UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

What do self-supervised speech models know about Dutch? Analyzing advantages of language-specific pre-training

de Heer Kloots, M.; Mohebbi, Hosein; Pouw, C.; Shen, Gaofei; Zuidema, W.; Bentum, Martijn

DOI
10.48550/arXiv.2506.00981
10.21437/Interspeech.2025-1526

Publication date
2025

Document Version
Final published version

Published in
Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH

License
Other

Link to publication

Citation for published version (APA):
de Heer Kloots, M., Mohebbi, H., Pouw, C., Shen, G., Zuidema, W., & Bentum, M. (2025). What do self-supervised speech models know about Dutch? Analyzing advantages of language-specific pre-training. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *26*, 256-260. https://doi.org/10.48550/arXiv.2506.00981, https://doi.org/10.21437/Interspeech.2025-1526

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:09 Sept 2026

Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

What do self-supervised speech models know about Dutch? Analyzing advantages of language-specific pre-training

Marianne de Heer Kloots[1], *Hosein Mohebbi*[2], *Charlotte Pouw*[1], *Gaofei Shen*[2], *Willem Zuidema*[1], *Martijn Bentum*[3]

[1]Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
[2]Cognitive Science and Artificial Intelligence, Tilburg University, The Netherlands
[3]Centre for Language Studies, Radboud University, Netherlands

m.l.s.deheerkloots@uva.nl, H.Mohebbi@tilburguniversity.edu, c.m.pouw@uva.nl, G.Shen@tilburguniversity.edu, w.h.zuidema@uva.nl, martijn.bentum@ru.nl

Abstract

How language-specific are speech representations learned by self-supervised models? Existing work has shown that a range of linguistic features can be successfully decoded from end-to-end models trained only on speech recordings. However, it's less clear to what extent pre-training on specific languages improves language-specific linguistic information. Here we test the encoding of Dutch phonetic and lexical information in internal representations of self-supervised Wav2Vec2 models. Pre-training exclusively on Dutch improves the representation of Dutch linguistic features as compared to pre-training on similar amounts of English or larger amounts of multilingual data. This language-specific advantage is well-detected by trained clustering or classification probes, and partially observable using zero-shot metrics. Furthermore, the language-specific benefit on linguistic feature encoding aligns with downstream performance on Automatic Speech Recognition.

Index Terms: interpretability, self-supervised learning, language-specificity, speech recognition

1. Introduction

In recent years, self-supervised learning (SSL) algorithms have been demonstrated to learn powerful representations of spoken language, in terms of both their downstream task performance and the richness of their embedding spaces. Despite being trained only on unlabeled speech recordings, neural SSL models vastly outperform acoustic baselines at encoding various levels of linguistic structure, including phonological [1, 2], syllabic [3] and lexical [4] features. Understanding the linguistic information these models encode is relevant not only for engineers interpreting model behaviour, but also for cognitive scientists studying language acquisition from spoken input [5, 6].

Despite the wide interest for interpreting SSL model functioning, research has so far primarily focused on English linguistic features encoded by English-only or multilingually pre-trained models, with some notable exceptions [7, 8, 9, 10]. As a result, it remains somewhat unclear whether measures detecting linguistic structures in SSL models rely on truly language-specific representations or more language-general acoustic correlates. For downstream automatic speech recognition (ASR), both multi- and monolingually trained Wav2Vec2 [11] models are known to generalize well even to languages not present in their training data [12, 13]. This could indicate that what these models learn about speech is not typically language-specific [14], in line with findings that internal activations of Wav2Vec2 models do not seem strongly influenced by the model's training language in distinguishing native vs. non-native phoneme contrasts [10]. However, the same models do show distinct patterns in predicting brain activity for their pre-training language vs. a different language [15], and crosslingual comparisons of Wav2Vec2 models trained on tonal vs. non-tonal languages have found language-specific advantages in the representation of suprasegmental features [7, 8].

