

Listening for Speaker Identity in Multilingual Speech Models: Phonetic Similarity Predicts Transfer in Cross-Lingual Speaker Verification

AriaRay Brown¹, Yusser Al Ghussin^{1,2}, Dietrich Klakow¹

¹Department of Language Science and Technology, Saarland University, Germany,

²German Research Center for Artificial Intelligence (DFKI), Germany
arbrown@lsv.uni-saarland.de

Abstract

Multilingual self-supervised transformer speech models build shared representations to encode speech, but how *speaker identity* is shared across languages, and what modulates its transfer, remains unclear. We probe XLS-R and WavLabLM for speaker information in the task of cross-lingual speaker verification using layer-weighted speech embeddings from ten languages spanning four families. Motivated by the Language Familiarity Effect (LFE) in human speech perception, we investigate whether phonetic similarity modulates the success of cross-lingual transfer. We introduce a model-internal method for computing *phonetic similarity*: applying the Sinkhorn distance between per-language distributions of contextualized phone embeddings at a phonetically prominent model layer. A linear mixed-effects model across all 90 language pairs shows that this phonetic distance significantly predicts cross-lingual speaker-verification accuracy in both XLS-R ($\beta = -3.13, p < 0.0001$) and WavLabLM ($\beta = -1.22, p < 0.0001$). These findings suggest that phonetic similarity is an important factor for facilitating cross-lingual transfer of speaker information in multilingual self-supervised speech models.

1 Introduction

Multilingual self-supervised learning (SSL) speech models such as XLS-R (Babu et al., 2022) and WavLabLM (Chen et al., 2023) build shared representations that encode both linguistic and paralinguistic properties of speech in multiple languages. Probing studies have shown that intermediate layers capture rich linguistic structure (Pasad et al., 2021; Chen et al., 2022a), and that paralinguistic properties such as speaker gender can transfer across languages (Gonen et al., 2022). Yet how *speaker identity*, in particular, is encoded and shared remains underexplored.

Speaker verification (SV), determining whether

two utterances belong to the same speaker, is a natural lens for this question: it is sensitive to both acoustic individuality and language-specific phonetic structure. In human speech perception, the **Language Familiarity Effect** (LFE) shows that listeners discriminate speakers more accurately in familiar languages, with performance depending on the phonological proximity of the target language to the listener’s native language (Goggin et al., 1991; Johnson et al., 2018; de Seyssel et al., 2022). We ask whether an analogous effect arises in multilingual speech models: does *phonetic similarity* between languages, measured directly from model embeddings, predict cross-lingual transfer in speaker verification?

Contributions. We make these contributions:

1. We probe XLS-R and WavLabLM for speaker information across all layers and ten phylogenetically diverse languages using layer-weighted utterance representations.
2. We introduce an embedding-based method for computing phonetic similarity between languages via Sinkhorn distances over contextualized phone embedding distributions, extracted from a phonetically prominent layer identified by canonical correlation analysis (CCA).
3. We demonstrate via Spearman correlations and a linear mixed-effects model that phonetic similarity significantly predicts cross-lingual speaker verification accuracy for XLS-R and WavLabLM.

2 Related Work

Speaker information in SSL models. Speaker identity has been found at intermediate and final layers of SSL speech models (Chen et al., 2022a; Pasad et al., 2021), with scalar-mixed layers outperforming single-layer probes for speaker verifica-

tion (Chen et al., 2022b; Chiu et al., 2026). In English models, speaker identity and phonetic content occupy largely orthogonal dimensions, enabling speaker normalization (van Niekerk et al., 2021; Liu et al., 2023). Our work extends such analysis to two *multilingual* SSL models across 24 layers and ten languages.

Cross-lingual transfer of paralinguistic properties. Gonen et al. (2022) showed that classifiers trained on multilingual text representations transfer gender detection across languages, suggesting that paralinguistic features can be shared. Cross-lingual speaker verification in neural models has received limited attention; prior work (Reuter and Jessen, 2025) offers a recent point of comparison for our methodology.

Phonetic similarity and the LFE. de Seyssel et al. (2022) demonstrated that a graded measure of language distance progressively increases the difficulty of human cross-lingual speaker verification. In their model of LFE, language distance largely pertained to similarities in phonology, prosody, and phonotactics of a perceived language. We adapt this insight by deriving language distance from model embeddings that target phonetic properties.

3 Methodology

Our methodology has three stages: (1) probing SSL models for speaker information via layer-weighted utterance representations in the SV task, (2) computing embedding-based phonetic similarity between language pairs, and (3) testing whether phonetic similarity predicts cross-lingual SV accuracy.

Models. **XLS-R 300M** (Babu et al., 2022) is a wav2vec 2.0 model trained on 436k hours of speech in 128 languages using a masked contrastive prediction objective against quantized targets. **WavLabLM-MS-40k** (Chen et al., 2023) is a smaller, HuBERT-style encoder trained on only 40k hours in 136 languages with a joint masked prediction and denoising objective. Both use 24-layer transformer encoders with 1024-dimensional hidden states and achieve comparable performance on several multilingual speech processing tasks.

Data. We use **GlobalPhone** (Schultz et al., 2013), a multilingual corpus of read speech recorded under uniform conditions, for ten languages spanning four families: German (de), French (fr), Czech (cs), Polish (pl), Bulgarian (bg),

Croatian (hr), Portuguese/Brazilian (pt), Russian (ru), Ukrainian (uk), and Vietnamese (vi). For probing, we sample 10 speakers per language, 50 utterances per speaker, and construct balanced sets of 1,000 same-speaker and 1,000 different-speaker pairs for training and evaluation.

