Phoneme- and Word-Level Metrics Using Self-Supervised Speech Representations for Forced Alignment Evaluation

V.S.D.S. Mahesh Akavarapu[1], Michael Daniel[2], Gerhard Jäger[1]
[1]University of Tübingen, [2]University of Jena
mahesh.akavarapu@uni-tuebingen.de, misha.daniel@gmail.com,
gerhard.jaeger@uni-tuebingen.de

Abstract

Forced alignment evaluation typically requires manually annotated timestamps, limiting large-scale and multilingual analysis. We introduce two corpus-level metrics based on self-supervised (SSL) speech representations for reference-free forced alignment evaluation: Phoneme-Cluster Mutual Information (PCMI) and Word Acoustic Consistency Score (WACS). PCMI measures agreement between aligned phoneme labels and clusters induced from SSL-speech representations, while WACS measures consistency of repeated word realizations using dynamic time warping similarity between word representation sequences. Using both random and systematic perturbations, we show that PCMI and WACS degrade consistently under alignment perturbations. We further analyze the metrics across multiple alignment systems on 85 languages from FLEURS, validate them against manually annotated alignments from 45 languages in DoReCo, and evaluate them on two phonologically complex low-resource languages. The metrics effectively separate high- and low-quality alignments and correlate strongly with timestamp-based alignment quality measures. Our results demonstrate that SSL-speech representations enable scalable, reference-free forced alignment evaluation. The metrics are available as an open-source Python package at https://github.com/mahesh-ak/forced-aligner-metrics.

1 Introduction

Forced alignment maps speech waveforms to transcript units such as words or phonemes and finds applications in subtitle generation, speech corpus segmentation, acoustic-phonetic analysis, and speech synthesis. Recent years have seen increasing adoption of transformer-based speech models for alignment (Bain et al., 2023; Rastorgueva et al., 2023; Shi et al., 2026). However, evaluation of forced alignment systems still largely depends on manually annotated timestamps, which are expensive to obtain and predominantly available for English. Consequently, aligners are commonly evaluated either solely for English or against outputs of existing aligners such as Montreal Forced Aligner (MFA) (McAuliffe et al., 2017). This substantially limits progress toward multilingual forced alignment.

Emb.	Alignment model	#Langs	PCMI ↑	WACS ↑
MMS	Baseline	85	0.09 ± 0.02	0.02 ± 0.01
	MFA	82	0.25 ± 0.11	0.14 ± 0.08
	MMS-300m-IPA*	85	0.24 ± 0.03	0.19 ± 0.02
	Wav2Vec2-IPA*	85	0.24 ± 0.03	0.19 ± 0.03
	Qwen3-FA	10	-	0.21 ± 0.02
XLSR	Baseline	85	0.09 ± 0.02	0.02 ± 0.01
	MFA	82	0.23 ± 0.09	0.13 ± 0.07
	MMS-300m-IPA*	85	0.23 ± 0.03	0.17 ± 0.02
	Wav2Vec2-IPA*	85	0.22 ± 0.02	0.17 ± 0.02
	Qwen3-FA	10	-	0.18 ± 0.01

Table 1: Multilingual evaluation of forced alignment using PCMI and WACS across 85 languages from FLEURS dataset. **Emb.** denotes representation model used. *Trained in this work.

To address this gap, we propose two reference-free corpus-level evaluation metrics based on representations from self-supervised (SSL) speech models: *Phoneme-Cluster Mutual Information (PCMI)* and *Word Acoustic Consistency Score (WACS)*. PCMI measures consistency between aligned phoneme labels and emergent clusters in SSL-speech representations, while WACS measures acoustic consistency of word occurrences using dynamic time warping (Sakoe and Chiba, 1978) over representation sequences. The proposed metrics do not require gold timestamp annotations.

We analyze the behavior of the metrics under random perturbations of Buckeye alignments (Pitt et al., 2005). Additionally, we investigate systematic adversarial perturbations such as vowel and silence absorption to characterize the scope and limitations of the metrics. We find that both PCMI and WACS degrade consistently under increasing perturbation, while WACS remains relatively insen-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 5865–5892
October 24-29, 2026 ©2026 Association for Computational Linguistics

sitive to silence absorption due to the properties of dynamic time warping.