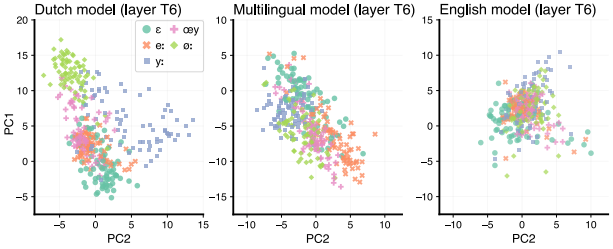

Figure 1: *A set of high front vowels (phonemically distinctive in Dutch) is most discriminably represented in a model trained exclusively on Dutch. Data points are vowel occurrences sampled across 3 speakers in Multilingual LibriSpeech.*

A key challenge in distinguishing language-specific vs. language-general speech processing by neural SSL models is the variability in measures and datasets used to identify language-specific representations. While some studies use 'zero-shot' embedding space distance metrics to measure the saliency of native vs. non-native phoneme contrasts [10], other studies use trained classifiers to detect how well language-specific information can be decoded from internal representations [7, 8]. Hence, the differences in reported results could be driven by differences in the methodologies used for testing the models' linguistic capabilities.

Here, we aim to investigate language-specific representations of Dutch linguistic information in self-supervised speech models, as well as the effects of different analysis pipelines on the detection of such information. For this purpose, we curated the *SSL-NL* evaluation set of publicly available Dutch speech data and alignments[1]. We designed the *SSL-NL* set to test the encoding of Dutch phonetic and lexical features in SSL speech representations, while allowing for comparisons across different analysis methods. We also pre-train a Wav2Vec2 model on the Dutch language exclusively[2] as an additional resource for studying Dutch-specific speech representations. Comparing the Dutch model to a monolingual English and a multilingual

[1]We release the *SSL-NL* dataset at http://doi.org/10.5281/zenodo.15548947

[2]We release *Wav2Vec2-NL* along with its training manifest and configuration at http://doi.org/10.5281/zenodo.15550628

10.21437/Interspeech.2025-1526

model, we find that both phone- and word-level features are best encoded in the Dutch model (e.g. Figure 1), which is also reflected in its downstream performance on Automatic Speech Recognition. Notably, the detection of language-specific benefits varies across analysis measures and datasets, underscoring the critical role of data selection and analysis methods in assessing the impact of language-specific pre-training on self-supervised speech representations.

2. Models

To investigate the effect of language-specific pre-training on linguistic feature encoding, we compare three Wav2Vec2 models with identical architectures (7 CNN + 12 Transformer layers), but with varying amounts of Dutch and other languages in their pre-training data.

We train **w2v2-nl** (`Wav2Vec2-NL`) on 960 hours of spoken Dutch, combining data from the Spoken Dutch Corpus (CGN; [16]), Multilingual Librispeech (MLS; [17]), and CommonVoice (CV; [18]). We sample 537 hours from CGN, including both Dutch and Flemish recordings but excluding telephone conversations (recordings with low sample rate) and sermons (poor recording quality). The remaining data includes recordings of spontaneous conversations and interviews, as well as read speech and news reports. We segment the CGN recordings into phrases, and use segments between 2 and 15 seconds in length as inputs for model training. From MLS, we sample 211 hours of the provided audiobook segments. The remaining 212 hours are sampled from CommonVoice, consisting of short segments of read aloud speech. Across the full training set, audio samples range between 2 and 32 seconds in length. We follow the original configuration for training Wav2Vec2 in [19], only modifying it to allow longer utterance length and per-device batch size. We use the fairseq toolkit [20] to pre-train w2v2-nl for 100k training steps using 4 Nvidia A100-40GB GPUs.

In addition to our Dutch model, we include two additional models trained on other languages, as well as a model trained on non-speech acoustics: **fb-en**, the self-supervised base model from the original Wav2Vec2 release, pre-trained on 960 hours of English audiobook recordings from LibriSpeech [19]; **fb-voxp-100k**, a multilingual base model pre-trained on 100k hours of European parliament recordings in 23 languages, including 4,5k hours of Dutch [21]; and a **nonspeech** base model trained on 600 hours of acoustic scenes from AudioSet [10][3].

3. Representational analysis methods

We create the *SSL-NL* evaluation set by sampling speech recordings from two different datasets of spoken Dutch. We use a separate subset of MLS audiobook segments (held-out from w2v2-nl training data), as well as the IFADV corpus [23] of face-to-face conversational speech. Based on the speech recordings and orthographic transcription pairs in each corpus, we obtain phone- and word-level forced alignments for both IFADV and MLS using the WebMAUS API. Phone- and word-level representations are extracted from each model layer by passing the full MLS audiobook segment or IFADV conversational turn as input to the model, and mean-pooling across frame representations within each phone or word respectively.