3.1 Speaker Verification: Probing Task

Paired utterance embeddings. For each utterance, we extract frame embeddings with $n = 1024$ features per frame from all 24 layers. We pool frames in a given layer via statistics pooling (Stafylakis et al., 2022), which concatenates the mean of all frames with the standard deviation. This yields a single-layer utterance vector with dimensionality $2n$. To create *utterance pairs* for classifier input, we then combine two utterance vectors u, v into a single $4n$ -dimensional vector through the formula $|u - v| \oplus (u \odot v)$.

Layer-weighted representations. We aggregate speaker information across layers using scalar mixing of layer embeddings (Gardner et al., 2018):

$$\tilde{e} = \sum_{i=1}^{24} w_i e_i, \quad w_i \geq 0, \quad \sum_i w_i = 1.$$

This serves two purposes: (1) classifiers consider all available speaker information in the model when training for SV, and (2) cross-lingual SV can be consistently performed across languages without compromising the choice of best layer.

Weighting methods. We test four: *uniform* (equal weights; baseline), *informed* (weights derived from per-layer monolingual SV accuracy), *amplified* (informed weights with layers above 0.5 scaled by 10), and *amp-all* (amplified weights derived from SV accuracy across all languages). The amplified method produces the highest mean *within-language* and *foreign-language* accuracies for both XLS-R and WavLabLM (Appendix B). We therefore use *amplified* layer-weight distributions per training language to best approximate speaker information.

Probing classifiers. We train a multi-layer perceptron classifier (MLPClassifier) and a linear classifier (SGDClassifier with squared hinge loss) on the weighted paired utterance embeddings.

3.2 Phonetic Similarity Pipeline

STEP 1. Phone alignment. Transcribed utterances are mapped to phone labels and aligned to

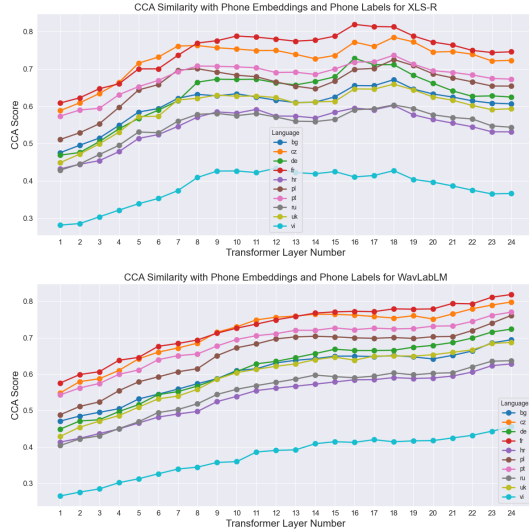


Figure 1: CCA between contextualized phone embeddings and one-hot IPA phone labels per language and transformer layer for XLS-R (top) and WavLabLM (bottom). Language plots from top to bottom in each figure: fr, cz, pt, pl, de, bg, uk, ru, hr, vi.

timestamps in the audio using the Montreal Forced Aligner (McAuliffe et al., 2017). This yields between 14,884 (vi) and 74,167 (hr) timestamped phone segments per language (Appendix D).

STEP 2. CCA phonetic layer selection. Following Pasad et al. (2023), we apply canonical correlation analysis (CCA) between phone embeddings and one-hot phone labels across layers to identify the layer whose representations correlate most strongly with phonetic content. This layer is used for all subsequent phonetic similarity calculations. CCA peaks at the same layer across nearly all ten languages¹: layer 18 for XLS-R and layer 24 for WavLabLM (Figure 1).²

STEP 3. Phone embeddings extraction. From the phonetic layer we extract the embedding vector for each utterance, select the aligned frames per phone segment, and use mean pooling to produce a single phone frame vector. For each language we thus obtain a distribution of phone embeddings.

STEP 4. Sinkhorn distance. The model-internal phonetic distance between languages L_i and L_j is then computed as the *Sinkhorn distance* (Curi, 2013), an optimization of the Wasserstein dis-

¹For German and French phone embeddings from XLS-R, layer 16 exhibits the highest CCA similarity, while nearby layers 17 and 18 show similarly high scores.

²Layer results are in accordance with Pasad et al. (2023) who report similar trends for the base models of XLS-R and WavLabLM, namely wav2vec 2.0 and HuBERT.

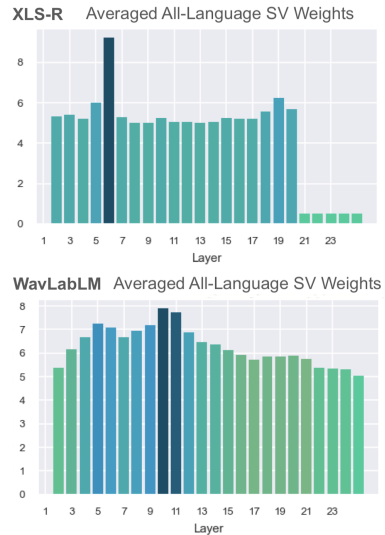


Figure 2: Distribution of layer weights averaged over all languages for models XLS-R (top) and WavLabLM (bottom). Depicts *layer-wise speaker information*.

tance³, between the two phone embedding distributions, using the Python Optimal Transport library (`ot.sinkhorn()`).

