sentence as input and outputs a predicted transcript with accent labels. To assign labels to the gold transcripts, we aligned the predicted transcripts with the gold ones. We permitted minor transcription errors if the predicted transcript could be corrected to the gold transcript via a manually defined substitution dictionary (see Appendix A for details). The dictionary mainly covers three types of errors: (1) single-mora confusions, (2) single-mora decompositions, and (3) two-mora compressions. Following this alignment procedure, 276 files for which alignment failed were removed.

To further validate the accent labels, we used fundamental frequency (F0) information predicted with CREPE (Kim et al., 2018). F0 was estimated at a 10ms resolution, and the average F0 was computed for each mora using the distributed text-speech alignments. Then, the F0 ratio was computed as follows:

$$\text{F0ratio}_i = \begin{cases} \log \dfrac{\text{F0}_{i+1}}{\text{F0}_i} & \text{if } m_{i+1} \neq \text{/cl/}, \\ \log \dfrac{\text{F0}_{i+2}}{\text{F0}_i} & \text{if } m_{i+1} = \text{/cl/}, \end{cases}$$

[1] https://github.com/Hiroshiba/jvs_hiho

[2] https://huggingface.co/AkitoP/whisper-large-v3-japense-phone_accent

Table 2: The number of within-/across-speaker/gender/accent triplets. The number of within-speaker triplets is relatively small because the JVS corpus contains speech from 100 speakers.

Test Type	Condition	#Test Items
Mora Discrimination	within-speaker	11,559
	across-speaker	180,441
	within-gender	101,562
	across-gender	90,438
	within-accent	120,578
	across-accent	71,422
Accent Discrimination	within-speaker	5,992
	across-speaker	92,008
	within-gender	52,698
	across-gender	45,302

where m_i and F0_i denote the i-th mora and its F0, respectively. Since the geminate consonant is not pronounced and thus does not have a corresponding F0, F0ratio is computed between the two adjacent morae. Then, based on F0 ratios and thresholds derived from previous accent type prediction studies (Kawai and Ishi, 1999; Short et al., 2011), accent labels that were inconsistent with the F0 patterns were discarded. Specifically, if F0ratio_i is less than or equal to -0.13, the accent labels of the i-th and $i+1$-th morae should be "HL", if it is bigger than or equal to 0.054, they should be "LH", and otherwise "HH" or "LL".

In the next step, all possible speech files of two-mora and three-mora sequences were extracted, and those starting or ending with a geminate consonant were excluded. Short files were also excluded from the dataset, specifically those shorter than 150 ms for two-mora sequences and 225 ms for three-mora sequences, as they did not sound like the transcripts. Using the resulting files, all valid ABX triplets were generated and categorized into the six contrast types for mora discrimination in Table 1 or the two types for accent discrimination. We sampled 32,000 triplets for each type of mora discrimination and 49,000 for accent, balancing the numbers of within-/across- speaker/gender/accent triplets as much as possible (see Table 2).

4 Human Evaluation

To assess the quality of the dataset, we conducted a human evaluation on a randomly sampled subset of 1000 triplets (500 mora discrimination + 500 accent discrimination). Three native Japanese speakers participated in each discrimination task. For each ABX test item, participants listened to three speech stimuli (A, B, and X) presented in this order, without altering the playback speed. Then, they were asked to indicate whether X had the same mora sequence or accent pattern as A or B by choosing one

of three answers: A, B, or "not sure". We asked participants to choose "not sure" when A and B were perceived as having the same mora sequence or accent pattern, or when X was perceived as differing from both A and B. Participants were allowed to adjust the playback volume and to listen to each stimulus as many times as they wished. Detailed instructions can be found in Appendix B.

We computed the accuracy of human participants as well as the agreement between participants measured by Krippendorff's α (Krippendorff, 2011). The accuracy was computed for each triplet first using data from all the participants, and then averaged across all the triplets, following the method used in Perceptimatic (Millet and Dunbar, 2020). The option "not sure" was treated as one of the valid options, not as a missing value.

The evaluation results are presented in Table 3. Overall, accuracy is high for both test types, indicating sufficient benchmark quality. For mora discrimination, accuracy exceeds previously reported human performance on existing phoneme discrimination datasets (0.797 for English and 0.840 for French (Millet and Dunbar, 2020)) in all contrast types except vowel length. The higher performance may also reflect longer stimulus or contrastive portion duration. Indeed, humans achieve accuracy of 0.952 for contrasts with two phoneme differences (e.g., **ra**ndo vs. **se**ndo), but only 0.870 for those with one phoneme difference (e.g., san**s**o vs. san**d**o). The lower accuracy on the vowel length contrast may reflect the fact that the length of long vowels depends on speaking rate, and that the difference in vowel length is difficult to perceive when spoken faster.

In terms of inter-rater agreement, most contrast types exhibit α values above 0.6, indicating reasonable reliability of the dataset. Exceptions are the moraic nasal and vowel length contrasts, whose lower agreement can be attributed to their reliance on durational cues. The reliability of these contrasts should be interpreted with caution.

5 Model Evaluation

5.1 Models

In addition to the human evaluation, we evaluated several pretrained S3Ms to benchmark performance on the whole dataset. We aim to investigate which layers encode mora and pitch accent information, and how this is affected by factors such as training objective and language. We limited the models to HuBERT (Hsu et al., 2021) and wav2vec 2.0 (Baevski et al., 2020), which have Japanese-trained versions. We downloaded six Japanese, five English, and three multilingual models from HuggingFace (see Appendix C for details) and used them as they are.[3] Japanese models were pretrained on Japanese TV broadcast audio with two different model sizes and three training corpus sizes. English models were pretrained on audiobook-based corpora (LibriSpeech (Panayotov et al., 2015) or Libri-Light (Kahn et al., 2020)) with two model sizes and two training corpus sizes. Multilingual models covered between 53 and 147 languages, with varying amounts of Japanese speech in their pretraining corpora. As acoustic baselines, we employed 13-dimensional MFCC and mel-spectrogram features computed over 25ms windows with a stride of 10ms.

We extracted representations of A, B, and X in two settings: *with-context*, where representations are extracted from sentence-level input, and *without-context*, where stimuli are fed directly to the model. In the latter setting, due to computational constraints, we used 421 mora and 495 accent discrimination ABX test items used in the human evaluation, excluding items with at least one "not sure" response to ensure comparability with human judgments. We report accuracy and item-wise Kendall's τ correlation score for every Transformer layer of base models and every two Transformer layers of large models. Kendall's τ is computed between each model's discrimination score D and human accuracy.

5.2 Mora Discrimination

5.2.1 Overview

Figure 2 shows the layer-wise accuracy in the with-context setting, and Table 4 shows the accuracy and Kendall's τ at the layer with peak accuracy in the without-context setting. In the following, the peak accuracy refers to the highest accuracy across all layers. Overall, S3Ms outperform MFCC and mel-spectrogram baselines at most layers, suggesting that S3Ms encode richer linguistic information. When compared to humans, S3Ms perform equally well and their discrimination scores correlate weakly but positively with human accuracy, regardless of their training language.

[3] We did not use imprt/izanami-wav2vec2-large, as our preliminary experiment showed low phoneme recognition performance in linear probing (14% accuracy in layer 12), suggesting limited phonemic information.

Table 3: Human evaluation results on the 500 mora discrimination test items and 500 accent discrimination ones randomly sampled from JPhonABX.

Test Type	Contrast Type	#Test Items	Avg. Acc.	Krippendorff's α	not sure ratio
Mora Discrimination	Initial mora	91	0.853	0.631	0.070
	Medial mora	84	0.917	0.737	0.044
	Final mora	82	0.967	0.876	0.024
	Moraic nasal	97	0.852	0.503	0.058
	Geminate consonant	79	0.928	0.792	0.038
	Vowel length	67	0.612	0.469	0.194
	Overall	500	0.862	0.665	0.067
Accent Discrimination	2-mora	253	0.881	0.724	0.007
	3-mora	247	0.865	0.717	0.001
	Overall	500	0.873	0.721	0.004

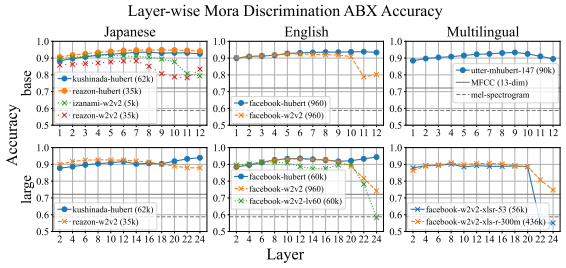


Figure 2: Accuracy in the with-context (n=192,000) setting. Figures in brackets show training data size (h).