For our *phonetic analyses* we collect samples of 37 Dutch phone categories, including 13 vowels (aː, ɑ, ɛ, eː, øː, ɪ, iː, yː, ʉ, uː, oː, ɔ, ə), 3 diphthongs (ɛi, œy, ɑu) and 21 consonants (p, b, t, d, k, g, ʋ, f, v, s, z, ʃ, x, ɣ, h, m, n, ŋ, r, l, j). From MLS, we collect 6,086 phone occurrences in total from 11 different speakers; from IFADV we collect 3,294 phone occurrences in total from 9 speakers. We sample 15 occurrences from MLS and 10 occurrences from IFADV per speaker per phone, except for [øː], which occurs more rarely. For our *lexical analyses* we sample a total of 1,338 word occurrences (307 unique words) from MLS and 734 word occurrences (124 unique words) from IFADV, from which we select smaller subsets for our clustering and RSA analyses, as explained below.

3.1. Phonetic analyses

3.1.1. Phone identity probing

Following previous studies on English (e.g. [24]), we first analyze the representation of Dutch phones in our models by fitting a linear transform to decode phone identity from model hidden layer representations. We fit separate multinomial logistic regression probes for each hidden layer on data from 8 MLS speakers and 6 IFADV speakers. We evaluate probe prediction accuracy on the remaining 3 speakers for each dataset. Chance accuracy for the phone identity probes is $\frac{1}{37}$, since we defined 37 phone categories.

3.1.2. Phone ABX

As an alternative to fitting trained classification probes, the encoding of phonetic categories in SSL models has also been investigated by directly measuring distances in model embedding spaces (e.g. [10, 25]). In the phone ABX task, we sample phone occurrence triplets A, B, and X, where A and X belong to the same phoneme category but B belongs to a different category. Phone representations are then evaluated on whether cosine similarity is greater for within-category embedding pairs (comparing A and X) than for between-category embedding pairs (comparing A and B). We construct phone ABX triplets based on a set of 59 unique Dutch phoneme contrasts (e.g. /oː/ vs. /ɔ/), while sampling A, B and X from different contexts and speakers. We evaluate phone ABX accuracy across a set of over 222K ABX triplets (3.7K per contrast) sampled from MLS, and over 66K triplets from IFADV (1.1K per contrast). Chance accuracy for the ABX measure is $\frac{1}{2}$, since accuracy for each ABX triplet is binary (testing if AX-similarity $>$ AB-similarity).

3.1.3. Phone clustering

Our final phone-level analysis method examines how well model embedding spaces cluster phone samples together according to our pre-defined Dutch phone categories. The degree of clustering can be measured using silhouette scores [26, 27], computed by comparing the mean distance between same-cluster samples with their mean distance to the next nearest cluster. We compute layerwise average silhouette scores across phone clusters after reducing model embedding dimensionality to 36 dimensions, either unsupervised (by applying Principal Component Analysis; PCA), or by optimizing a linear transform for separability between phone categories (by applying Linear Discriminant Analysis; LDA). We evaluate the dimensionality reduced projections on a held-out set of speakers, after fitting them on a train set (applying the same split used for phone identity probing). Silhouette scores range between -1 and 1, with positive values indicating better cluster separability.

[3] HuggingFace [22] identifiers: `amsterdamNLP/Wav2Vec2-NL` (w2v2-nl); `facebook/wav2vec2-base` (fb-en); `facebook/wav2vec2-base-100k-voxpopuli` (fb-voxp-100k); `ewandunbar/humanlike-speech-2022` (nonspeech)