4 Results

4.1 Where Is Speaker Information Encoded?

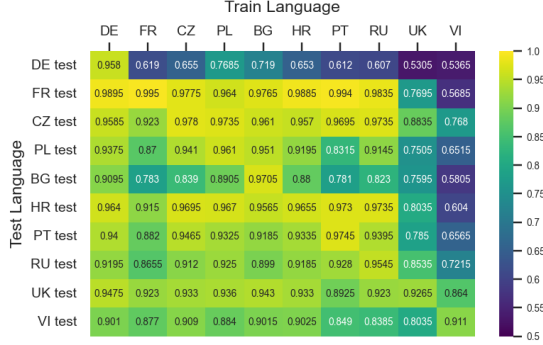
Monolingual SV accuracy peaks at early and late transformer layers for all ten languages in both models, consistent with prior findings on SSL speech models (Chen et al., 2022a). Per-language layer-weight profiles, computed from the *amplified* weighting scheme, are similar over languages within each model, indicating that speaker information occupies multilingual shared representational structure in these conditions (Figure H.10). Layer-weight distributions averaged over all languages are shown in Figure 2.

4.2 Cross-Lingual Speaker Verification

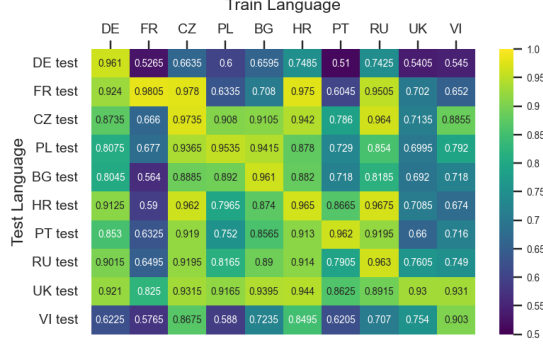
Training on one language and evaluating on another (cross-lingual SV) reveals variance in transfer performance, ranging from near-monolingual accuracy for some language pairs to near-chance⁴ for others. Slavic language pairs (e.g. Polish–Czech, Bulgarian–Ukrainian) tend to transfer well; pairs including Vietnamese show the weakest performance (Figure 3).

³Also known as Earth Mover’s Distance.

⁴Chance level here defines an accuracy of 50% since classifiers are trained to predict 1 or 0, for same-speaker or different-speaker pairs.



(a) XLS-R



(b) WavLabLM

Figure 3: Cross-lingual speaker-verification accuracy using *amplified* layer weights, optimized for speaker information in the training language. Results from maximum accuracy over MLP and SGD probing classifiers. Rows: test language. Columns: train language.

Results between classifiers are overall consistent. The MLPClassifier performs slightly better on all language pairs except when trained on Vietnamese for foreign-language tests, where the SGDClassifier performs best.

4.3 Phonetic Similarity Predicts Transfer

The Sinkhorn distance matrices (Figure 4) show minimal distance along the diagonal for same-language pairs, as expected. Within each model, the matrix captures linguistically intuitive trends between pairs of Slavic languages with lower distances, and shows Vietnamese as the most distant language from every other language.

Per-test-language Spearman correlations. For each test language we compute Spearman’s ρ between cross-lingual SV accuracy (across all training languages) and the corresponding phonetic distance. Negative correlations, indicating that lower distance yields higher accuracy, are statistically significant for some but not all test languages in both models ($p < 0.05$), with effect size varying by language (Table 1).

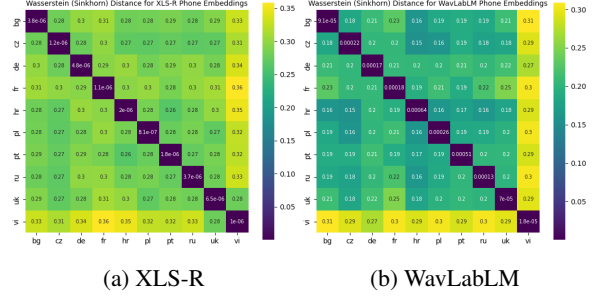


Figure 4: Sinkhorn distance between per-language distributions of contextualized phone embeddings at the CCA-located phonetic layer. Darker cells indicate more similar languages.

Test lang.	XLS-R		WavLabLM	
	r_s	p	r_s	p
Bulgarian	−0.41	0.244	−0.80	0.005
Czech	−0.58	0.080	−0.77	0.009
German	−0.50	0.138	−0.55	0.098
French	−0.71	0.022	−0.50	0.138
Croatian	−0.72	0.019	−0.60	0.067
Polish	−0.58	0.082	−0.82	0.004
Portuguese	−0.53	0.117	−0.49	0.150
Russian	−0.43	0.214	−0.65	0.043
Ukrainian	−0.18	0.613	−0.37	0.293
Vietnamese	−0.13	0.726	−0.65	0.043

Table 1: Per-test-language Spearman r_s between phonetic distance and cross-lingual speaker-verification accuracy, where accuracy is measured across all ten source-language trained classifiers. **Bold:** $p < 0.05$.

Overall analysis: all language pairs. To estimate the overall association between phonetic distance and cross-lingual SV performance, while accounting for test-language variability, we fit a linear mixed-effects model

$$\text{accuracy} \sim \text{phonetic_distance} + (1 \mid \text{test_language})$$

with phonetic distance as a fixed effect and per-test-language random intercepts. All 90 cross-lingual language pairs are pooled per model. In both XLS-R and WavLabLM, phonetic distance is a significant negative predictor of cross-lingual SV accuracy ($p < 10^{-13}$ and $p < 10^{-4}$):

$$\begin{aligned} \beta_{\text{XLS-R}} &= -3.13 \quad [95\% \text{ CI} : -3.93, -2.34], \\ \beta_{\text{WavLabLM}} &= -1.22 \quad [95\% \text{ CI} : -1.82, -0.61]. \end{aligned}$$

confirming that phonetic similarity, as measured in the model’s own embedding space, significantly improves cross-lingual transfer of speaker information (Figure 5).⁵

⁵Coefficient magnitudes should be interpreted separately, as the phonetic distance predictors are model-specific.