Table 4: Accuracy and Kendall's τ in the layer with peak accuracy for each model for mora discrimination in the without-context (n=421) setting.

Lang.	Model Size	Model	Acc.	Kendall's τ
JA	base	kushinada-hubert (62k)	0.957	0.200
		reazon-hubert (35k)	0.948	0.199
		izanami-w2v2 (5k)	0.931	0.204
		reazon-w2v2 (35k)	0.891	0.214
	large	kushinada-hubert (62k)	0.950	0.209
		reazon-w2v2 (35k)	0.926	0.184
EN	base	facebook-hubert (960)	0.933	0.222
		facebook-w2v2 (960)	0.914	0.193
	large	facebook-hubert (60k)	0.943	0.195
		facebook-w2v2 (960)	0.926	0.194
		facebook-w2v2-lv60 (60k)	0.943	0.215
Multi	base	utter-mhubert-147 (90k)	0.931	0.216
	large	facebook-w2v2-xlsr-53 (56k)	0.926	0.212
		facebook-w2v2-xls-r-300m (436k)	0.933	0.215
-	-	MFCC	0.774	0.139
-	-	mel-spectrogram	0.587	0.024
-	-	Human	0.929	-

HuBERT and wav2vec 2.0 show a clear difference in how accuracy evolves across layers, regardless of training language and model size: HuBERT achieves its peak at a deeper layer and shows a smaller accuracy drop in the final layers. This trend has been observed in previous work on phoneme information (Pasad et al., 2023), and can be attributed to iterative refinement in HuBERT's training (Huo and Dunbar, 2025). HuBERT is trained for two iterations; in the second iteration, pseudo-labels are derived from an intermediate layer of the first-iteration model. Since intermediate layers are known to capture more abstract linguistic information than the final layers (Pasad et al., 2021), these pseudo-labels are more linguistically informative than the acoustic pseudo-labels used by wav2vec 2.0. This may explain why HuBERT maintains higher accuracy in its latter layers, as they are directly used to predict the pseudo-labels.

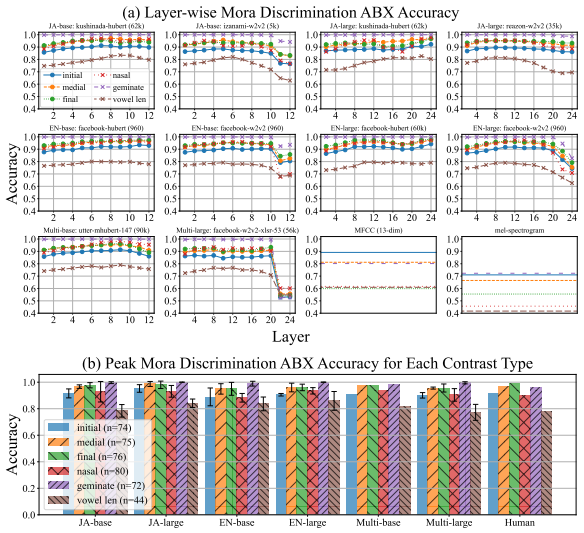


Figure 3: Comparison between the six contrast types. (a) Results in the with-context (n=192,000) setting. See Appendix D for full results. (b) Peak ABX accuracy in the without-context (n=421) setting averaged within each model group. Error bars indicate standard deviation across models.

5.2.2 Difference between Contrast Types

Figure 3 compares the six contrast types in Table 1. For most S3M layers, geminate consonant contrasts are distinguished most easily, and vowel length contrasts are the most challenging. This trend holds in the without-context setting, and humans also show a similar trend. The difficulty of vowel length contrasts may reflect the nature of DTW: since the additional morae in long vowels are acoustically nearly identical to the preceding short vowels, DTW may align them completely with the corresponding short vowel frames, contributing little to the cumulative distance. The quality of vowel length contrasts may

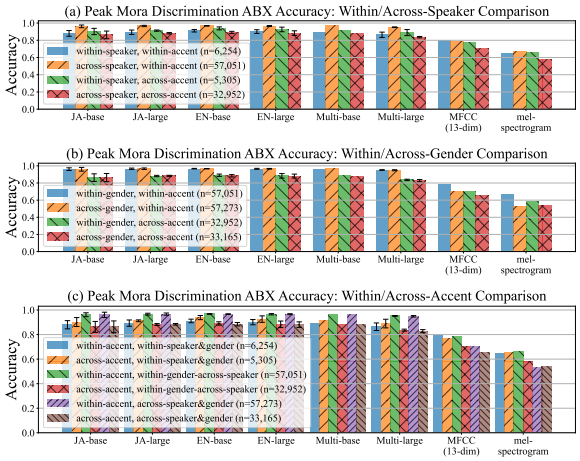


Figure 4: Comparison between within- and across- (a) speaker, (b) gender, and (c) accent conditions in the with-context (n=192,000) setting. In (a), gender condition is within and in (b), speaker condition is across. Accuracy is averaged within each model group. Error bars indicate standard deviation across models.

also be a factor, as suggested by the relatively low human accuracy.

The layer-wise trend of each contrast type follows the overall trend in Figure 2 for most S3Ms, with one exception: HuBERT models achieve nearly 100% accuracy on geminate consonant contrasts across all layers, regardless of training language and model size.

Figure 4 compares peak ABX accuracy between within- and across- speaker, gender, and accent conditions in the with-context setting. Across-speaker accuracy exceeds within-speaker accuracy in the within-accent condition, and within-accent accuracy exceeds across-accent accuracy in the within-speaker&gender condition. These reversals may reflect the influence of context on the representations. Gender condition has little effect on accuracy, supporting the view that S3M representations are largely invariant to speaker gender.

5.2.3 Effect of Training Language, Training Data Size, and Model Size

Comparing Japanese and English models with the same training objective, architecture, and similar training data size (Japanese large kushinada-hubert (62k) and English large facebook-hubert (60k) in Figure 2), peak performance is nearly identical. The EN-hard subset shows a similar pattern (Figure 3), with the only notable difference being slightly lower accuracy for English models on the vowel length contrast. These results suggest that S3M representations are not strongly phonemically language-specific.

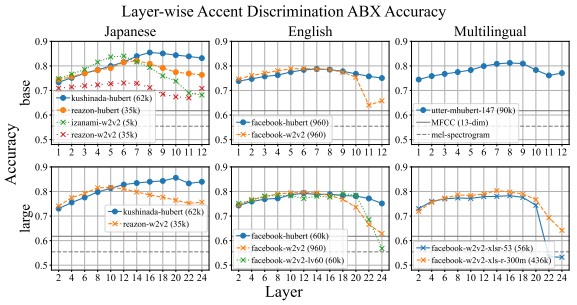


Figure 5: Accuracy in the with-context (n=98,000) setting. Figures in brackets show training data size (h).

Table 5: Accuracy and Kendall's τ in the layer with peak accuracy for each model for accent discrimination in the without-context (n=495) setting.

Lang.	Model Size	Model	Acc.	Kendall's τ
JA	base	kushinada-hubert (62k)	0.752	0.144
		reazon-hubert (35k)	0.747	0.109
		izanami-w2v2 (5k)	0.766	0.158
		reazon-w2v2 (35k)	0.673	0.109
	large	kushinada-hubert (62k)	0.812	0.139
		reazon-w2v2 (35k)	0.731	0.101
EN	base	facebook-hubert (960)	0.701	0.061
		facebook-w2v2 (960)	0.711	0.093
	large	facebook-hubert (60k)	0.713	0.087
		facebook-w2v2 (960)	0.743	0.116
		facebook-w2v2-lv60 (60k)	0.768	0.141
Multi	base	utter-mhubert-147 (90k)	0.756	0.132
	large	facebook-w2v2-xlsr-53 (56k)	0.762	0.101
		facebook-w2v2-xls-r-300m (436k)	0.824	0.181
-	-	MFCC	0.630	0.058
-	-	mel-spectrogram	0.572	0.128
-	-	Human	0.877	-

Increasing training data does not improve performance, indicating that mora contrasts are well captured even with a corpus of 5k hours. In contrast, increasing model size consistently improves performance, suggesting that larger models are better able to capture moraic information.