To study multilingual behavior, we train multilingual IPA-based phoneme recognition models on 85 FLEURS languages (Conneau et al., 2023) and generate alignments using connectionist temporal classifier (CTC) segmentation (Kürzinger et al., 2020). Comparing these alignments against MFA and evenly spaced phoneme baselines, the proposed metrics reveal a clear bimodal distribution for MFA alignments across languages, separating failed and successful alignments, while CTC-based aligners exhibit greater stability (Table 1). We additionally validate the metrics on 45 languages from the DoReCo corpus (Paschen et al., 2020; Seifart et al., 2024), demonstrating strong agreement with manually annotated word-level and phone-level alignments.

Finally, we validate the proposed metrics against manually annotated alignments for two low-resource as well as phonologically complex languages Archi and Rutul, using recent phoneme models and data for these languages (Akavarapu et al., 2026). These experiments further validate the robustness of PCMI and clarify the operational limitations of WACS under silence-related perturbations.

Our contributions are threefold:

1. We propose PCMI and WACS, reference-free metrics for forced alignment evaluation based on SSL-speech representations.

2. We study the robustness and adversarial behavior of the metrics under both random and systematic perturbations.

3. We conduct large-scale multilingual analysis across 85 FLEURS languages, validate against manually annotated alignments from 45 DoReCo languages and two phonologically complex low-resource languages. We release multilingual phoneme recognition models used for alignment as a byproduct.[1]

2 Related Work

Forced alignment quality is traditionally evaluated against manually annotated timestamps using boundary displacement or overlap-based measures. Prior work commonly reported the proportion of boundaries within fixed tolerances such as 10–50 ms from manual annotations (Kelley et al., 2024; Rousso et al., 2024), with standard thresholds for accurate alignment ranging from 20 ms (Hosom, 2009) upto 200 ms (Bain et al., 2023; Rastorgueva et al., 2023). Other work evaluates overlap between manually annotated and automatically aligned intervals using measures such as overlap rate (Gonzalez et al., 2018) or average temporal shift between aligned intervals (Shi et al., 2022). These approaches require gold timestamp annotations and are difficult to scale to multilingual settings.

Information-theoretic measures such as mutual information and cluster purity have previously been used in speech processing for phoneme-feature selection (Omar et al., 2002), speaker clustering (Cohen and Lapidot, 2021), and speaker recognition evaluation using normalized mutual information (Mridha et al., 2021). Pointwise mutual information between aligned strings of sound classes has also been employed to derive similarity scores for phonetic sequence alignment and phylogenetic inference (Jäger, 2013, 2015). Dynamic time warping based similarity measures are also widely used in spoken term detection and retrieval (Sakoe and Chiba, 1978; Tejedor et al., 2012). Motivated by these approaches, we employ mutual information and dynamic time warping to define reference-free metrics for forced alignment evaluation.

3 Problem Formulation

We consider a transcript sequence $x_1, x_2, \ldots, x_L$ where each x_l corresponds to a word or phoneme token. Silence tokens (contiguous) may optionally be inserted between adjacent units as well as at the beginning and end of the utterance. An alignment for an utterance of duration T performed by an alignment model $\mathcal{A}$ is defined by a sequence of boundary times $0 = t_0 \leq t_1 \leq \cdots \leq t_L = T$ where token x_l occupies the interval $[t_{l-1}, t_l]$. Given gold boundary times $0 = t'_0 \leq t'_1 \leq \cdots \leq t'_L = T$ alignment quality is commonly measured using Average Accumulated Shift (AAS) i.e., temporal deviation between postulated and gold boundaries (Shi et al., 2022, 2026):

$$\mathrm{AAS} = \frac{1}{L} \sum_{l=1}^{L} |t_l - t'_l|.$$

Our goal is to introduce reference-free evaluation metrics whose behavior correlates with alignment

[1] https://hf.co/mahesh27/mms-300m-ipa-fleurs
https://hf.co/mahesh27/mms-300m-xsampa-doreco

quality in a manner similar to AAS, without requiring gold timestamp annotations. AAS may be computed with respect to either word-level or phoneme-level boundaries. In practice, these values differ by less than 5 ms across the datasets considered in this work. Larger discrepancies arise primarily under artificial perturbations that modify one level of annotation while leaving the other unchanged. Unless otherwise stated, we therefore report AAS as the maximum of the word-level and phoneme-level AAS values.

4 PCMI and WACS

4.1 Phoneme-Cluster Mutual Information (PCMI)

SSL-speech representations learned through predictive and clustering-based pretraining objectives (Baevski et al., 2020; Hsu et al., 2021) are known to encode phonetic structure, including articulation and phonological attributes (Cormac English et al., 2022; de la Fuente and Jurafsky, 2024). We exploit this property to evaluate forced alignments by measuring consistency between aligned phoneme labels and clusters induced from these representations.