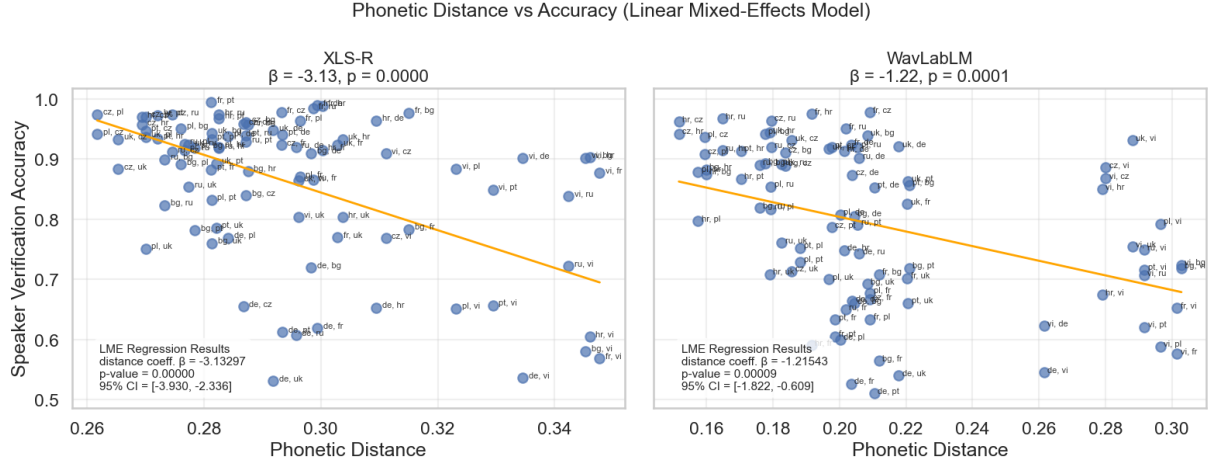


Figure 5: Phonetic distance vs. cross-lingual speaker-verification accuracy, with linear mixed-effects fit per model. Each point is a (test, train) language pair; same-language pairs are excluded.

5 Discussion

Our results support a computational parallel to the Language Familiarity Effect where multilingual SSL speech models transfer speaker verification more successfully between phonetically similar languages. That is, the closer two languages’ phone-embedding distributions are at the model’s phonetic layer, the better an SV classifier trained on one language transfers to the other. Critically, our phonetic similarity measure is derived from the same model embeddings used for the SV task.

The variation in effect strength across test languages suggests that phonetic distance may not be the sole predictor of cross-lingual transfer. This could be due to the limited scale of test languages and speakers, or the influence of other properties in the speech signal. Future work should assess the robustness of the measure across different phonetic layer choices, larger speaker sets, and additional downstream tasks.

Model differences in the results of cross-lingual SV suggest that WavLabLM exhibits weaker transfer than XLS-R. This may indicate that speaker information in WavLabLM is less consistently shared across languages and thus less transferable. Whether this difference persists across models with similar training objectives or larger language and speaker sets remains for future work. Nonetheless, phonetic distance was a significant predictor of cross-lingual SV in both models.

6 Conclusion

We investigated speaker representations in the multilingual SSL speech models XLS-R and

WavLabLM across ten languages. Speaker information is similarly encoded across layers and languages within each model. When language-specific dimensions were emphasized, within-language speaker-verification accuracy increased, while cross-lingual accuracy varied between language pairs. To account for this, we introduced a model-internal measure of phonetic similarity based on the Sinkhorn distance between contextualized phone-embedding distributions extracted from CCA-identified phonetic layers. Across all 90 language pairs, phonetic distance significantly predicts cross-lingual speaker verification accuracy in both XLS-R ($\beta = -3.13, p < 0.0001$) and WavLabLM ($\beta = -1.22, p < 0.0001$), broadly reflecting the Language Familiarity Effect for self-supervised speech representations. These findings suggest that phonetic similarity should be considered for improving cross-lingual transfer in speaker-related tasks involving multilingual SSL speech models. We encourage the use of contextualized phonetic similarity for conducting further fine-grained analyses.

Limitations

Our analysis is limited to ten languages and ten speakers each from the GlobalPhone corpus. The small per-language speaker pool may inflate variance in cross-lingual SV results. Conclusions about phonetic similarity effects could be strengthened with more speakers, the inclusion of spontaneous speech, and a wider language sample for more variation in phonetic differences. Dialect variation within a language, for example in German, may

also affect the cross-lingual results and resulting phonetic similarity measures. Additionally, our phonetic similarity metric depends on the quality of forced alignment for the resulting phone embeddings, and the chosen phonetic layer; ablations over these choices remain as future work.

Ethical Considerations

This work analyzes an existing licensed speech corpus, GlobalPhone, released for use in research with anonymous speakers and consented recordings. We use anonymous speaker data in this work to probe model representations, without attempting to uncover personal speaker identity. We encourage future work that pairs improved cross-lingual speaker verification with advances in voice anonymization.

Acknowledgments

This research is supported by funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB/CRC 1102.