5.3 Accent Discrimination

5.3.1 Overview

Figure 5 shows the layer-wise accuracy in the with-context setting, and Table 5 shows the accuracy and Kendall's τ at the layer with peak accuracy in the without-context setting. All in all, S3Ms outperform MFCC and mel-spectrogram baselines at most layers, while all S3Ms perform worse than native Japanese speakers with little positive correlation, suggesting that even Japanese S3Ms do not adequately encode pitch accent patterns.

As in mora discrimination, HuBERT and wav2vec 2.0 show different layer-wise trends, likely due to iterative refinement. However, overall accuracy is up to 20% lower than mora discrimination, suggesting the relative difficulty of accent

discrimination. In addition, the accuracy of HuBERT models peaks at earlier layers than in mora discrimination.

Comparing Japanese and English models with the same training objective, architecture, and a similar training data size (Japanese large kushinada-hubert (62k) and English large facebook-hubert (60k) in Figure 5), the Japanese model achieves much higher peak accuracy. In addition, the peak performance of the multilingual model trained on Japanese data (utter-mHuBERT-147 (90k)) falls between Japanese and English models with the same architecture, while the other multilingual models perform similarly to English models. These results suggest that exposure to Japanese speech is necessary for better pitch accent discrimination, given the language-specific nature of the pitch accent system.

Increasing training data improves peak performance for the HuBERT model but not for wav2vec 2.0 models, while increasing model size benefits both. The wav2vec 2.0 models may learn most of the useful cues for accent discrimination from relatively small amounts of data.

5.3.2 Difference between Contrast Types

Figure 6 compares the contrasts between two pitch accent patterns that are both possible as word accent patterns in Standard Japanese (i.e., HL, LH, HLL, LHH, LHL) against contrasts involving at least one impossible pattern. Most S3M layers, acoustic baselines, and humans achieve higher accuracy on the former contrasts in both with- and without- context settings. This suggests that these contrasts are acoustically more salient.

Figure 7 compares peak accuracy between within- and across- speaker and gender conditions in the with-context setting. Across-speaker accuracy exceeds within-speaker accuracy, possibly reflecting the effect of context. Gender condition has little effect on accuracy, consistent with the mora discrimination results.

6 Conclusion

In this work, we introduced JPhonABX, a large-scale Japanese phonological discrimination benchmark targeting moraic and accentual contrasts. Human evaluation by native Japanese speakers yielded high discrimination accuracy and inter-rater agreement, indicating sufficient quality of our benchmark. Model evaluation revealed a clear asymmetry between the two tasks. For mora discrimination, S3Ms achieved accuracy comparable to that of native Japanese speakers regardless of their training language, suggesting that moraic contrasts in Japanese are well captured by S3M representations even without Japanese-specific training. For accent discrimination, however, S3M accuracy was markedly lower than in mora discrimination, particularly for models not trained on Japanese data. Notably, even Japanese S3Ms performed below human level, suggesting that pitch accent contrasts pose a greater challenge and require exposure to Japanese speech to be better encoded. Future work includes incorporating our benchmark to investigate methods to improve pitch accent encoding in S3Ms.

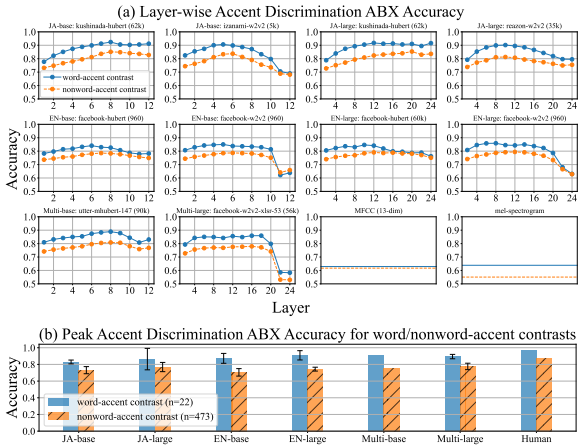


Figure 6: Comparison between word- and nonword-accent contrasts. Word-accent contrasts are those between two pitch accent patterns that are both possible as word accent patterns in Standard Japanese (i.e., HL, LH, HLL, LHH, LHL), and nonword-accent contrasts are those involving at least one impossible pattern. (a) Results in the with-context (n=4,337 (word) + 93,663 (nonword)) setting. See Appendix D for full results. (b) Peak ABX accuracy in the without-context setting averaged within each model group. Error bars indicate standard deviation across models.

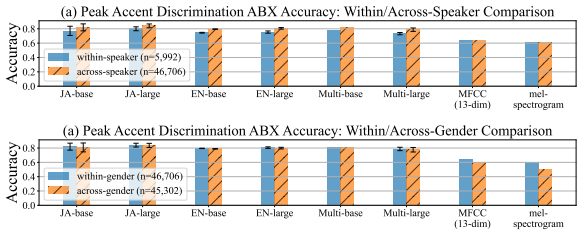


Figure 7: Comparison between within- and across- (a) speaker and (b) gender conditions in the with-context (n=98,000) setting. In (a), gender condition is within and in (b), speaker condition is across. Accuracy is averaged within each model group. Error bars indicate standard deviation across models.

Limitations

We identify three limitations regarding the quality of our benchmark. First, our human evaluation involved three participants per task, which may be insufficient for reliable estimation of human performance. Second, each stimulus (A, B, X) was extracted automatically using a publicly available text-speech alignment data. This was produced by a speech recognition model and may therefore contain mora boundary misalignments. We did not manually verify whether each stimulus sounded exactly as the corresponding transcript, and the relatively low human accuracy and inter-rater agreement on the vowel length contrast suggest that some stimuli may be of insufficient quality. Third, our benchmark is constructed from read speech in Standard Japanese and may not fully reflect the properties of dialectal or spontaneous speech.

Regarding our analysis, we evaluated a limited number of S3Ms, and other models may show different trends. In addition, the without-context evaluation and human evaluation were conducted on only 500 items per task, which limits the reliability of fine-grained analyses such as per-contrast-type comparisons for mora discrimination and within/across condition comparisons.

Ethical Considerations

All models used in our experiments were used in accordance with their respective licenses. For the human evaluation, participants' responses were initially linked to email addresses for data collection purposes. After collection, the email address information was removed, and all analyses were conducted on anonymized data that contained no personally identifiable information.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

José Atria and Valerie Hazan. 2015. Development of accentual categories in Japanese as a second language. Conference Name: 18th International Congress of Phonetic Sciences,, 10-14 August 2015, Glasgow ISSN: 241-0669 Meeting Name: 18th International Congress of Phonetic Sciences,, 10-14 August 2015, Glasgow Volume: 2015.

Madhavaraj Ayyavu, Bharathi Pilar, and Ramakrishnan A. G. 2022a. Knowledge-driven Subword Grammar Modeling for Automatic Speech Recognition in Tamil and Kannada. *arXiv preprint*. ArXiv:2207.13333 [eess.AS].

Madhavaraj Ayyavu, Bharathi Pilar, and Ramakrishnan A. G. 2022b. Subword Dictionary Learning and Segmentation Techniques for Automatic Speech Recognition in Tamil and Kannada. *arXiv preprint*. ArXiv:2207.13331 [eess.AS].

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. pages 2278–2282.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

Martijn Bentum, Louis ten Bosch, and Tomas O. Lentz. 2025. Word stress in self-supervised speech models: A cross-linguistic comparison. pages 251–255.

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. ISSN: 2472-7695.

Won Ik Cho, Seok Min Kim, Hyunchang Cho, and Nam Soo Kim. 2021. kosp2e: Korean Speech to English Translation Corpus. pages 3705–3709.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. pages 2426–2430.

Patrick Cormac English, John D. Kelleher, and Julie Carson-Berndsen. 2022. Domain-Informed Probing of wav2vec 2.0 Embeddings for Phonetic Features. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 83–91, Seattle, Washington. Association for Computational Linguistics.