Consider utterances $S_1, \ldots, S_N$, where each utterance S_n is associated with aligned phoneme tokens $p_1, \ldots, p_{L_n}$ including silences, together with predicted boundary times from an aligner $\mathcal{A}$ (§ 3). Given a self-supervised model $\mathcal{M}$, we extract frame-level representations $\mathbf{emb}_1, \ldots, \mathbf{emb}_F$ from a fixed layer over a sample of frames of size F and assign each frame the phoneme label corresponding to the aligned interval containing the frame timestamp. We then cluster the representations using K-Means to obtain cluster assignments $c_1, \ldots, c_F$. PCMI is defined as the normalized mutual information between phoneme labels P and representation clusters C:

$$\mathrm{PCMI} = \frac{I(P;C)}{\sqrt{H(P)H(C)}},$$

where H is entropy and I denotes mutual information given by:

$$I(P;C) = H(P) - H(P \mid C)$$

High PCMI indicates that the alignment induces phoneme assignments consistent with emergent phonetic structure in the representation space. Under sufficiently discriminative representations and approximately stationary phoneme realizations, increasing alignment boundary error introduces phoneme-label corruption near segment boundaries, which monotonically increases conditional entropy $H(P \mid C)$ and therefore decreases PCMI.

4.2 Word Acoustic Consistency Score (WACS)

SSL-speech models are known to encode substantial word-level information and support acoustic word discrimination tasks (Pasad et al., 2024; Meghanani and Hain, 2024). Such representations have also been used for speech alignment through dynamic time warping (Zhu et al., 2024; Sakoe and Chiba, 1978). Motivated by these observations, WACS measures whether repeated occurrences of the same aligned word exhibit greater acoustic consistency than unrelated word pairs.

For each aligned word occurrence, we extract the sequence of frame-level representations spanning its aligned interval, $\mathbf{e}_w = \{\mathbf{emb}_1, \ldots, \mathbf{emb}_{|w|}\}$. Given two word occurrences w_1 and w_2, we compute a similarity score $\mathrm{dtw}(\mathbf{e}_{w_1}, \mathbf{e}_{w_2})$ using dynamic time warping (DTW) with cosine similarity between frame-level representations (Appendix B).

Let $\mathcal{F}$ denote the set of all unique word forms in the corpus, and let Σ_f^+ denote the set of all word occurrences having word form $f \in \mathcal{F}$. Similarly, let Σ_f^- denote the set of word occurrences that have non-identical word form as f (negative samples, $\Sigma_f^+ \cap \Sigma_f^- = \varnothing$). WACS is defined as:

$$\mathrm{WACS} = \mathbb{E}_{f \sim \mathcal{F},\, (w_1,w_2) \sim \Sigma_f^+ \times \Sigma_f^+}[\mathrm{dtw}(\mathbf{e}_{w_1}, \mathbf{e}_{w_2})]$$

$$-\mathbb{E}_{f \sim \mathcal{F},\, (w_1,w_2) \sim \Sigma_f^+ \times \Sigma_f^-}[\mathrm{dtw}(\mathbf{e}_{w_1}, \mathbf{e}_{w_2})]$$

Positive pairs are sampled from distinct occurrences of the same word form. High WACS indicates strong acoustic consistency within identical word forms compared to non-identical pairings.

5 Experiments

We evaluate the proposed metrics using SSL-speech representations from MMS (Pratap et al., 2024) and XLSR (Conneau et al., 2021). We additionally compare against conventional MFCCs[2] in the perturbation experiments as a classical acoustic baseline. We use representations from layer 15 selected based on the perturbation analysis described below. For computational efficiency, PCMI and WACS are computed on random subsets of utterances, where each subset contains 50 utterances in

[2]Computed using a 25 ms window, 10 ms hop size, and 20 coefficients.

the case of PCMI and 200 utterances in the case of WACS. Reported perturbation results are averaged over 5 independent samples. Additional details are provided in Appendix A.

5.1 Perturbation Analysis

We study the behavior of PCMI and WACS under controlled perturbations of manually annotated alignments from the Buckeye corpus (Pitt et al., 2005), which contains word- and phoneme-level annotations for 40 speakers. We segment the recordings into utterances of duration 3–20 s, yielding 2935 utterances totaling about 7.5 hours.