References

- Arun Babu, Changan 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](#). In *Proceedings of Inter-speech 2022*, pages 2278–2282.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022a. [WavLM: Large-scale self-supervised pre-training for full stack speech processing](#). *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- William Chen, Jiatong Shi, Brian Yan, Dan Berrebbi, Wangyou Zhang, Yifan Peng, Xuankai Chang, Soumi Maiti, and Shinji Watanabe. 2023. [Joint prediction and denoising for large-scale multilingual self-supervised learning](#). In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.
- Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. 2022b. Large-scale self-supervised speech representation learning for automatic speaker verification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6147–6151. IEEE.
- Aemon Yat Fei Chiu, Kei Ching Fung, Roger Tsz Yeung Li, Jingyu Li, and Tan Lee. 2026. [A large-scale probing analysis of speaker-specific attributes in self-supervised speech representations](#). *Preprint*, arXiv:2501.05310.
- Marco Cuturi. 2013. [Sinkhorn distances: Lightspeed computation of optimal transportation distances](#). *Advances in Neural Information Processing Systems*, 26:2292–2300.
- Maureen de Seyssel, Guillaume Wisniewski, and Emmanuel Dupoux. 2022. [Is the Language Familiarity Effect gradual? A computational modelling approach](#). *Proceedings for the Annual Meeting of the Cognitive Science Society*.
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, 28th edition. SIL International, Dallas, Texas.
- Otmar Ertl. 2020. [ProbMinHash – A Class of Locality-Sensitive Hash Algorithms for the \(Probability\) Jaccard Similarity](#). *IEEE Transactions on Knowledge and Data Engineering*.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. [AllenNLP: A deep semantic natural language processing platform](#). In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.
- Judith P. Goggin, Charles P. Thompson, Gerhard Strube, and Liza Raquel Simental. 1991. [The role of language familiarity in voice identification](#). *Memory & Cognition*, 19:448–458.
- Hila Gonen, Shauli Ravfogel, and Yoav Goldberg. 2022. [Analyzing gender representation in multilingual models](#). In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 67–77, Dublin, Ireland. Association for Computational Linguistics.
- Elizabeth K. Johnson, Laurence Bruggeman, and Anne Cutler. 2018. [Abstraction and the \(misnamed\) language familiarity effect](#). *Cognitive Science*, 42(2):633–645.
- Oli Danyi Liu, Hao Tang, and Sharon Goldwater. 2023. [Self-supervised Predictive Coding Models Encode Speaker and Phonetic Information in Orthogonal Subspaces](#). In *Interspeech 2023*, pages 2968–2972.
- Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. [Montreal forced aligner: Trainable text-speech alignment using kaldi](#). In *Interspeech 2017*, pages 498–502.
- Ryan Moulton and Yunjiang Jiang. 2018. [Maximally Consistent Sampling and the Jaccard Index of Probability Distributions](#). In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*,

pages 1094–1103, Los Alamitos, CA, USA. IEEE Computer Society.

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. IEEE.

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. IEEE.

Paul M. Reuter and Michael Jessen. 2025. [On the influence of language similarity in non-target speaker verification trials](#). In *Proceedings of Interspeech 2025*, pages 3998–4002.

Tanja Schultz, Ngoc Thang Vu, and Tim Schlippe. 2013. [GlobalPhone: A multilingual text & speech database in 20 languages](#). In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8126–8130.

Themis Stafylakis, Ladislav Mošner, Sofoklis Kakourous, Oldřich Plchot, Lukáš Burget, and Jan Černocký. 2022. [Extracting speaker and emotion information from self-supervised speech models via channel-wise correlations](#). In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1136–1143.

Gonzalo Travieso, Alexandre Benatti, and Luciano da F. Costa. 2024. [An Analytical Approach to the Jaccard Similarity Index](#). *Preprint*, arXiv:2410.16436.

Benjamin van Niekerk, Leanne Nortje, Matthew Baas, and Herman Kamper. 2021. [Analyzing Speaker Information in Self-Supervised Models to Improve Zero-Resource Speech Processing](#). In *Interspeech 2021*, pages 1554–1558.

A Phonetic Similarity Pipeline

Further details for computing phonetic similarity are provided below. Figure 6 illustrates our method.

1. Phone alignment. Each utterance is force-aligned to its IPA phone sequence with MFA (McAuliffe et al., 2017), yielding between 14,884 (vi) and 74,167 (hr) phone segments per language (Appendix D).

2. CCA phonetic layer selection. For each language and each transformer layer we run canonical correlation analysis on (phone embedding, one-hot phone label) pairs, following Pasad et al. (2023). Across all ten languages, CCA similarity peaks at **layer 18 for XLS-R** and at **layer 24 for WavLabLM** (Figure 1). We treat these as the “phonetic layer” of each model.

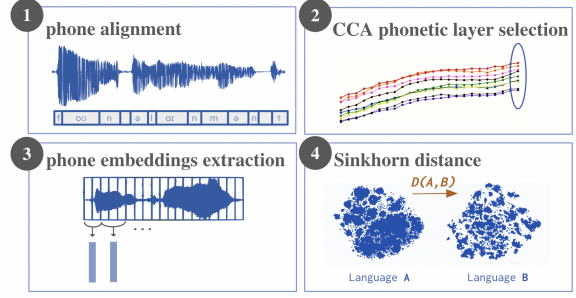


Figure 6: Our method for calculating *phonetic similarity* via Sinkhorn distance between contextualized phone embeddings extracted from the CCA-identified phonetic layer of the self-supervised multilingual speech model.

3. Phone embedding extraction. For each language we extract contextualized phone embeddings at the phonetic layer by mean-pooling the segment frames produced by MFA. The result is a per-language set of phone embeddings whose distribution we treat as a phonetic profile of the language in the model.

