Self-supervised speech representations display *some* human-like cross-linguistic perceptual abilities

Joselyn Rodriguez[1], Kamala Sreepada[1],
Ruolan Leslie Famularo[1], Sharon Goldwater [2], Naomi H. Feldman[1]
[1] University of Maryland
[2] University of Edinburgh
Correspondence: jrodri20@umd.edu

Abstract

State of the art models in automatic speech recognition have shown remarkable improvements due to modern self-supervised (SSL) transformer-based architectures such as wav2vec 2.0 (Baevski et al., 2020). However, how these models encode phonetic information is still not well understood. We explore whether SSL speech models display a linguistic property that characterizes human speech perception: language specificity. We show that while wav2vec 2.0 displays an overall language specificity effect when tested on Hindi vs. English, it does not resemble human speech perception when tested on finer-grained differences in Hindi speech contrasts.

1 Introduction

Human listeners become attuned to the speech sounds of their native language already in their first year of life (Werker, 1995; Jusczyk, 2000; Kuhl et al., 2006). By adulthood, nonnative contrasts which they were once able to discriminate are no longer discriminable (Miyawaki et al., 1975; Cutler, 2000; Best and Tyler, 2007). Acquiring a second language as an adult can thus be marred by difficulty acquiring certain phonetic contrasts. This *language specificity* effect is a core property of human speech perception.

For human listeners, the difficulty of acquiring a particular non-native contrast depends both on whether the acoustic-phonetic dimension used to distinguish the contrasts is used in the native language and on the perceptual similarity of the non-native categories to native categories (see Best and Tyler, 2007). For example, while English only has two categories for the coronal stop series (/t/ and /d/), Hindi has eight (/ɖ/,/ɖʰ/,/ʈ/, /ʈʰ/,/t̪/,/t̪ʰ/,/d̪/,/d̪ʰ/). Because of the relationship between the acoustic dimensions used and the perceptual similarity to existing English categories, Hindi contrasts that differ along aspiration (/t/ vs. /tʰ/) are easier for English listeners to acquire than along place of articulation (dental vs. retroflex; /t/ vs. /ʈ/) or voicing (/t/ vs. /d/) (Werker et al., 1981; Tees and Werker, 1984; Pederson and Guion-Anderson, 2010; Hayes-Harb and Barrios, 2022).

Whether computational models of speech perception share certain properties with humans has been a subject of recent interest. Previous work has suggested that like humans, some speech models with non-transformer architectures display language specificity effects (Millet et al., 2019; Matusevych et al., 2020; Schatz et al., 2021).

However, less work has examined this effect in transformer architectures. Millet and Dunbar (2022) suggest that self-supervised speech transformers do not display a cross-linguistic difference in predicting human performance, but their measures aggregate across all contrasts, and given the complex relationship between native and non-native contrasts, this makes interpretation of the results difficult. In fact, previous work with non-transformer speech models found that for vowels, while the model displayed an overall language specificity effect, the direction of the effect was in the opposite direction than expected (Millet et al., 2019): a non-native model better predicted native speakers' discrimination. It is not known to what extent the specific perceptual similarity space in transformers is similar to humans.

In this paper, we test whether self-supervised transformer speech models (wav2vec 2.0) display an effect of language specificity. We do so by examining specific patterns of cross-linguistic differences, using Hindi contrasts as a case study. We explore a series of contrast that are known be be difficult for native English listeners. For these listeners, as noted above, place (/t/ vs. /ʈ/) and voicing (/t/ vs. /d/) are more difficult dimensions to discriminate along than aspiration (/t/ vs. /tʰ/) (Werker et al., 1981; Tees and Werker, 1984; Pederson and Guion-Anderson, 2010; Hayes-Harb and Barrios,

2022). These behavioral results provide a test case to explore targeted fine-grained categorization patterns of speech models to determine whether the models' representations are structured similarly to human listeners.

