How Bilingual Are SSL Speech Models? Cross-Lingual Probing of Articulatory Encoding with Finnish and Russian EMA

Ailín Pollio San Pedro[1,2], *Tomi Kinnunen*[1], *Alexandre Nikolaev*[1], *Ruchi Pandey*[1]

[1]University of Eastern Finland, Computational Speech Group, Finland
[2]CNRS, Gipsa-Lab, France

ailin.pollio-san-pedro@gipsa-lab.grenoble-inp.fr

Abstract

Self-supervised learning (SSL) models such as wav2vec 2.0 [1], HuBERT [2], and WavLM [3] have become foundational in modern speech research. Trained on large-scale unlabeled audio, they learn structured representations that transfer effectively to diverse downstream tasks [4]. However, the precise nature of these representations remains only partially understood. Probing studies have revealed that SSL features encode phonetic categories [5, 6, 7], speaker characteristics [8], and suprasegmental structure (e.g. prosody and tone) [9], yet these models are often treated as opaque feature extractors.

In this work, we investigate articulatory encoding in SSL representations using bilingual Finnish–Russian electromagnetic articulography (EMA) data from the FROST-EMA corpus [10], enabling one of the first systematic analyses of articulatory encoding across bilingual conditions (L1, L2, and accent imitation). We extract layer-wise representations from five SSL models that differ in their "familiarity" with Finnish and Russian (Wav2Vec 2.0 Large[1]; MMS-300m[2]; XLSR-53[3]; XLSR-53 fine-tuned on Russian[4]; and XLS-R fine-tuned on Finnish[5]). Linear regression probes [11] are trained to predict articulatory trajectories from latent features for five articulators—tongue tip, tongue body, tongue dorsum, upper lip, and lower lip—each represented in horizontal and vertical dimensions. Performance is evaluated using Pearson correlation between predicted and reference EMA trajectories.

Table 1 summarizes five complementary experiments showing that SSL models encode substantial articulatory information across Finnish and Russian. In cross-model comparison (E1), multilingual and language-adapted models (MMS-300m, XLS-R fine-tuned) achieve the highest performance (mean $r \approx 0.69$), outperforming Wav2Vec 2.0 Large ($r = 0.641$) and XLSR-53 ($r = 0.62$), which also exhibits a sharp late-layer degradation. Layer–sensor analyses (E2), shown in Figures 1(a)–(c), reveal that articulatory prediction performance varies systematically across encoder depth. Across models, performance peaks at intermediate layers, with X/Z axes generally easier to predict than Y. This pattern supports a hierarchical organization in which intermediate layers capture articulatory structure, while deeper layers encode more abstract representations [12, 13]. As revealed in Figure 3, training-size experiments (E3) show high data efficiency, with performance rapidly increasing and saturating after ∼5 minutes of paired data. Speaker generalization (E4) reaches up to $r \approx 0.78$, with tongue articulators (Tongue Tip/Tongue Blade) more reliably predicted than lips, while upper-lip vertical motion remains consistently challenging. As observed in Figure 2, task and language analyses (E5) further indicate higher correlations for controlled speech ($r \approx 0.70$–0.74) than spontaneous speech ($r \approx 0.58$–0.62), suggesting that reduced linguistic variability facilitates articulatory probing, while L2 and accented speech remain robustly decodable (up to $r \approx 0.76$), indicating accent-robust encoding. Together, these results point to a shared articulatory subspace underlying SSL representations across languages and conditions.

These findings support the hypothesis that SSL representations encode partially language-agnostic articulatory structure, even across typologically distant languages. In turn, articulatory constraints appear to be recoverable from acoustics through large-scale representation learning, opening practical avenues for applications in low-resource articulatory modeling, clinically oriented speech analysis, and pronunciation assessment of L2 learners. Future work will investigate fine-tuning multilingual encoders on target languages, combining representations across models, and employing nonlinear or sequential probing methods to better capture gestural timing.

Index Terms: Self-supervised learning, articulatory representations, cross-language analysis

Table 1: *Overview of experimental configurations. LOSO = Leave-One-Speaker-Out.*

ID	Experimental configuration and evaluation
E1	**Cross-model comparison**. 18 speakers (combined). Models: WV2-L, MMS-300m, XLSR-53, RU-FT, FI-FT. Output: Mean articulatory score across layers and EMA dimensions.
E2	**Sensor–layer mapping**. 18 speakers (combined). Models: WV2-L, MMS-300m, XLSR-53. Output: Pearson r per EMA dimension and layer (1–24).
E3	**Training-size sensitivity**. 18 speakers (combined). Model: WV2-L. Output: Pearson r vs. probe training duration (20 s–20 min).
E4	**Speaker generalization (LOSO)**. $S-1$ train / 1 test. Model: MMS-300m. Output: Pearson r per speaker and EMA dimension.
E5	**Task & proficiency effects**. LOSO within Finnish/Russian. Model: MMS-300m. Output: Pearson r per speaker, EMA dimension, task, and condition (L1, L1+accent, L2).