Marianne de Heer Kloots, Hosein Mohebbi, Charlotte Pouw, Gaofei Shen, Willem Zuidema, and Martijn

Bentum. 2025. What do self-supervised speech models know about Dutch? Analyzing advantages of language-specific pre-training. pages 256–260.

Anton de la Fuente and Dan Jurafsky. 2024. A layer-wise analysis of Mandarin and English suprasegmentals in SSL speech models. pages 1290–1294.

Ewan Dunbar, Mathieu Bernard, Nicolas Hamilakis, Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Eugene Kharitonov, and Emmanuel Dupoux. 2021. The Zero Resource Speech Challenge 2021: Spoken Language Modelling. pages 1574–1578.

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 323–330.

Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. 2021. Exploring wav2vec 2.0 on Speaker Verification and Language Identification. pages 1509–1513.

Marco Fonseca. 2023. THE PERCEPTION OF JAPANESE LEXICAL CONTRASTS BY L2 LEARNERS: THE ROLE OF THE LEXICON. *Proceedings of the 20th International Congress of Phonetic Sciences*, pages 2606–2611.

Mark J. F. Gales, Kate M. Knill, Anton Ragni, and Shakti P. Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED. pages 16–23.

Parismita Gogoi, Sishir Kalita, Wendy Lalhminghlui, Viyazonuo Terhiija, Moakala Tzudir, Priyankoo Sarmah, and S. R. M. Prasanna. 2025. Tone recognition in low-resource languages of North-East India: peeling the layers of SSL-based speech models. pages 4173–4177.

Jung-Woo Ha, Kihyun Nam, Jingu Kang, Sang-Woo Lee, Sohee Yang, Hyunhoon Jung, Hyeji Kim, Eunmi Kim, Soojin Kim, Hyun Ah Kim, Kyoungtae Doh, Chan Kyu Lee, Nako Sung, and Sunghun Kim. 2020. ClovaCall: Korean Goal-Oriented Dialog Speech Corpus for Automatic Speech Recognition of Contact Centers. pages 409–413.

Linyang He, Qiaolin Wang, Xilin Jiang, and Nima Mesgarani. 2025. Layer-wise Minimal Pair Probing Reveals Contextual Grammatical-Conceptual Hierarchy in Speech Representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35338–35353, Suzhou, China. Association for Computational Linguistics.

Staffan Hedström, Ragnheiður Þórhallsdóttir, David Erik Mollberg, Smári Freyr Guðmundsson, Ólafur Helgi Jónsson, Sunneva Þorsteinsdóttir, Judy Y. Fong, Eydís Huld Magnúsdóttir, and Jon Gudnason. 2022. Samrómur Unverified 22.07. Reykjavik University: Language and Voice Lab.

John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Robin Huo and Ewan Dunbar. 2025. Iterative Refinement, Not Training Objective, Makes HuBERT Behave Differently from wav2vec 2.0. pages 261–265.

Katsuya Iida. 2022. Kokoro Speech Dataset. `https://github.com/kaiidams/Kokoro-Speech-Dataset`.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-Light: A Benchmark for ASR with Limited or No Supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. ISSN: 2379-190X.

Goh Kawai and Carlos Toshinori Ishi. 1999. A system for learning the pronunciation of Japanese pitch accent. pages 177–182.

Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. Crepe: A Convolutional Representation for Pitch Estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165. ISSN: 2379-190X.

Rostislav Kolobov, Olga Okhapkina, Olga Omelchishina, Andrey Platunov, Roman Bedyakin, Vyacheslav Moshkin, Dmitry Menshikov, and Nikolay Mikhaylovskiy. 2021. MediaSpeech: Multilanguage ASR Benchmark and Dataset. *arXiv preprint*. ArXiv:2103.16193 [eess.AS].

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability. *http://www.asc.upenn.edu/Krippendorff/*.

Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Nigel G. Ward. 2023. On the Utility of Self-Supervised Models for Prosody-Related Tasks. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1104–1111.

Danni Ma, Neville Ryant, and Mark Liberman. 2021. Probing Acoustic Representations for Phonetic Properties. In *ICASSP 2021 - 2021 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 311–315. ISSN: 2379-190X.

Kikuo Maekawa. 2003. Corpus of spontaneous Japanese: its design and evaluation. page paper MMO2.

Kinan Martin, Jon Gauthier, Canaan Breiss, and Roger Levy. 2023. Probing Self-supervised Speech Models for Phonetic and Phonemic Information: A Case Study in Aspiration. pages 251–255.

Josh Meyer, David Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack, Julian Weber, Salomon Kabongo Kabenamualu, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Chinenye Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Olanrewaju Samuel, Jesujoba Alabi, and Shamsuddeen Hassan Muhammad. 2022. BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus. pages 2383–2387.

Juliette Millet and Ewan Dunbar. 2020. Perceptimatic: A Human Speech Perception Benchmark for Unsupervised Subword Modelling. pages 4881–4885.

Juliette Millet and Ewan Dunbar. 2022. Do self-supervised speech models develop human-like perception biases? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7591–7605, Dublin, Ireland. Association for Computational Linguistics.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-Supervised Speech Representation Learning: A Review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Ryoko Mugitani, Ferran Pons, Laurel Fais, Christiane Dietrich, Janet F. Werker, and Shigeaki Amano. 2009. Perception of vowel length by Japanese- and English-learning infants. *Developmental Psychology*, 45(1):236–247.

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 2020. The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *arXiv preprint*. ArXiv:2011.11588 [cs].

Daniel van Niekerk, Charl van Heerden, Marelie Davel, Neil Kleynhans, Oddur Kjartansson, Martin Jansche, and Linne Ha. 2017. Rapid Development of TTS Corpora for Four South African Languages. pages 2178–2182.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. ISSN: 2379-190X.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What Do Self-Supervised Speech Models Know About Words? *Transactions of the Association for Computational Linguistics*, 12:372–391.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-Wise Analysis of a Self-Supervised Speech Representation Model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative Layer-Wise Analysis of Self-Supervised Speech Models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. ISSN: 2379-190X.

Maxime Poli, Thomas Schatz, Emmanuel Dupoux, and Marvin Lavechin. 2024. Modeling the initial state of early phonetic learning in infant.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. pages 2757–2761.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Honghao Ren. 2022. Perception of Japanese singleton and geminate contrasts: A case of Chinese learners with different dialectal backgrounds. *Frontiers in Psychology*, 13:1070107.

Aisikaer Rouzi, Yin Shi, Zhang Zhiyong, Wang Dong, Askar Hamdulla, and ZHENG Fang. 2017. THUYG-20: A free Uyghur speech database. *Journal of Tsinghua University (Science and Technology)*, 57(2):182–187.

Ramon Sanabria, Hao Tang, and Sharon Goldwater. 2023. Analyzing Acoustic Word Embeddings from Pre-Trained Self-Supervised Speech Models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. ISSN: 2379-190X.

Yuta Sasaki. 2024a. japanese-wav2vec2-base.

Yuta Sasaki. 2024b. japanese-wav2vec2-large.

Yuta Sasaki. 2025. japanese-hubert-base-k2.

Thomas Schatz and Naomi H. Feldman. 2018. Neural network vs. HMM speech recognition systems as models of human cross-linguistic phonetic perception. *Proceedings of the Conference on Cognitive Computational Neuroscience*.

Thomas Schatz, Naomi H. Feldman, Sharon Goldwater, Xuan-Nga Cao, and Emmanuel Dupoux. 2021. Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences of the United States of America*, 118(7):e2001844118.

Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. 2013. Evaluating speech features with the minimal-pair ABX task: analysis of the classical MFC/PLP pipeline. pages 1781–1785.

Tanja Schultz. 2002. Globalphone: a multilingual speech and text database developed at karlsruhe university. pages 345–348.

Gaofei Shen, Michaela Watkins, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała. 2024. Encoding of lexical tone in self-supervised models of spoken language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4250–4261, Mexico City, Mexico. Association for Computational Linguistics.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2021. AISHELL-3: A Multi-Speaker Mandarin TTS Corpus. pages 2756–2760.

Greg Short, Keikichi Hirose, and Nobuaki Minematsu. 2011. Rule-based method for pitch level classification for a Japanese pitch accent CALL system. pages 45–48.

Keshan Sodimana, Pasindu De Silva, Supheakmungkol Sarin, Oddur Kjartansson, Martin Jansche, Knot Pipatsrisawat, and Linne Ha. 2018. A Step-by-Step Process for Building TTS Voices Using Open Source Data and Frameworks for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese. pages 66–70.