In Experiment 1, we find that wav2vec 2.0 displays an overall language specificity effect: a native-trained model performs better on native categorization task than a non-native model. In Experiment 2, we examine specific contrasts where second language learners are attested to struggle, and find that both the native and non-native model show high accuracy in categorization across the most difficult dimensions for humans – place and voicing. We additionally find that where human listeners have been shown to have the least difficulty overall, the models show the largest cross-linguistic difference in accuracy. This suggests that wav2vec 2.0 is encoding language specific information, but structured in ways that differ from human listeners.

2 Experiments

2.1 Models

The following experiments were performed on two models based on the wav2vec 2.0 architecture with 7 CNN encoder layers and 12 transformer layers (Baevski et al., 2020). Throughout this paper we display results from all 12 transformer layers, as previous work has shown that different layers may produce different results depending on the task (Pasad et al., 2021).

The first model we use is the English pre-trained wav2vec 2.0 base model available through the `fairseq` repository.[1] This model is pre-trained on approximately 1000 hours of English from the Librispeech Corpus of read English (this model is referred to as `wav2vec-english`).

The second is a Hindi pre-trained model available through the `Vakyansh` toolkit (Chadha et al., 2022).[2] The Hindi model is trained on 4200 hours of Hindi starting from the base fairseq wav2vec 2.0 model with continued pre-training (this model is referred to as `wav2vec-hindi`).

2.2 Data Preparation and Classifier Setup

For Hindi evaluation, we used the Hindi Common Voice corpus,[3] a crowd-sourced corpus of read speech. To acquire time-aligned phoneme transcriptions, we force-aligned the speech with the Montreal Forced Aligner (McAuliffe et al., 2017). We used the validated subset of the corpus which totaled 13 hours. For English evaluation, we used the Wall Street Journal corpus (Paul and Baker, 1992), a read corpus of English. To match the sizes of the corpora, we randomly sampled utterances until we reached 13 hours.

For each utterance in both Hindi and English, we extracted embeddings from each of the 12 transformer layers from both the Hindi-trained and English-trained models. For each embedding, we average over the frames composing a single phoneme according to the forced alignments. Thus, each phoneme is represented by a single embedding vector of size 768.

Given all the Hindi and English embedded phonemes, we additionally sub-sampled both datasets to get roughly an equal number of instances in each target category. We performed this step because the distribution of phonemes differs between English and Hindi, especially for the phonemes of interest. The number of individual tokens was determined by the smallest class of interest (N=67 for Hindi /ɖʰ/). Taking the entire set of embeddings and phoneme labels, we randomly sampled from the set of phoneme embeddings until the desired number of tokens was reached. In all experiments, each phone category contains at most 67 tokens.[4] This step was performed to ensure that any difference in classification accuracy was due to the learned representations, and not to a frequency effect in the classifier.

Classification was performed using `sklearns`'s Logistic Regression function with a multinomial loss to get a measure of overall phone multi-way classification accuracy across layers. The classifier is trained to predict the correct phone label from all possible labels (English=42 labels, Hindi=72 labels). In order to get a measure of standard error, we utilized 5-fold cross validation.

2.3 Experiment 1: Global Language Specificity

In the first experiment, we explored whether the models display an overall global language specificity effect by examining the cross-linguistic classification accuracy aggregated over all phonemes

[1] https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec

[2] https://github.com/Open-Speech-EkStep/vakyansh-models

[3] https://commonvoice.mozilla.org/en/datasets

[4] Some rare phonemes (e.g., /ʒ/) occurred fewer than 67 times but were not the focus of the current work