4. Sinkhorn distance. For each pair of languages (A, B) within a model, we compute the Sinkhorn approximation of the Wasserstein distance between their phone-embedding distributions (Cuturi, 2013) via the Python Optimal Transport library:

$$W_\varepsilon(A, B) = \min_{\pi \in \Pi(A, B)} \left[\sum_{i,j} \pi_{ij} \|a_i - b_j\|^2 + \varepsilon \sum_{i,j} \pi_{ij} \log \pi_{ij} \right],$$

where $\Pi(A, B)$ is the set of couplings between the empirical distributions of phone embeddings in A and B . This yields a 10×10 phonetic distance matrix per model (Figure 4). As a label-only baseline we also compute a normalized weighted Jaccard similarity over the *categorical* IPA phone-label distributions (Moulton and Jiang, 2018; Travieso et al., 2024) discussed in Appendix E.

B Layer-Weighting Methods Comparison

Table 2 reports mean accuracies for all four layer-weighting methods (see Section 3, Methodology). *Amplified weights* results in the highest within-language and foreign-language accuracy for both models.

C Per-Language Spearman Correlations

We additionally provide per-language Spearman correlations between phonetic distance and cross-

		Uniform	Inform.	Amp.	Amp-All
XLS-R	within	0.920	0.920	0.959	0.957
	foreign	0.855	0.854	0.861	0.848
	overall	0.861	0.861	0.871	0.859
WavLabLM	within	0.935	0.939	0.955	0.954
	foreign	0.696	0.683	0.789	0.759
	overall	0.720	0.709	0.806	0.778

Table 2: Mean speaker verification accuracies across layer-weighting methods. **Bold**: highest in row.

lingual speaker-verification accuracy from the *perspective of the training language* in Table 3. For the test-language perspective reported in the main results, see Table 1.

Train	XLS-R		WavLabLM	
	r_s	p	r_s	p
bg	-0.07	0.86	-0.60	0.07
cz	-0.49	0.15	-0.48	0.16
de	-0.14	0.70	-0.44	0.20
fr	-0.40	0.25	-0.19	0.60
hr	-0.58	0.08	-0.36	0.31
pl	-0.39	0.26	-0.73	0.016
pt	-0.58	0.08	-0.56	0.09
ru	-0.05	0.89	-0.60	0.07
uk	-0.38	0.28	-0.52	0.13
vi	-0.79	0.006	-0.15	0.68

Table 3: Per-train-language Spearman correlations (r_s), fixing the training language, between phonetic distance and cross-lingual speaker verification accuracy across all ten test languages. **Bold**: $p < 0.05$.

D MFA Alignment Statistics

Table 4 summarizes the per-language phone segment counts and unique phoneme label counts produced by the Montreal Forced Aligner on the 10-speaker $\times$ 50-utterance training subset of Global-Phone. Vietnamese has the smallest ratio of segment count to label inventory, which may account for its relatively low CCA scores (Figure 1).

E Phone Label Similarity

As a symbolic baseline complementing our embedding-based phonetic distance, we computed normalized weighted Jaccard similarity over phone label inventories per language pair (Moulton and Jiang, 2018; Ertl, 2020; Travieso et al., 2024). However, the measure of phone inventory overlap, or phone-label distribution (see Figure 7), only compares a categorical similarity based on a perceptual dimension, namely phones. Using phones with only one dimension of representation could prevent

Language	Phone segments	Unique labels
Bulgarian	40,773	46
Czech	41,862	44
German	35,782	52
French	51,214	42
Croatian	74,167	63
Polish	42,865	51
Portuguese	45,262	40
Russian	50,319	79
Ukrainian	14,884	56
Vietnamese	15,789	166

Table 4: Phone-segment and unique-phone-label counts per language after MFA alignment.

a more fine-grained comparison of the contextualized phone *embeddings*, one that makes use of underlying speech information therein which is also accessible to the model.