Haitong Sun, Stephen McIntosh, Kwanghee Choi, Eunjung Yeo, Daisuke Saito, and Nobuaki Minematsu. 2026. Prosodic ABX: A Language-Agnostic Method for Measuring Prosodic Contrast in Speech Representations.

Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari. 2019. JVS corpus: free Japanese multi-speaker voice corpus. *arXiv preprint*. ArXiv:1908.06248 [cs].

T Toda. 1998. Perceptual categorization of the durational contrasts by japanese learners. *Studies in language and literature*, 33:65–82.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy Liu, Cheng-I Lai, Jiatong Shi, Xuankai Chang, Phil Hall, Hsuan-Jui Chen, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2022. SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8479–8492, Dublin, Ireland. Association for Computational Linguistics.

Jörgen Valk and Tanel Alumäe. 2021. VOXLINGUA107: A Dataset for Spoken Language Recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658.

Maarten Versteegh, Roland Thiollière, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. 2015. The zero resource speech challenge 2015. pages 3169–3173.

Elena Voita and Ivan Titov. 2020. Information-Theoretic Probing with Minimum Description Length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Dong Wang and Xuewei Zhang. 2015. THCHS-30 : A Free Chinese Speech Corpus. *arXiv preprint*. ArXiv:1512.01882 [cs.CL].

Mu Yang, Ram C. M. C. Shekar, Okim Kang, and John H. L. Hansen. 2023. What Can an Accent Identifier Learn? Probing Phonetic and Prosodic Information in a Wav2vec2-based Accent Identification Model. pages 1923–1927.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I. Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. pages 1194–1198.

Yifan Yang, Jianheng Zhuo, Zengrui Jin, Ziyang Ma, Xiaoyu Yang, Zengwei Yao, Liyong Guo, Wei Kang, Fangjun Kuang, Long Lin, Daniel Povey, and Xie

Chen. 2025. k2SSL: A Faster and Better Framework for Self-Supervised Speech Representation Learning. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. ISSN: 1945-788X.

Yue Yin, Daijiro Mori, and Seiji Fujimoto. 2023. Reazonspeech: A free and massive corpus for japanese asr.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mHuBERT-147: A Compact Multilingual HuBERT Model. pages 3939–3943.

A Accent Label Correction

We corrected automatically generated transcriptions with a manually crafted substitution dictionary. For each sentence, we aligned the predicted transcription with the gold one and consulted the dictionary to determine whether the predicted transcription could be converted to the gold one in within each accent phrase. The dictionary mainly covers three types of systematic errors produced by the model: (1) single-mora confusions, (2) decompositions of a single mora into two separate morae, and (3) compressions of two separate morae into a single mora. Note that we only take into account errors that actually occurred.

(1) Single-Mora Confusion

This occurs where the model wrongly transcribes one-mora part of speech as another mora. In this case, the predicted accent label can be aligned directly with the gold mora.

(2) Single-Mora Decomposition

This occurs where the model wrongly transcribes one-mora part of speech as a two-mora sequence. When the two morae are labeled as LH, we set the accent label for the gold mora as L, and when LL or HH, we set it as L or H, respectively.

(3) Two-Mora Compression

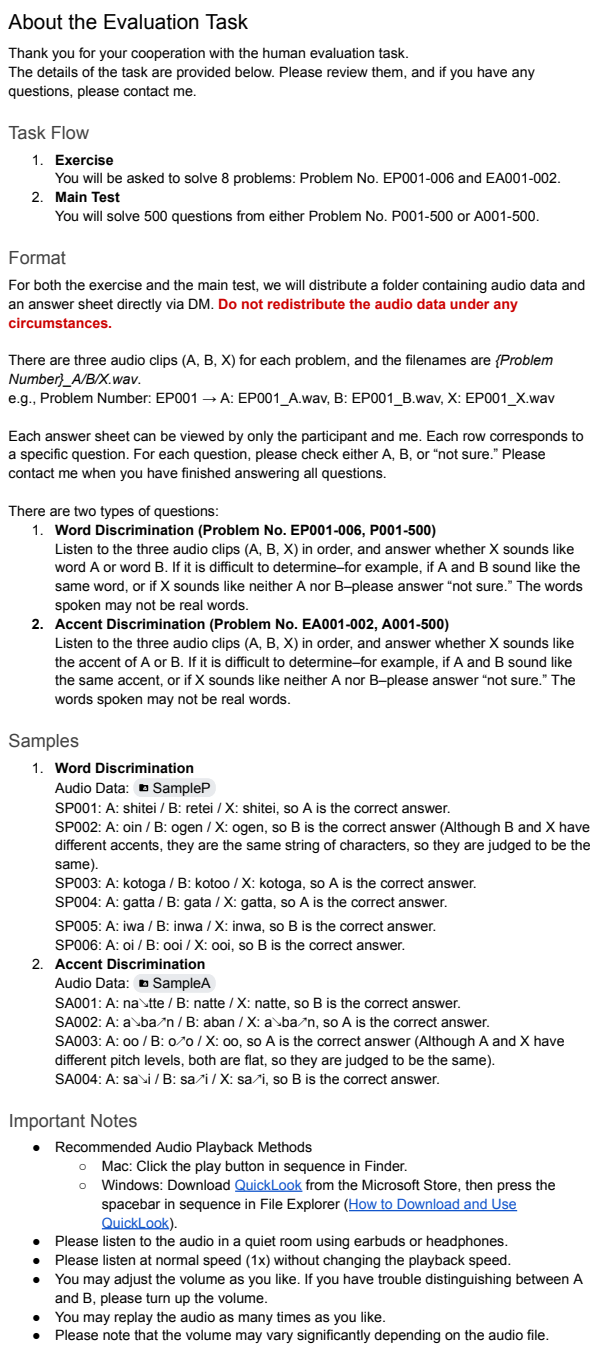
Thisoccurs where the model wrongly transcribes two-mora part of speech as another mora. We copy the predicted accent label for each of the gold two morae.

Table 6: Examples of predicted and gold transcript alignment. Pred = Predicted.

Type	Pred Transcript with Pred Accent Label		Gold transcript	Gold transcript with Pred Accent Label	
(1)	/u/	L/H	/we/	/we/	L/H
(2)	/ki yu/ /ki yu/	LH LL/HH	/kyu/ /kyu/	/kyu/ /kyu/	L L/H
(3)	/wa/	L/H	/o a/	/o a/	LL/HH

B Human Evaluation Details

We recruited six native Japanese students researching either natural language processing or speech processing. When recruiting, we informed them that the collected data would be used for research on Japanese speech. The participants were paid 1,500 Japanese yen per hour, which was determined based on the rules in our affiliation. We show the instructions presented to the participants in Figure 8.



About the Evaluation Task

Thank you for your cooperation with the human evaluation task.
The details of the task are provided below. Please review them, and if you have any questions, please contact me.

Task Flow

1. **Exercise**
   You will be asked to solve 8 problems: Problem No. EP001-006 and EA001-002.
2. **Main Test**
   You will solve 500 questions from either Problem No. P001-500 or A001-500.

Format

For both the exercise and the main test, we will distribute a folder containing audio data and an answer sheet directly via DM. **Do not redistribute the audio data under any circumstances.**

There are three audio clips (A, B, X) for each problem, and the filenames are *{Problem Number}_A/B/X.wav*.
e.g., Problem Number: EP001 → A: EP001_A.wav, B: EP001_B.wav, X: EP001_X.wav

Each answer sheet can be viewed by only the participant and me. Each row corresponds to a specific question. For each question, please check either A, B, or "not sure." Please contact me when you have finished answering all questions.

There are two types of questions:

1. **Word Discrimination (Problem No. EP001-006, P001-500)**
   Listen to the three audio clips (A, B, X) in order, and answer whether X sounds like word A or word B. If it is difficult to determine–for example, if A and B sound like the same word, or if X sounds like neither A nor B–please answer "not sure." The words spoken may not be real words.
2. **Accent Discrimination (Problem No. EA001-002, A001-500)**
   Listen to the three audio clips (A, B, X) in order, and answer whether X sounds like the accent of A or B. If it is difficult to determine–for example, if A and B sound like the same accent, or if X sounds like neither A nor B–please answer "not sure." The words spoken may not be real words.