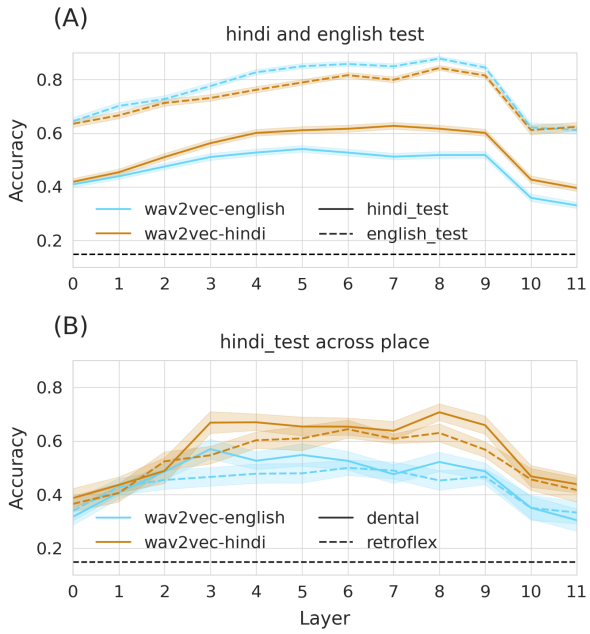

Figure 1: Wav2vec-hindi performs better than wav2vec-english on global multi-way classification for Hindi test (A) as well as for Hindi dental and retroflex sounds (B)

for the Hindi- and English-trained models' performance on both Hindi and English test data.

If the models display a native language specificity effect, we would expect that the aggregated classification accuracy across phonemes for the Hindi-trained model should be higher on Hindi test data than the English-trained model on Hindi test data and vice versa for English test data.

2.3.1 Results

Examining the overall classification accuracy in the best performing layer (layer=8), we find that the wav2vec-hindi has 10% higher categorization accuracy than wav2vec-english on Hindi test data. When tested on English, wav2vec-english outperforms wav2vec-hindi by 3% (Figure 1a). This suggests that the models do display a language specificity effect at a global level.

To determine whether the cross-linguistic differences are due to predicted difficulty in place of articulation for Hindi test data, we further examined the multi-way classification results averaged across only the set of Hindi dental and retroflex test phonemes (Figure 1b). As expected, the effect remains. The non-native wav2vec-english displays more difficulty with the dental and retroflex sounds in Hindi than the native model (wav2vec-hindi).

2.3.2 Discussion

In the global classification, we found an overall difference in wav2vec-english and wav2vec-hindi cross-linguistic classification accuracy collapsed across phonemes for Hindi and English test data. When we examined the aggregated classification accuracy of dental and retroflex sounds in the Hindi test data only, the effect remains – the Hindi model performs better on both dental and retroflex classification than the English model.

This suggests that through self-supervised training, wav2vec 2.0 is encoding language specific information. This has downstream consequences on phoneme encoding causing language-dependent patterns of categorization. However, the current experiment is limited to multi-way classification in which the model identifies the correct phoneme out of all possible labels (e.g., /d/ vs all other labels {/dʰ/, /tʰ/, /ɖ/, /b/, /p/,...}).

To determine whether the model is encoding information in a similar way to human listeners, it is of interest what the possible errors in this categorization are. For example, while the model makes errors in classification of dental or retroflex sounds, it is unknown whether the error is due to mistaking a dental sound as a retroflex sound or for some unrelated sound such as a vowel or fricative.

Therefore, in the following experiment, we examine finer-grained categorization performance limited to just the distinctions across dimensions of interest in the Hindi test data: place (dental or retroflex), voicing (voiced or unvoiced), and aspiration (aspirated or unaspirated). This task is also more directly comparable to the kind of perceptual tasks used with human listeners.

2.4 Experiment 2: Local Language Specificity

We simulate a two-alternative forced choice task where the model must categorize a sound into one of two categories while other features are kept constant. In a behavioral two-alternative forced choice task, listeners are given a sound, and asked to determine whether the sound belongs to category A or B. We simulate this by reducing the multi-way classification of Experiment 1 to a two-way classification task where the probability of a class y for a feature vector x_l from layer l is equal to

$$p(y = A) = \frac{exp(W_A^T x_l)}{exp(W_A^T x_l) + exp(W_B^T x_l)} \quad (1)$$

W_A refers to the classifier weights for class A and W_B refers to the weights for class B in the classifier

trained on representations from a given layer l.