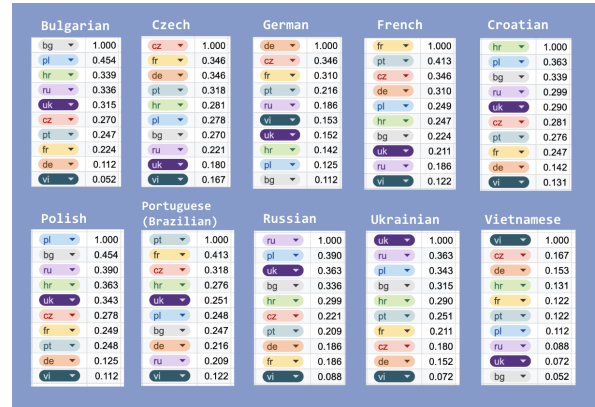


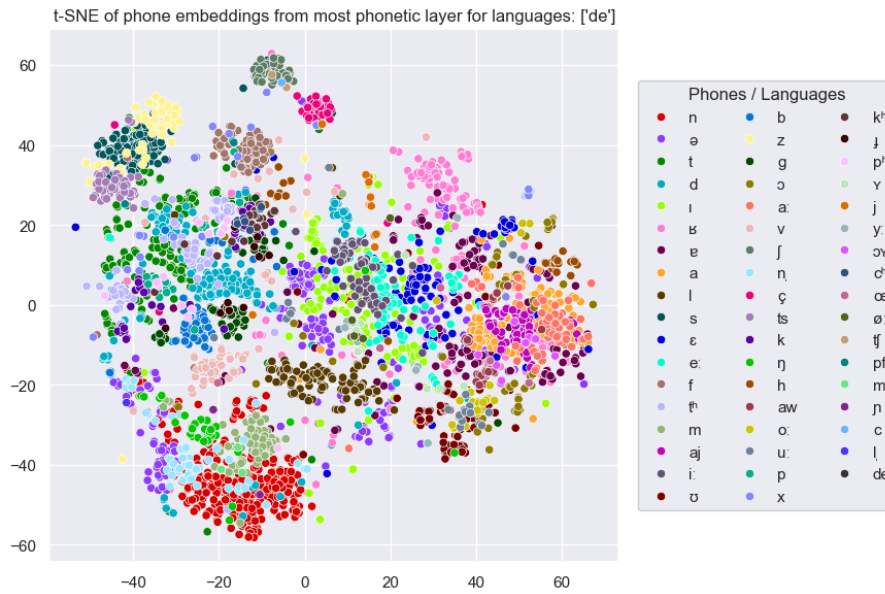
Figure 7: *Phone-label* distribution similarity rankings, using normalized weighted Jaccard similarity. For each language, the most similar languages are listed from top to bottom with scores adjacent to each language pairing.

F t-SNE Visualization: Phonetic Layer Phone Embeddings

t-SNE projections of phone embeddings extracted from the phonetic layer confirm that phones cluster meaningfully by articulatory class within each language, and that the degree of cross-language cluster overlap corresponds qualitatively to phonetic distance scores. Figure 8 provides an example.

G t-SNE Visualization: Embedding Distributions Compared

t-SNE projections of phone embeddings, as seen as overlapping distributions, are depicted in one example each for XLS-R in Figure 9a and WavLabLM in Figure 9b.



H Per-Language Layer-Weight Distributions

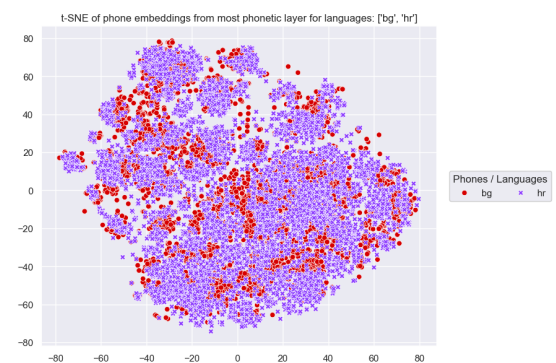
The weight distributions calculated across layers for each of the ten languages are summarized in Figure 10.

I Language Families

Speaker data are selected from ten languages from various language families, as depicted in the phylogenetic tree in Figure 11. Relations represent estimated evolutionary relatedness, derived from comparative studies of language change and regular sound correspondences.

J Layer-wise Speaker and Phonetic Information

We compare in Figure 12 the layer-wise speaker and phonetic information mapped to XLS-R and WavLabLM through experiments in speaker verification and CCA.



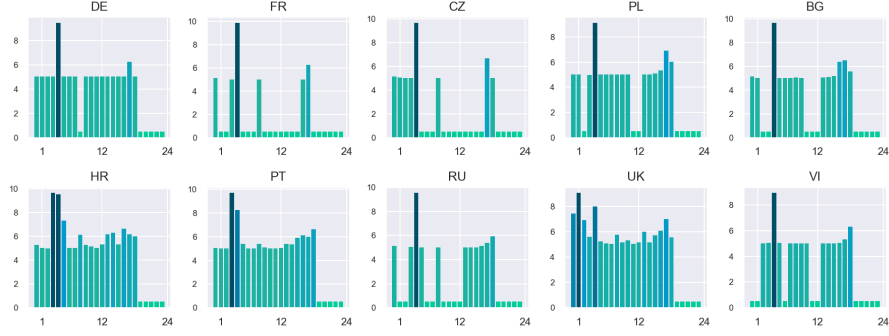
(a) t-SNE visualization of the phone embedding distributions for phylogenetically related languages Bulgarian and Croatian from XLS-R.



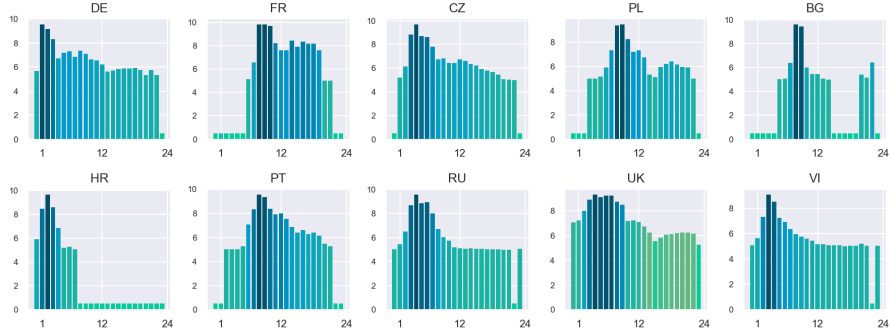
(b) t-SNE vis. of the phone embedding distributions for the Slavic language of Bulgarian and the Germanic language of German from WavLabLM.

Figure 9: t-SNE visualization of phone embedding distributions between language pairs, for BG-HR (from XLS-R, layer 18) and BG-DE (from WavLabLM, layer 24).

Per-Language Layer-Weight Distributions



(a) XLS-R layer weight distributions per language.



(b) WavLabLM layer weight distributions per language.

Figure 10: Per-language layer-weight distributions for models XLS-R and WavLabLM.