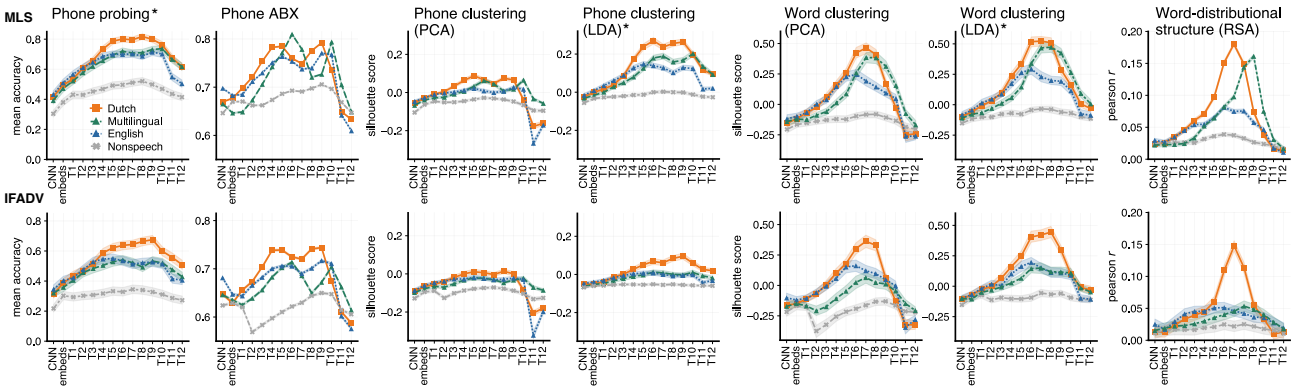

Figure 2: *Layerwise phonetic and lexical analyses, across a read speech (MLS, top row) and a dialogue (IFADV, bottom row) dataset of spoken Dutch. Measures marked * involve optimized linear transforms, whereas others are computed zero-shot; shading indicates 95% confidence intervals. The Dutch* ***w2v2-nl*** *model achieves highest scores across most analyses of Dutch phone and word encoding, though the size of this language-specific advantage varies considerably across analyses.*

3.2. Lexical analyses

3.2.1. Word clustering

For word clustering analyses, we chose the 50 most frequent words from the BAK list of basic Dutch preschooler vocabulary [28] within each evaluation subset. We sample up to 9 occurrences for every word, each from a different speaker and recording (448 occurrences in total for IFADV; 450 for MLS). We use two thirds of this set (6 out of 9 speakers) to compute and fit PCA and LDA transformations (reducing the embedding dimensionality to 49), and we use the final third for computing silhouette scores to evaluate the degree of clustering by word type.

3.2.2. Word-distributional structure (Fasttext RSA)

To analyze the encoding of Dutch word-distributional structure in our set of speech models, we evaluate representational similarity [29] of speech-based word embeddings to static text-based word embeddings from a Dutch Fasttext [30] model[4]. We compute representational similarity across a set of commonly occurring words from the same basic preschooler vocabulary list, sampling up to 3 occurrences from different speakers for every word (297 word types, 889 word tokens from MLS; 116 word types, 341 word tokens from IFADV). We measure representational similarity as the Pearson correlation between pairwise cosine distances for each of these word tokens and the corresponding distances in the text-based Fasttext word embeddings. In computing pairwise distances, we exclude word token pairs of the same type, to avoid effects of word identity.

4. Results of representational analyses

4.1. Phonetic and lexical representations improve with language-specific pre-training

Across most of our phonetic and lexical analyses, we observe a moderate to substantial advantage of the w2v2-nl model over its English and multilingual counterparts (Figure 2). This suggests that Dutch-specific pre-training enhances the encoding of Dutch phonetic and lexical features in self-supervised speech models. In the phonetic analyses, the small benefit reflects the fact that Dutch is phonetically relatively similar to English, as well as other European languages in the multilingual model's pre-training data — many of the phone categories in our analysis set are shared by these languages. However, when zooming in on phones more specific to Dutch, such as the high-font vowels [yː], [øː] and diphthong [œy], we see that they are more distinctively represented along the first principal components of the Dutch model's hidden layer representations, as compared to the multilingual and English models (Figure 1).

The English model also shows significantly higher scores than our non-speech baseline model on the lexical analyses. The majority of words in our lexical analysis sets are not particularly similar to English words, with some exceptions (e.g. [bɛt], [hart]). Rather than recognizing Dutch word forms, it is more likely that the English model better encodes acoustic differences between words than the nonspeech model. Nevertheless, language-specific information does seem to help the encoding of lexical structure, as reflected by the higher scores of the Dutch and multilingual model (which included Dutch in its pre-training data) over the English one.