2.4.1 Results

If the models are encoding language specific information during training, we would expect the English model to struggle in classification of the Hindi sounds relative to the Hindi model primarily along the dimensions of place (dental vs. retroflex), secondarily along voicing, and rarely along aspiration, as these are the relative difficulties experienced by human second language learners of Hindi whose native language is English (Werker et al., 1981; Tees and Werker, 1984; Pruitt et al., 2006; Hayes-Harb and Barrios, 2022). What we found instead is that **both** the English and the Hindi trained models perform well in correctly categorizing sounds as either dental or retroflex (Figure 2, top plot). Thus, despite the cross-linguistic difference for the global multi-way classification task from Experiment 1, this effect is no longer present when we compare between only dental and retroflex sounds, where human data would most predict it. Similarly, both the Hindi and English models perform well when tested on categorization along voicing (Figure 2, middle plot). While voicing seems to be marginally easier to categorize along, this holds for both the English and the Hindi-trained models. Therefore, unlike human English listeners who perform worse than native Hindi listeners when categorizing these sounds, the monolingual native English model and the Hindi-trained model perform well overall on the Hindi contrasts *regardless* of training language.

Further, when testing categorization accuracy along aspiration, where we expect the least amount of language specificity (i.e., best performance for the wav2vec-english), we find the opposite effect. In the best performing layer (layer=7), we find the largest cross-linguistic difference in which the the categorization accuracy for wav2vec-hindi is 25% higher than the wav2vec-english. Further examination of the pattern of classification errors in the multi-way classification from Experiment 1 shows that the confusions for both wav2vec-hindi and wav2vec-english were indeed primarily across aspiration and only secondarily across place Figure 3.

2.4.2 Discussion

In this experiment, we limited the task to a two-alternative forced choice in order to explore classification accuracy across specific phonetic dimensions. We found that neither the monolingual wav2vec-english nor wav2vec-hindi displayed any

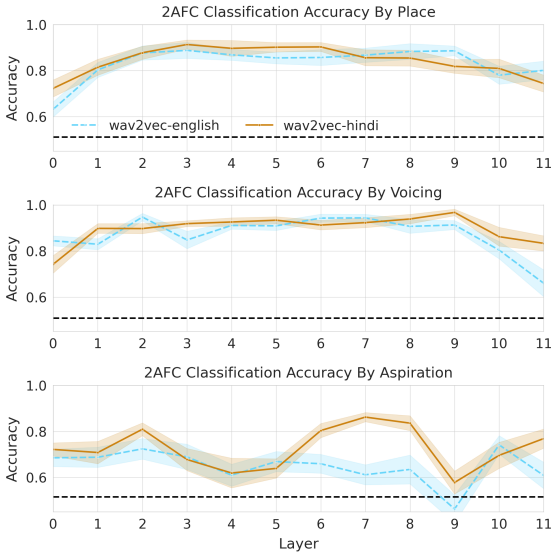

Figure 2: There is no difference in performance for the models on contrasts differing along place or voicing. Wav2vec-hindi outperforms wav2vec-english only on contrasts differing along aspiration.