Samples

1. **Word Discrimination**
   Audio Data: SampleP
   SP001: A: shitei / B: retei / X: shitei, so A is the correct answer.
   SP002: A: oin / B: ogen / X: ogen, so B is the correct answer (Although B and X have different accents, they are the same string of characters, so they are judged to be the same).
   SP003: A: kotoga / B: kotoo / X: kotoga, so A is the correct answer.
   SP004: A: gatta / B: gata / X: gatta, so A is the correct answer.
   SP005: A: iwa / B: inwa / X: inwa, so B is the correct answer.
   SP006: A: oi / B: ooi / X: ooi, so B is the correct answer.
2. **Accent Discrimination**
   Audio Data: SampleA
   SA001: A: na↘tte / B: natte / X: natte, so B is the correct answer.
   SA002: A: a↘ba↗n / B: aban / X: a↘ba↗n, so A is the correct answer.
   SA003: A: oo / B: o↗o / X: oo, so A is the correct answer (Although A and X have different pitch levels, both are flat, so they are judged to be the same).
   SA004: A: sa↘i / B: sa↗i / X: sa↗i, so B is the correct answer.

Important Notes

- Recommended Audio Playback Methods
  - Mac: Click the play button in sequence in Finder.
  - Windows: Download QuickLook from the Microsoft Store, then press the spacebar in sequence in File Explorer (How to Download and Use QuickLook).
- Please listen to the audio in a quiet room using earbuds or headphones.
- Please listen at normal speed (1x) without changing the playback speed.
- You may adjust the volume as you like. If you have trouble distinguishing between A and B, please turn up the volume.
- You may replay the audio as many times as you like.
- Please note that the volume may vary significantly depending on the audio file.

Figure 8: The instruction document for human evaluation participants.

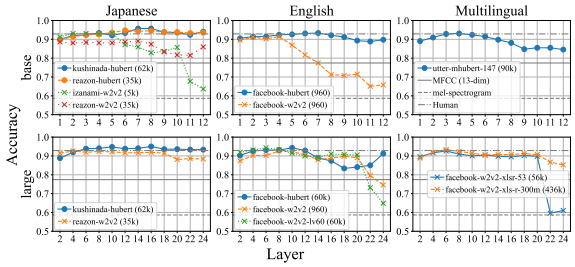


Figure 9: Layer-wise ABX accuracy for mora discrimination in the without-context (n=421) setting.

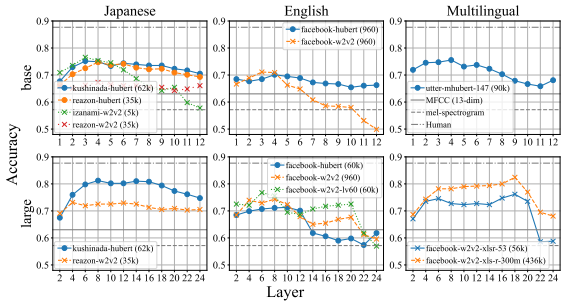


Figure 10: Layer-wise ABX accuracy for accent discrimination in the without-context (n=495) settings.

C Evaluated Models

We show the details of the S3Ms used in model evaluation in Table 7. Corpora for training multilingual models include multilingual corpora such as BABEL (17 Asian and African languages), CV (11 languages for facebook/wav2vec2-large-xlsr-53, 60 languages for facebook/wav2vec2-xls-r-300m, and 100 languages for utter-project/mHuBERT-147), MLS (eight European languages), VL (107 languages), VP (23 European languages), IISc-MILE (two Indian languages), B-TTS (ten Sub-Saharan African languages), and G-TTS (four South African languages and six low-resource languages), and monolingual corpora such as Japanese ones (JVS and Kokoro), Chinese ones (Aishells, and THCHS-30), Korean ones (ClovaCall and Kosp2e), an Icelandic one (Samrómur), an Uyghur one (THUYG-20), and a Turkish one (MS).

D Additional Analysis

D.1 Overall Results in the Without-Context Setting

Figure 9 and Figure 10 show overall layer-wise mora and accent discrimination results. Comparing the with- and without- context settings, no consistent difference in peak accuracy is observed for mora discrimination. For accent discrimination, however, most models achieve lower peak accuracy in the without-context setting. This suggests that pitch accent contrasts are more easily distinguished when S3Ms are able to use information from surrounding context, such as relative pitch height.

In both tasks, the peak layer tends to appear earlier in the without-context condition than in the with-context condition. We leave investigation of the reason for this for future work.

D.2 Model-Wise Results

Figure 11 shows full comparison results between the six contrast types in mora discrimination, and Figure 12 shows full comparison results between word-accent and nonword-accent contrasts in accent discrimination.

Figure 13 compares the layer-wise accuracy of two types of contrasts: those involving one phoneme difference (e.g., san**s**o vs. san**d**o) and those involving two phoneme differences (e.g., **ra**ndo vs. **se**ndo). We only use contrasts where A and B have the same number of phonemes here. The results indicate that those with two phoneme differences are better discriminated by representations of most model layers, probably due to longer difference in sound stimuli. This trend is also seen in human accuracy.

Figure 14 and Table 8 shows model-wise comparison results between within and across conditions in mora discrimination. Overall, S3M representations do not show a consistent benefit for any of the within conditions: within-speaker, within-gender, and within-accent accuracy are not consistently higher than their across counterparts. Humans seem to be sensitive to speaker identity, but show no consistent pattern for gender or accent conditions. Note that some of these results should be interpreted with caution, given the small number of items.

Figure 15 and Table 8 shows model-wise comparison results between within and across conditions in accent discrimination. As in mora discrimination, S3M representations do not consistently show higher accuracy in the within conditions than in the across conditions. Humans achieve higher accuracy in the across-speaker condition than the within-speaker condition, which is counter-intuitive, but gender has little effect on their accuracy.

Table 7: Details of the models used for evaluation. All the models are taken from HuggingFace. Kushinada, reazon-research, and facebook models are licensed under Apache 2.0, imprt/izanami-wav2vec2-base under imprt-license, and utter-project/mHuBERT-147 under CC-BY-NC-SA-4.0. Lang. = Language. Aishells = AISHELL-1 (Bu et al., 2017) and AISHELL-3 (Shi et al., 2021), B-TTS = BibleTTS (Meyer et al., 2022), CV = CommonVoice (Ardila et al., 2020), MLS = Multilingual LibriSpeech (Pratap et al., 2020), MS = MediaSpeech (Kolobov et al., 2021), TH-data = THCHS-30 (Wang and Zhang, 2015) and THUYG-20 (Rouzi et al., 2017), VL = VoxLingua107 (Valk and Alumäe, 2021), VP = VoxPopuli (Wang et al., 2021).

Lang.	Model	Model Size	Training Loss	#Params	Corpus Size (h)	Training Corpus
JA	imprt/kushinada-hubert-base (Hsu et al., 2021)	base	Predictive	90M	62k	Japanese TV broadcast audio
	imprt/kushinada-hubert-large (Hsu et al., 2021)	large	Predictive	300M		
	imprt/izanami-wav2vec2-base (Baevski et al., 2020)	base	Contrastive	95M	5k	
	reazon-research/japanese-hubert-base-k2 (Sasaki, 2025; Yang et al., 2025)	base	Predictive	94.4M	35k	ReazonSpeech v2 (Yin et al., 2023)
	reazon-research/japanese-wav2vec2-base (Sasaki, 2024a)	base	Contrastive	94.4M		
	reazon-research/japanese-wav2vec2-large (Sasaki, 2024b)	large	Contrastive	300M		
EN	facebook/hubert-base-ls960 (Hsu et al., 2021)	base	Predictive	90M	960	LibriSpeech
	facebook/hubert-large-ll60k (Hsu et al., 2021)	large	Predictive	317M	60k	Libri-Light
	facebook/wav2vec2-base (Baevski et al., 2020)	base	Contrastive	95M	960	LibriSpeech
	facebook/wav2vec2-large (Baevski et al., 2020)	large	Contrastive	317M		
	facebook/wav2vec2-large-lv60 (Baevski et al., 2020)	large	Contrastive	317M	60k	Libri-Light
Multi	facebook/wav2vec2-large-xlsr-53 (Conneau et al., 2021)	large	Contrastive	317M	56k	BABEL (Gales et al., 2014), CV, MLS
	facebook/wav2vec2-xls-r-300m (Babu et al., 2022)	large	Contrastive	300M	436k	BABEL, CV, MLS, VL, VP
	utter-project/mHuBERT-147 (Zanon Boito et al., 2024)	base	Predictive	94.4M	90k	Aishells, B-TTS, ClovaCall (Ha et al., 2020), CV, G-TTS (Niekerk et al., 2017; Sodimana et al., 2018), IISc-MILE (Ayyavu et al., 2022a,b), JVS, Kokoro (Iida, 2022), kosp2e (Cho et al., 2021), MLS, MS, Samrómur (Hedström et al., 2022), TH-data, VL, VP

Table 8: Human accuracy of within- and across- speaker, gender, and accent contrasts. The letters 's', 'g', and 'a' denote speaker, gender, and accent, respectively.