4.2. Differences across analysis measures and datasets

While SSL representations have been analyzed for linguistic structure before, many such analyses have been performed on read speech datasets. Moreover, studies differ in the type of analysis method applied, for example using either zero-shot ABX tasks or trained probing classifiers. We observe variability across both the different analysis measures and the different datasets we performed our analyses on (Figure 2). Across measures which make use of linear projections to optimize phone identity decoding (probing, LDA), Dutch phonetic encoding is clearly enhanced in the Dutch model compared to the other models. However, this is less clear in measures analyzing phonemic contrasts by directly comparing distances within each model's representation space (ABX) or by analyzing the degree of phone clustering along a subset of its most variance-explaining components (PCA). The contrast between these results suggests that language-specific phonetic information may be encoded in a small subspace of model internal representations, which is decodable after a linear transformation, but not very prominently featured across the full embedding space.

In contrast, our lexical analyses show that the word-level benefit of language-specific pre-training is observable using both zero-shot (PCA, RSA) and optimized (LDA) analysis mea-

[4] `facebook/fasttext-nl-vectors`

sures. While phone clustering silhouette scores significantly improve after fitting LDA projections, word-level silhouette scores are relatively similar between PCA and LDA clustering metrics. Possibly, learned word identities are quite saliently represented in model embedding space, especially when mean-pooling across all 20 ms frame representations within words. Work on word representations in English SSL models has shown that word identity can be decoded from both mean-pooled and individual frame representations [4], suggesting that this is not an effect of the mean-pooling operation itself.

5. Downstream ASR performance

We fine-tune our SSL models for speech-to-text transcription to examine whether the language-specific advantages in self-supervised representations also lead to improved performance on downstream ASR tasks. Each SSL model is fine-tuned on Dutch read-aloud speech from the CGN (component o), using 78 hours of training data while reserving 10 hours each for development and testing. ASR performance is evaluated using word error rate (WER) on the held-out CGN-o test set. Additionally, we assess performance on test sets from CV and MLS (both containing read-aloud speech), IFADV (dialogue speech), and N-Best [31], a Dutch ASR benchmark featuring conversational telephone speech and news broadcast recordings. While we have no expectation of reaching state-of-the-art transcription performance with our models and fine-tuning set-up, we are interested in comparing relative model performance at this task to the observed differences in representational quality.

Across all test sets, the Dutch pre-trained **w2v2-nl** model consistently achieves lower WER than the English **fb-en** and multilingual **fb-voxp-100k** models (see Table 1). This suggests that the advantages of language-specific pre-training extend beyond self-supervised representations to downstream ASR performance. The performance of the Dutch model is followed by the multilingual and English models, aligning with the patterns observed in our phonetic and lexical analyses.

Models	CGN-o	IFADV	MLS	CV	N-Best
Dutch	10.4	65.6	15.4	21.0	25.2
English	21.5	84.4	32.3	47.9	53.1
Multilingual	12.7	78.8	23.1	34.2	37.2
Nonspeech	43.5	94.9	58.6	78.4	75.7

Table 1: *WER results for models fine-tuned on the read-aloud component of the CGN and evaluated on held out test set (CGN-o), the test sets of two read-aloud corpora (MLS and CV), a dialogue corpus (IFADV) and a Dutch benchmark (N-Best).*

6. Discussion & Conclusions

We introduced the *SSL-NL* evaluation set, and used it to compare a new monolingual Dutch Wav2Vec2 model against existing English and multilingual models. What do these models represent about the phonetic and lexical structure of Dutch?

We find that linguistic information at both levels can be accurately decoded from each model's hidden layer representations, using a variety of methods. While accuracy is similar across models at the output of the CNN feature encoder, a language-specific advantage for the Dutch model emerges across the model's Transformer module. Thus, language-specific pre-training does benefit the encoding of phonetic and lexical structure in Wav2Vec2 hidden layer representations.

We observe interesting differences across analysis methods. Phone ABX tasks are widely used to evaluate self-supervised speech representations, including in studies of native- vs. non-native speech processing [10, 25]. However, we find that they may be a less sensitive measure for detecting language-specific information that is decodable by training classification or clustering probes on model representations. This is an important consideration for future work aiming to study the language-specificity vs. -generality of self-supervised representations — especially in models with relatively high representational dimensionalities. For example, large multilingually pre-trained Wav2Vec2 models may make use of language-specific subspaces to achieve similar linguistic encoding accuracy for individual languages as monolingually pre-trained models [32, 13].