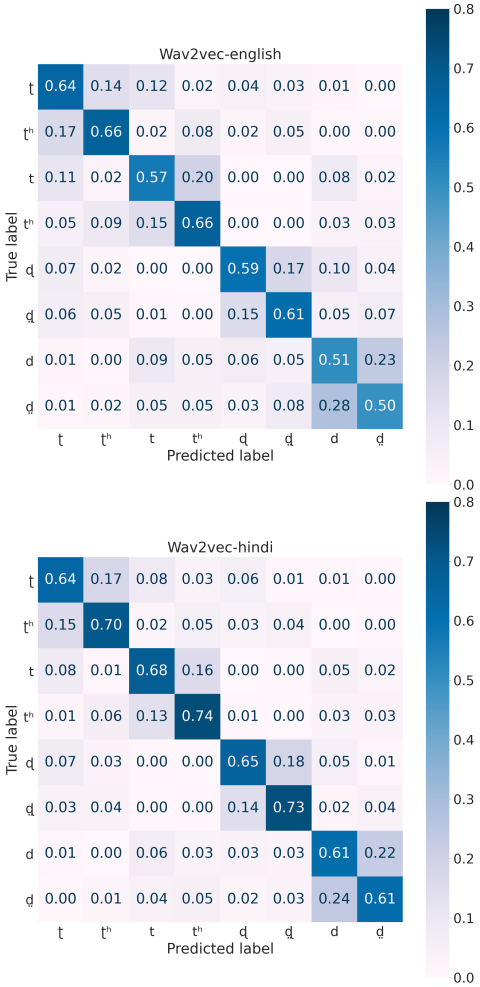

Figure 3: Confusion matrix of Hindi phoneme classification for English and Hindi trained models

difficulty in distinguishing between the Hindi contrasts of place (dental vs. retroflex) or voicing (voiced vs. unvoiced) in phoneme classification. We also found that for distinctions along aspiration (aspirated vs. unaspirasted), wav2vec-hindi outperforms wav2vec-english, displaying a fine-grained effect of language specificity. These results show that the effect of overall language specificity that was found in Experiment 1 was driven primarily by the wav2vec-english model's errors along aspiration when categorization the Hindi phonemes.

Native English listeners who are learning Hindi as a second language primarily struggle with place and voicing distinctions rather than aspiration (Werker et al., 1981; Tees and Werker, 1984; Pruitt et al., 2006; Pederson and Guion-Anderson, 2010; Hayes-Harb and Barrios, 2022). The difficulty in discrimination along place for native English listeners is thought to be due to issues in attending to the relevant acoustic cues differentiating the contrast (Flege and Bohn, 2021; Strange, 2011). The models' performance suggests they are not weighting the relevant cues to category identification in a way similar to humans. This is in line with recent work that has found that wav2vec 2.0 displays different weighting of dimensions than humans in noisy listening environments (Jurov, 2024).

While large speech models may have high performance on downstream speech recognition tasks, they are not learning speech representations in a way comparable to humans. The difference in the learned representations could be because the current pre-trained models are trained in a self-supervised manner without any information regarding category identity, unlike human learners who have knowledge the phonological structure of their native language. This could indicate that the necessary information for creating native-like cue-weighting patterns is guided by higher-level category knowledge that is not present in the current models. Of interest in future work is investigating this differential weighting of acoustic cues to better understand how the learned perceptual spaces differ between humans and speech models and how this may impact global and fine-grained cross-linguistic in categorization and discrimination performance.

3 Conclusion

In this work we explored both global and fine-grained cross-linguistic patterns of categorization in wav2vec 2.0. We found that models perform better overall at test on a language they have been trained on, displaying a global language specificity effect similar to humans. However, when we examined specific contrasts differing along certain phonetic features, the models pattern differently than humans. This result provides evidence of fundamental differences in the structure of representations learned by wav2vec 2.0 and human listeners.

4 Limitations

One limitation of the current work is the reliance on pre-trained models which limited the balance between amount and kind of training data for the wav2vec-hindi and wav2vec-english. Wav2vec-hindi was trained on a greater amount of data than wav2vec-english, but had been trained using the weights from the wav2vec BASE as the starting point for continued pre-training. Thus, the model may be better described as a bilingual Hindi-English model. The current work also displayed results from only wav2vec 2.0, leaving open the question of whether transformer models trained with a different objective would display the same patterns of language specificity.

5 Acknowledgements

We thank Bill Idsardi and the computational cognitive science group for helpful comments and discussion. This work was supported by NSF grant BCS-2120834.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.