Test Type	Condition	#Test Items	Acc.
Mora Discrimination	within-s-g-a	14	1.000
	within-s-g-across-a	6	1.000
	within-g-a-across-s	133	0.927
	within-g-across-s-a	76	0.930
	within-a-across-s-g	125	0.936
	across-s-g-a	67	0.896
Accent Discrimination	within-s-g	24	0.833
	within-g-across-s	220	0.882
	across-s-g	251	0.876

E JPhonABX License

We will release the code for creating the JPhonABX benchmark upon acceptance. The benchmark itself will not be released, in accordance with the license of the JVS corpus.

The audio data in JPhonABX is derived from the JVS corpus and is therefore restricted to the following uses: research by academic institutions, non-commercial research (including research conducted within commercial organizations), and personal use (including blog posts) (Takamichi et al., 2019). Regarding the text and annotation data, the text data follows the original license of the JVS text corpus[4], and the tags are licensed under CC BY-SA 4.0. The text and tags in JPhonABX will be distributed under the same terms.

[4] https://sites.google.com/site/shinnosuketakamichi/publication/jsut?authuser=0#h.p_YexdnC0CAtY8

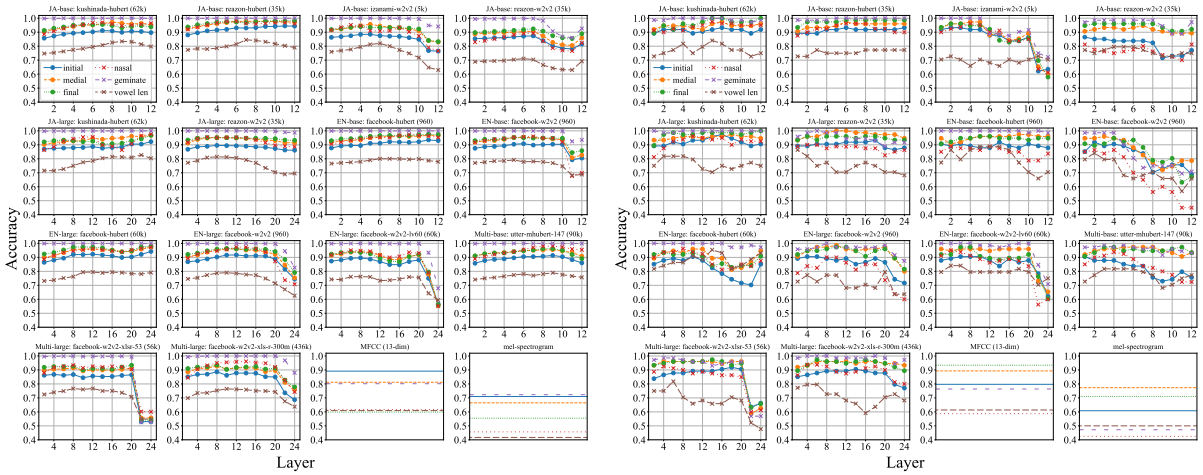


Figure 11: Layer-wise ABX accuracy of each contrast type for mora discrimination in the (left) with-context (n=192,000) and (right) without-context (n=421) settings.

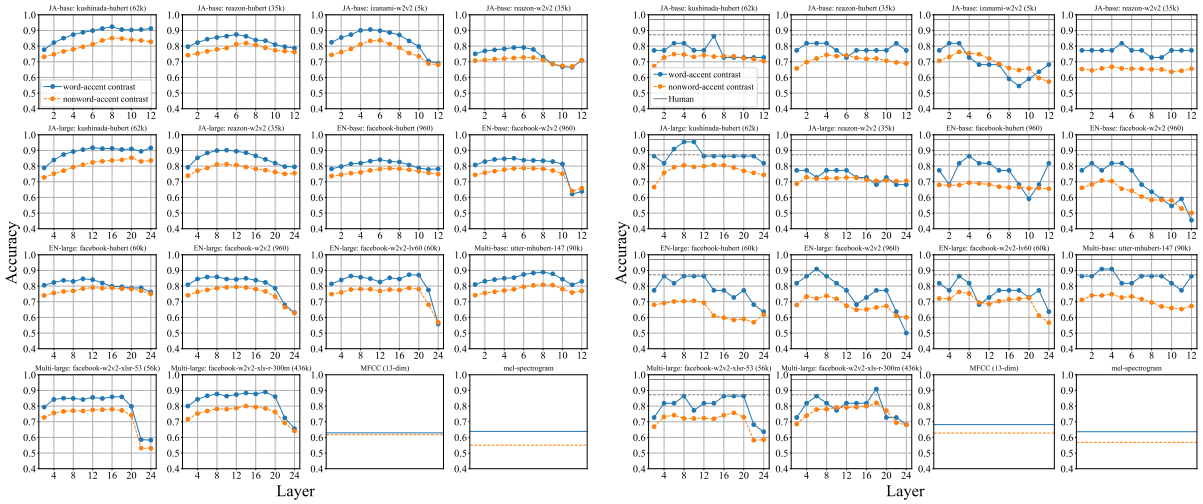


Figure 12: Layer-wise ABX accuracy comparison for accent discrimination between word- and nonword-accent contrasts. Word-accent contrasts are those between two pitch accent patterns that are both possible as word accent patterns in Standard Japanese (i.e., HL, LH, HLL, LHH, LHL), and nonword-accent contrasts are those involving at least one impossible pattern. The left panel shows results in the with-context (n=4,337 (word) + 93,663 (nonword)) setting and the right panel shows results in the without-context (n=22 (word) + 473 (nonword)) setting.

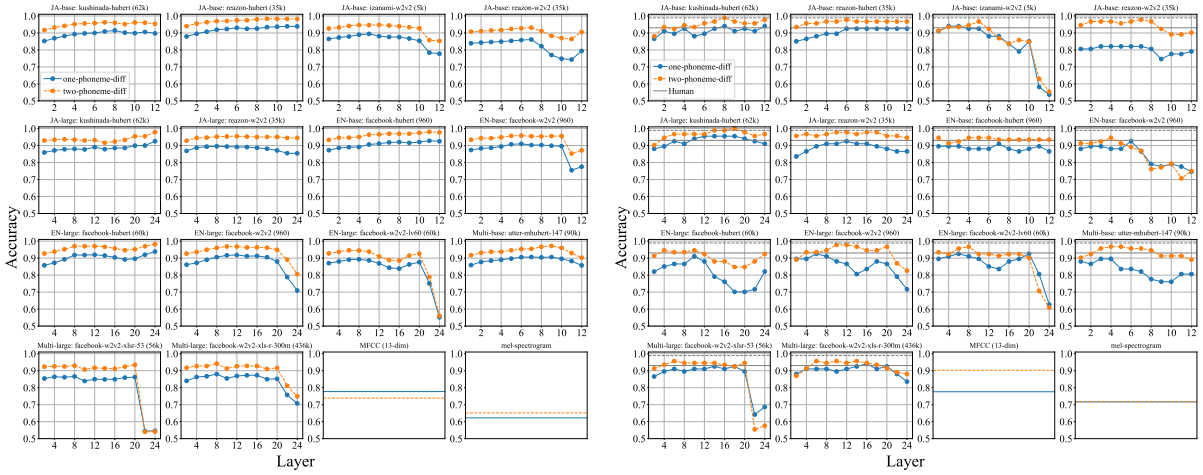


Figure 13: Layer-wise ABX accuracy comparison for mora discrimination between contrasts involving two phoneme differences and those involving a single phoneme difference. The left panel shows results in the with-context (n=29,226 (one-phoneme-diff) + 35,031 (two-phoneme-diff)) setting and the right panel shows results in the without-context (n=67 (one-phoneme-diff) + 92 (two-phoneme-diff)) setting.