Between the Dutch **w2v2-nl** model and the other models, we generally observe larger differences on the IFADV dialogue dataset than on the MLS read speech dataset. We believe this reflects an effect of the pre-training data domain beyond its language-specificity: while the Dutch model's pre-training data included conversational speech, the English and multilingual models were respectively trained on read-aloud books and less spontaneous speech styles (parliament debates). The effect is particularly prominent for word-distributional structure, as measured by representational similarity between the speech models and the text-based Fasttext word embedding model. Word-distributional patterns are known to significantly differ between spoken and written language corpora [33], potentially limiting the generalizability of models trained on read speech [34]. Our analyses show the benefit of training on conversational speech not only for enhancing the representation of conversation-level structures, but also for the encoding of smaller linguistic units such as phones and words.

When fine-tuning our set of SSL models on ASR, we find that the benefit of language-specific pre-training is also reflected in transcription accuracy. However, probe performance and downstream task accuracy are not necessarily directly related, and rankings of self-supervised models have shown stark differences when models are used as feature extractors (i.e. probed) vs. when they are fine-tuned for the evaluation task [35, 36]. An interesting direction of future research could explore how the representation of linguistic features causally affects downstream text-transcription performance, for example using feature removal techniques to manipulate model representation spaces, as has been explored for the manipulation of gender representation in ASR models [37].

In this study, we compared models trained on languages with relatively high phonetic similarity, but still found observable benefits of language-specific pre-training on encoding Dutch phone categories, and stronger effects on linguistically more complex features such as word-distributional structure. We note that we have not tested for specific kinds of lexical knowledge such as syntactic or semantic categories — to the extent that self-supervised models represent such higher-level information, we would expect stronger language-specific effects, but this remains to be tested. We hope that the w2v2-nl model and the SSL-NL evaluation set provide valuable resources for such further investigations into language-specific representations in self-supervised speech models.

7. Acknowledgements

We would like to thank David van Leeuwen and Nik Vaessen for sharing the N-Best evaluation dataset. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-8324.

8. References

[1] L. ten Bosch, M. Bentum, and L. Boves, "Phonemic competition in end-to-end ASR models," in *Interspeech 2023*, pp. 586–590.

[2] M. Bentum, L. ten Bosch, and T. Lentz, "The Processing of Stress in End-to-End Automatic Speech Recognition Models," in *Interspeech 2023*, pp. 2350–2354.

[3] M. de Heer Kloots and W. Zuidema, "Human-like Linguistic Biases in Neural Speech Models: Phonetic Categorization and Phonotactic Constraints in Wav2Vec2.0," in *Interspeech 2024*, pp. 4593–4597.

[4] A. Pasad, C.-M. Chien, S. Settle, and K. Livescu, "What Do Self-Supervised Speech Models Know About Words?" *TACL*, vol. 12, pp. 372–391, 2024.

[5] M. Lavechin, M. de Seyssel, H. Titeux, G. Wisniewski, H. Bredin, A. Cristia, and E. Dupoux, "Simulating Early Phonetic and Word Learning Without Linguistic Categories," *Developmental Science*, vol. 28, no. 2, p. e13606, 2025.

[6] E. Dupoux, "Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner," *Cognition*, vol. 173, pp. 43–59, 2018.

[7] G. Shen, M. Watkins, A. Alishahi, A. Bisazza, and G. Chrupała, "Encoding of lexical tone in self-supervised models of spoken language," in *NAACL 2024*, Mexico City, Mexico, pp. 4250–4261.

[8] A. de la Fuente and D. Jurafsky, "A layer-wise analysis of Mandarin and English suprasegmentals in SSL speech models," in *Interspeech 2024*, pp. 1290–1294.

[9] Z. Dugonjić, A. Pupier, B. Lecouteux, and M. Coavoux, "What Has LeBenchmark Learnt about French Syntax?" in *LREC-COLING 2024*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., pp. 17 493–17 499.

[10] J. Millet and E. Dunbar, "Do self-supervised speech models develop human-like perception biases?" in *ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., pp. 7591–7605.

[11] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, vol. abs/2006.11477, 2020.