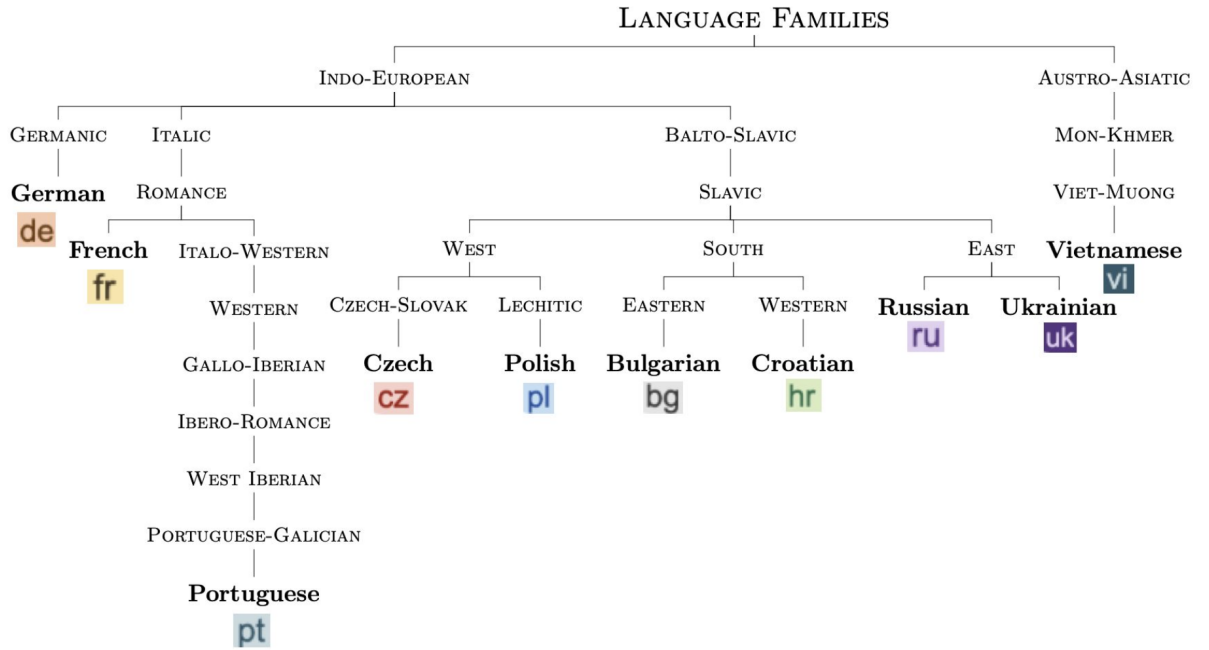


Figure 11: Language families of the sampled languages in a phylogenetic tree. Data retrieved from Ethnologue (28th ed.) (Eberhard et al., 2025).

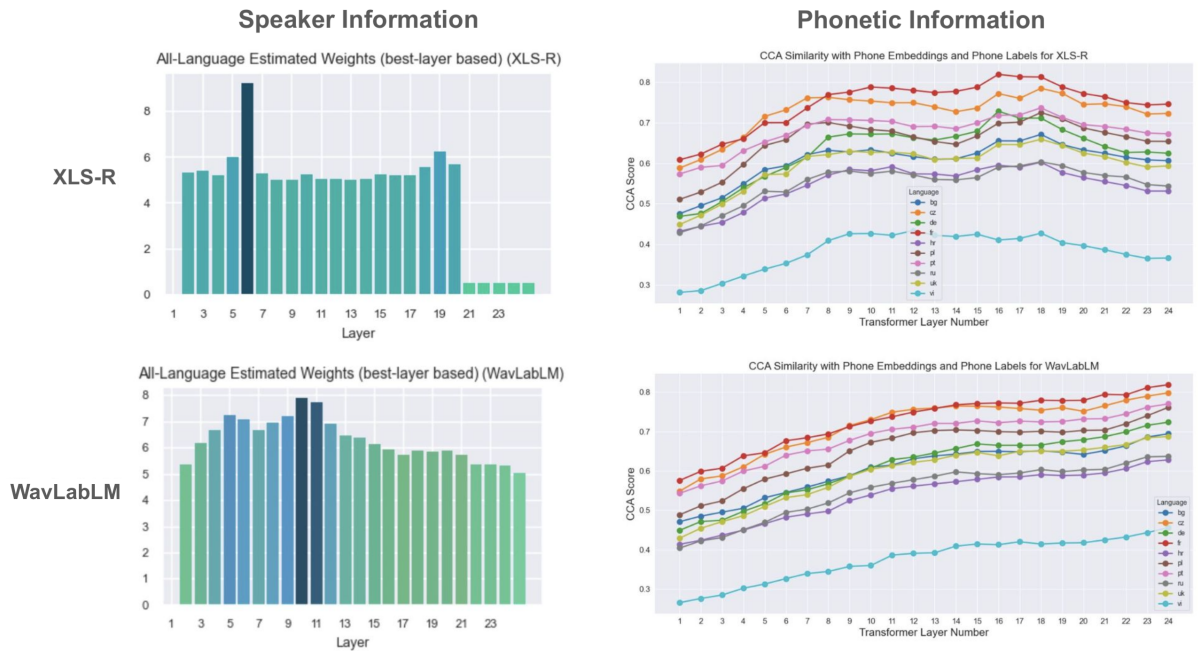


Figure 12: Comparison of layer-wise analyses of *speaker information* (left) and *phonetic information* (right) across the two models XLS-R and WavLabLM. Speaker information is derived from the layer-weight coefficients found through speaker verification probes, averaged at each layer across the 10 languages. Phonetic information is derived from canonical correlation analysis (CCA), which correlates contextualized phone embeddings with phone labels at each encoder layer.