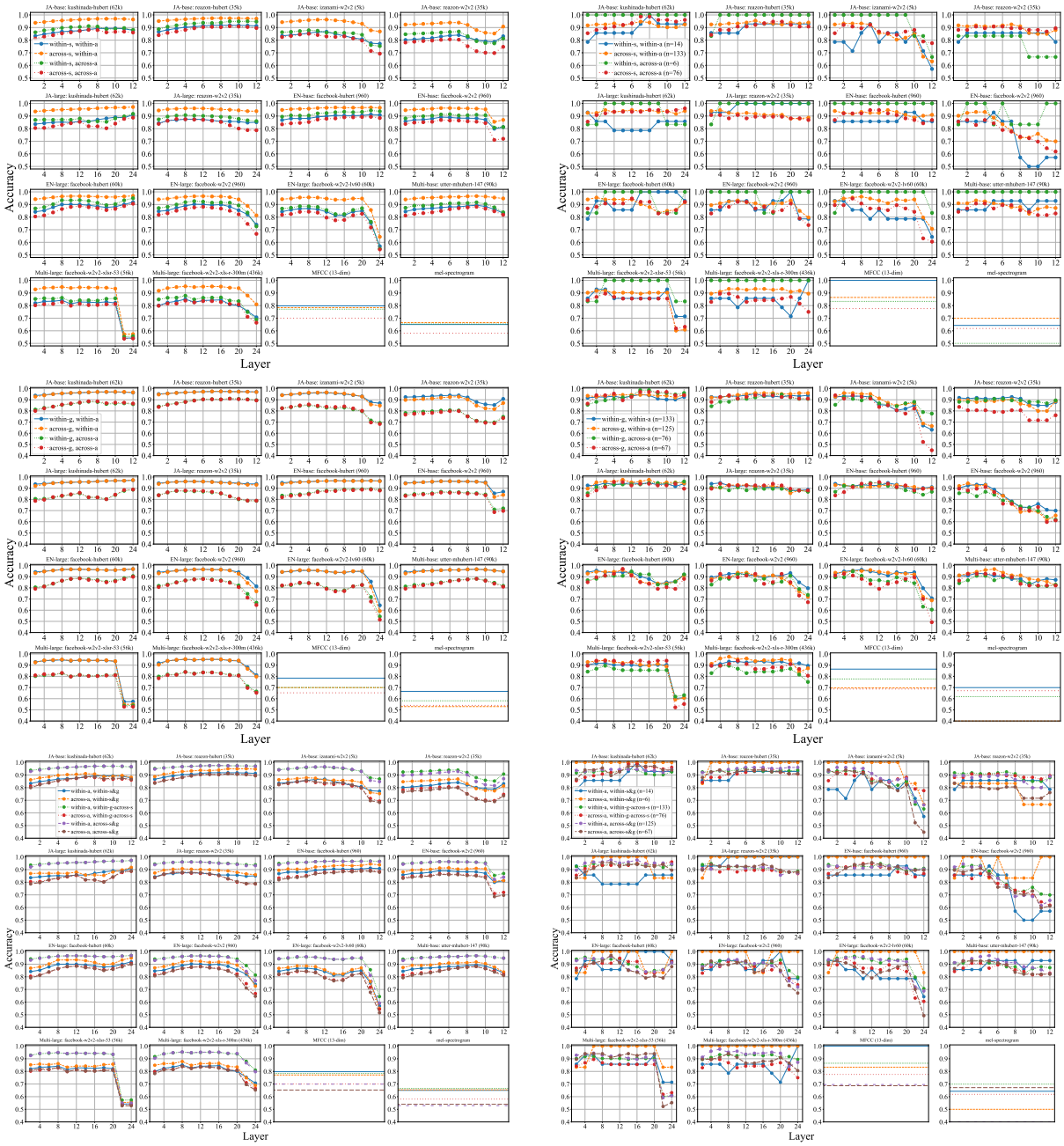


Figure 14: Layer-wise ABX accuracy comparison for mora discrimination between within- and across- (top) speaker, (middle) gender, and (bottom) accent conditions in the (left) with-context (n=192,000) and (right) without-context (n=421) settings. The letters “s”, “g”, and “a” denote speaker, gender, and accent, respectively.

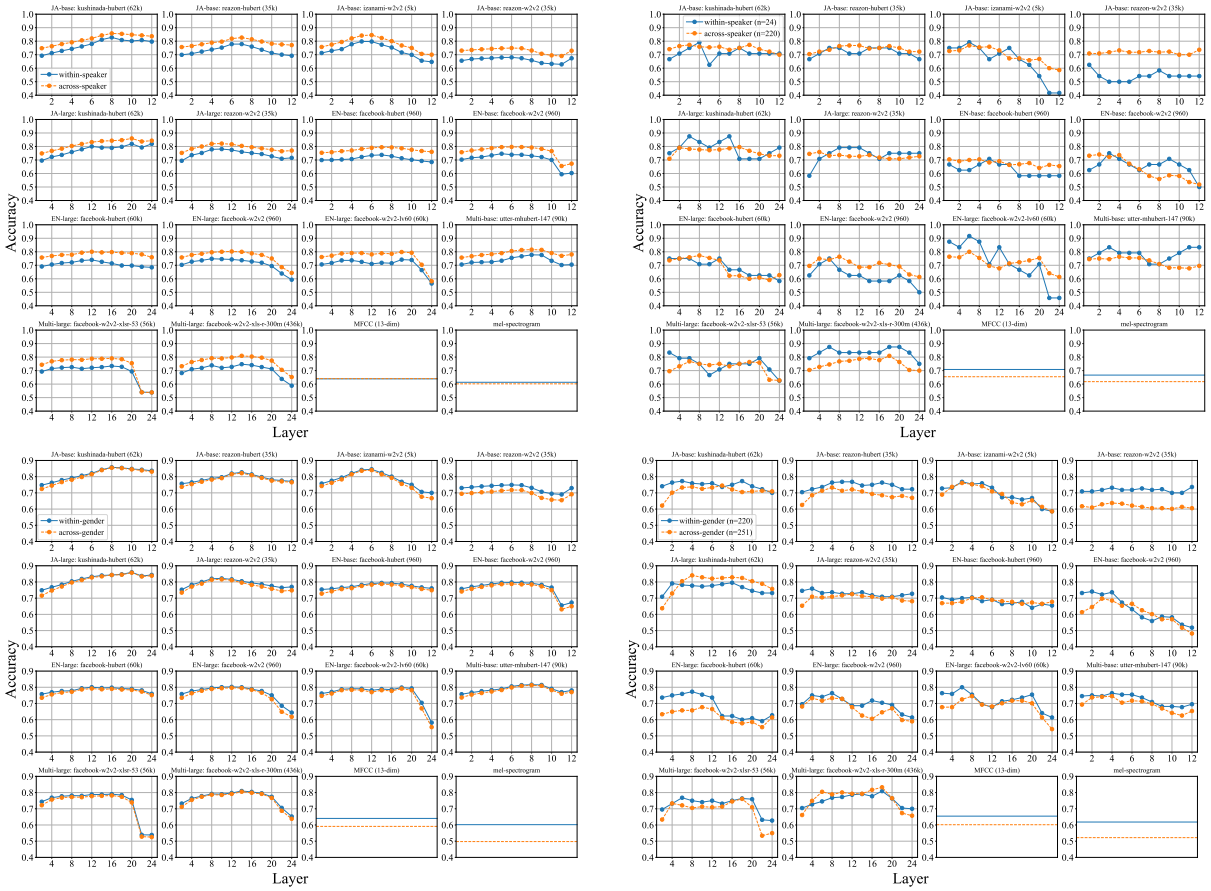


Figure 15: Layer-wise ABX accuracy comparison for accent discrimination between within- and across- (top) speaker and (bottom) gender conditions in the (left) with-context (n=98,000) and (right) without-context (n=495) settings.