[12] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised Pretraining Transfers Well Across Languages," in *ICASSP 2020*, pp. 7414–7418, iSSN: 2379-190X.

[13] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, pp. 2278–2282.

[14] T. tom Dieck, P. A. Pérez-Toro, T. Arias, E. Noeth, and P. Klumpp, "Wav2vec behind the scenes: How end2end models learn phonetics," in *Interspeech 2022*, pp. 5130–5134.

[15] J. Millet, C. Caucheteux, P. Orhan, Y. Boubenec, A. Gramfort, E. Dunbar, C. Pallier, and J.-R. King, "Toward a realistic model of speech processing in the brain with self-supervised learning," *NeurIPS 2022*, vol. 35, pp. 33 428–33 443.

[16] I. Schuurman, M. Schouppe, H. Hoekstra, and T. van der Wouden, "CGN, an annotated corpus of spoken Dutch," in *LINC-03 at EACL 2003*.

[17] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Interspeech 2020*, pp. 2757–2761.

[18] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," 2020.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *NeurIPS 2020*, vol. 33, pp. 12 449–12 460.

[20] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "Fairseq: A Fast, Extensible Toolkit for Sequence Modeling." ACL, 2019, pp. 48–53.

[21] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation," in *ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., pp. 993–1003.

[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. v. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," 2020, arXiv.

[23] R. van Son, W. Wesseling, E. Sanders, and H. van den Heuvel, "The IFADV Corpus: a Free Dialog Video Corpus," in *LREC 2008*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, and D. Tapias, Eds. ELRA.

[24] P. C. English, J. Kelleher, and J. Carson-Berndsen, "Domain-informed probing of wav2vec 2.0 embeddings for phonetic features," in *SIGMORPHON 2022*, pp. 83–91.

[25] M. Poli, T. Schatz, E. Dupoux, and M. Lavechin, "Modeling the initial state of early phonetic learning in infants," *Language Development Research*, vol. 5, no. 1, 2024.

[26] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *JMLR*, vol. 12, pp. 2825–2830, 2011.

[28] F. Mulder, Y. Timman, and S. Verhallen, *Handreiking bij de Basiswoordenlijst Amsterdamse Kleuters (BAK)*. ITTA, 2009.

[29] N. Kriegeskorte, M. Mur, and P. A. Bandettini, "Representational similarity analysis - connecting the branches of systems neuroscience," *Frontiers in Systems Neuroscience*, vol. 2, no. 4, pp. 1–28, 2008.

[30] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *TACL*, vol. 5, pp. 135–146, 2017.

[31] J. M. Kessens and D. A. van Leeuwen, "N-best: the northern- and southern-dutch benchmark evaluation of speech recognition technology." in *Interspeech 2007*, 2007, pp. 1354–1357.

[32] B. M. Abdullah, M. M. Shaik, B. Möbius, and D. Klakow, "An information-theoretic analysis of self-supervised discrete representations of speech," in *Interspeech 2023*, pp. 2883–2887.

[33] M. Bentum, L. Ten Bosch, A. Van den Bosch, and M. Ernestus, "Do speech registers differ in the predictability of words?" *International Journal of Corpus Linguistics*, vol. 24, no. 1, pp. 98–130, 2019.

[34] M. Dingemanse and A. Liesenfeld, "From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology," in *ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland, pp. 5614–5633.

[35] S. Zaiem, Y. Kemiche, T. Parcollet, S. Essid, and M. Ravanelli, "Speech self-supervised representation benchmarking: Are we doing it right?" in *Interspeech 2023*, pp. 2873–2877.

[36] T. Parcollet, H. Nguyen, S. Evain, M. Zanon Boito, A. Pupier, S. Mdhaffar, H. Le, S. Alisamir, N. Tomashenko, M. Dinarelli, S. Zhang, A. Allauzen, M. Coavoux, Y. Estève, M. Rouvier, J. Goulian, B. Lecouteux, F. Portet, S. Rossato, F. Ringeval, D. Schwab, and L. Besacier, "LeBenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of French speech," *Computer Speech & Language*, vol. 86, p. 101622, 2024.

[37] A. Krishnan, B. M. Abdullah, and D. Klakow, "On the encoding of gender in transformer-based asr representations," in *Interspeech 2024*, pp. 3090–3094.