[1]https://huggingface.co/facebook/wav2vec2-large
[2]https://huggingface.co/facebook/mms-300m
[3]https://huggingface.co/facebook/wav2vec2-large-xlsr-53
[4]https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-russian
[5]https://huggingface.co/aapot/wav2vec2-xlsr-300m-finnish-lm

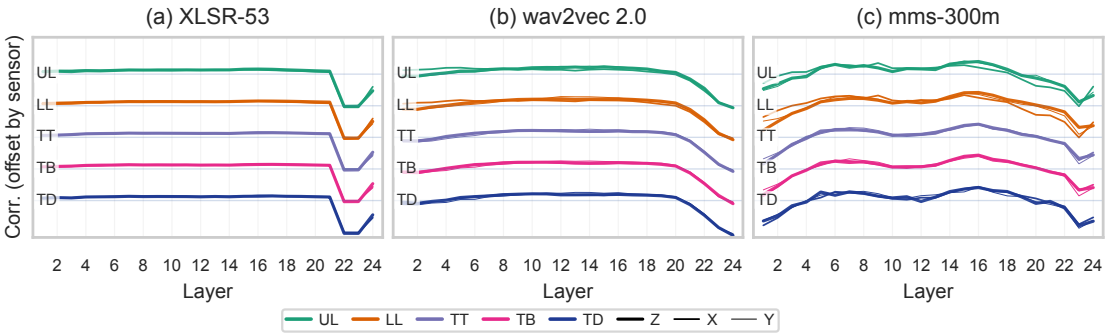

Figure 1: *Layer-wise articulatory prediction performance (average Pearson correlation r) across encoder depth for (a) XLSR-53, (b) wav2vec 2.0, and (c) MMS-300m. Curves are vertically offset by sensor for visual clarity. While distinct correlation trajectories are observed across spatial axes, the overall layer-wise profiles remain highly similar across articulators and coordinate axes, suggesting consistent probing dynamics throughout the network.*

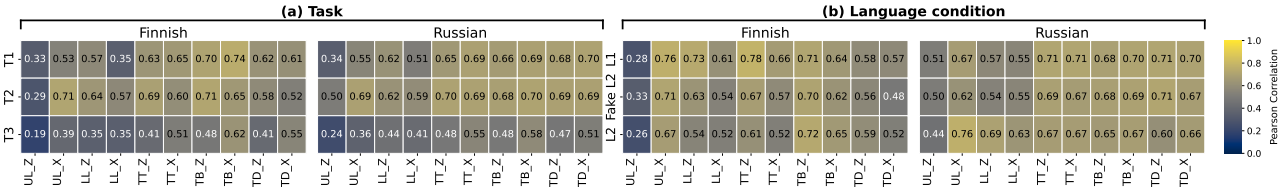

Figure 2: *Layer-wise articulatory prediction performance (Pearson correlation r) for MMS-300m under a LOSO evaluation. (a) Task comparison across Finnish and Russian speakers, showing higher correlations for controlled reading tasks (T1-T2) than for spontaneous picture description (T3). (b) Language-condition comparison (L1, L2, and simulated accent conditions.)*

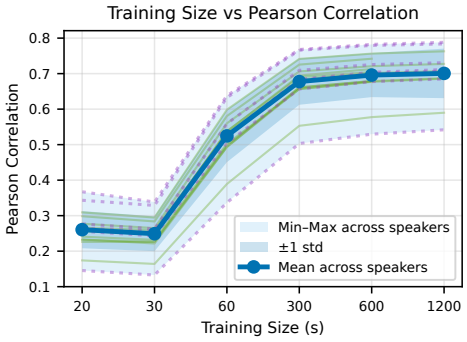

Figure 3: *Performance increases with training size and stabilizes around 300 seconds. The solid green lines correspond to Finnish speakers, while the dashed purple lines indicate Russian speakers.*

1. References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv*, 2020.

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

[3] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, D. Yu, S. Yu, Z. Yao, Y. Qian, X. Huang, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE JSTSP*, vol. 16, no. 6, pp. 1505–1518, 2022.

[4] S.-w. Yang, P.-H. Chi, Y.-S. Chuang *et al.*, "SUPERB: Speech processing universal performance benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

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

[6] K. Martin, J. Gauthier, C. Breiss, and R. Levy, "Probing self-supervised speech models for phonetic and phonemic information: A case study in aspiration," *arXiv preprint arXiv:2306.06232*, 2023.

[7] N. Venkateswaran, K. Tang, and R. Wayland, "Probing for phonology in self-supervised speech representations: A case study on accent perception," *arXiv preprint arXiv:2506.17542*, 2025.

[8] A. Y. F. Chiu, K. C. Fung, R. T. Y. Li, J. Li, and T. Lee, "A large-scale probing analysis of speaker-specific attributes in self-supervised speech representations," 2025.

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

[10] S. Hopponen, T. Kinnunen, A. Nikolaev, R. González Hautamäki, L. Tavi, and E. Meister, "FROST-EMA: Finnish and Russian Oral Speech Dataset of Electromagnetic Articulography Measurements with L1, L2 and Imitated L2 Accents," in *Proc. Interspeech 2025*, 2025, pp. 364–368.

[11] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.

[12] C. J. Cho, P.-c. Wu, A. Mohamed, and G. K. Anumanchipalli, "Evidence of vocal tract articulation in self-supervised learning of speech," in *ICASSP 2023*, 2023, pp. 1–5.

[13] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *ASRU 2021*, 2021, pp. 914–921.