Catherine T. Best and Michael D. Tyler. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In Ocke-Schwen Bohn and Murray J. Munro, editors, *Language Learning & Language Teaching*, volume 17, pages 13–34. John Benjamins Publishing Company, Amsterdam.

Harveen Singh Chadha, Anirudh Gupta, Priyanshi Shah, Neeraj Chhimwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan. 2022. Vakyansh: Asr toolkit for low resource indic languages. *Preprint*, arXiv:2203.16512.

Anne Cutler. 2000. Listening to a second language through the ears of a first. *Interpreting. International Journal of Research and Practice in Interpreting*, 5(1):1–23.

James Emil Flege and Ocke-Schwen Bohn. 2021. The Revised Speech Learning Model (SLM-r). In Ratree Wayland, editor, *Second Language Speech Learning*, 1 edition, pages 3–83. Cambridge University Press.

Rachel Hayes-Harb and Shannon Barrios. 2022. Native English speakers and Hindi consonants: From cross-language perception patterns to pronunciation teaching. *Foreign Language Annals*, 55(1):175–197.

Nika Jurov. 2024. *Modeling Adaptability Mechanisms of Speech Perception*. Ph.D. thesis, University of Maryland.

Peter W. Jusczyk. 2000. *The Discovery of Spoken Language*. The MIT Press.

Patricia K. Kuhl, Erica Stevens, Akiko Hayashi, Toshisada Deguchi, Shigeru Kiritani, and Paul Iverson. 2006. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, 9(2):F13–F21. Https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-7687.2006.00468.x.

Yevgen Matusevych, Thomas Schatz, Herman Kamper, Naomi H. Feldman, and Sharon Goldwater. 2020. Evaluating computational models of infant phonetic learning across languages. *arXiv:2008.02888 [cs, eess]*. ArXiv: 2008.02888.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Juliette Millet and Ewan Dunbar. 2022. Do self-supervised speech models develop human-like perception biases? *arXiv preprint*. ArXiv:2205.15819 [cs, eess].

Juliette Millet, Nika Jurov, and Ewan Dunbar. 2019. Comparing unsupervised speech learning directly to human performance in speech perception. In *CogSci 2019 - 41st Annual Meeting of Cognitive Science Society*, Montréal, Canada.

Kuniko Miyawaki, James J. Jenkins, Winifred Strange, Alvin M. Liberman, Robert Verbrugge, and Osamu Fujimura. 1975. An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, 18(5):331–340.

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.

Douglas B. Paul and Janet M. Baker. 1992. The design for the Wall Street Journal-based CSR corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Eric Pederson and Susan Guion-Anderson. 2010. Orienting attention during phonetic training facilitates learning. *The Journal of the Acoustical Society of America*, 127(2):EL54–EL59.

John S. Pruitt, James J. Jenkins, and Winifred Strange. 2006. Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *The Journal of the Acoustical Society of America*, 119(3):1684–1696. Publisher: Acoustical Society of America.

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*, 118(7):e2001844118.

Winifred Strange. 2011. Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics*, 39(4):456–466.

Richard C. Tees and Janet F. Werker. 1984. Perceptual flexibility: Maintenance or recovery of the ability to discriminate non-native speech sounds. *Canadian Journal of Psychology / Revue canadienne de psychologie*, 38(4):579–590.

Janet F. Werker. 1995. Exploring developmental changes in cross-language speech perception. In *An Invitation to Cognitive Science, Volume 1: Language*. The MIT Press. Https://direct.mit.edu/book/chapter-pdf/2306128/9780262273916_cad.pdf.

Janet F. Werker, John H. V. Gilbert, Keith Humphrey, and Richard C. Tees. 1981. Developmental Aspects of Cross-Language Speech Perception. *Child Development*, 52(1):349–355. Publisher: [Wiley, Society for Research in Child Development].