Random perturbations. To simulate progressively degraded alignments, we perturb word boundaries by adding Gaussian noise with target standard deviation σ. Boundary constraints are then reimposed to avoid overlaps and intervals with zero-duration, after which phoneme boundaries are warped proportionally within each word interval. Although perturbations were generated with target scales ranging from 50–2000 ms, the resulting average accumulated shift (AAS) after constraint enforcement ranged from approximately 45–210 ms. Figure 1 reports the resulting behavior of PCMI and WACS using MMS-300M and MFCC representations.

Both PCMI and WACS degrade consistently with increasing perturbation severity. The metrics decrease approximately linearly at lower perturbation levels before gradually plateauing at larger shifts. While the overall degradation trends remain similar across MMS-300M and MFCC representations, notable differences emerge in score ranges. PCMI retains a comparatively broad operating range under MFCC features, suggesting that local phonetic structure remains sufficiently recoverable from conventional spectral representations. In contrast, WACS values under MFCC representations occupy a substantially narrower range, whereas self-supervised MMS representations produce considerably stronger separations. This suggests that WACS particularly benefits from the richer lexical and contextual structure captured by SSL-speech representations.

We additionally analyze sensitivity across network layers using representations extracted from hidden states of different transformer blocks of MMS and XLSR. As shown in Figure 2, middle layers exhibit the clearest separation between low- and high-AAS alignments for both PCMI and WACS, whereas early layers may be dominated by local acoustic variation and later layers may become increasingly contextualized. This behavior is consistent with prior analyses showing that intermediate layers of SSL-speech models encode enhanced phonetic structure (Cormac English et al., 2022; de la Fuente and Jurafsky, 2024).

Across both metrics, layers approximately between 5 and 16 show the greatest perturbation sensitivity. We therefore select layer 15 for all subsequent experiments by maximizing the separation between clean alignments and the most severely perturbed alignments.

We further analyze the sensitivity of PCMI to the number of K-means clusters used for representation quantization. The separation between clean and severely perturbed alignments remains stable, varying only from approximately 0.26 to 0.25 for MMS and from 0.24 to 0.22 for XLSR representations across $25 \leq \text{n_cluster} \leq 100$ (see Appendix Figure 4). We therefore select $\text{n_cluster} = 50$ for all subsequent experiments, which approximately matches the average phoneme inventory (labels) (Appendix Table 15).

Systematic perturbations. Random perturbations alone do not capture structured alignment errors that may interact favorably with representation-based metrics. We therefore construct adversarial perturbations motivated by phonological regularities and DTW behavior.

First, we introduce vowel absorption perturbations by merging vowel-consonant transitions, motivated by the tendency of neighboring phonetic segments to exhibit similar acoustic structure. These perturbations primarily affect PCMI by altering phoneme assignments while preserving coarse acoustic continuity.

Second, we construct silence absorption perturbations by merging preceding silence intervals into neighboring words. Since DTW is relatively insensitive to leading or trailing silence, this perturbation particularly targets WACS.

The resulting perturbations achieve moderate AAS while producing substantially weaker degradation than random perturbations at comparable shift levels. In particular, silence absorption emerges as a prominent adversarial failure mode for WACS.

Nevertheless, both metrics remain informative under realistic alignment deviations: PCMI continues to reflect phonetic consistency under moderate

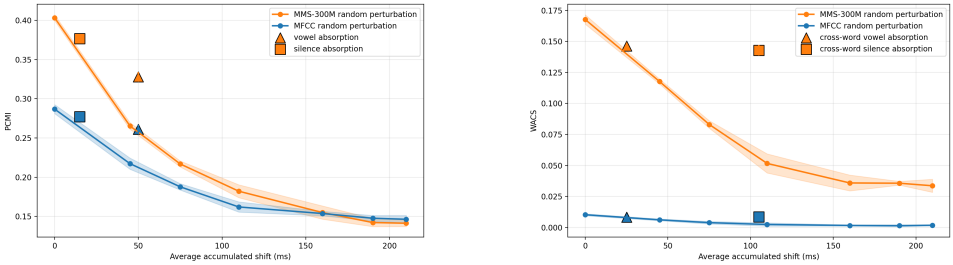

Figure 1: Behavior of PCMI and WACS under random and systematic perturbations of Buckeye alignments using MMS-300M and MFCC representations. Curves show mean scores over 5 random perturbation samples, with shaded regions denoting standard deviation.

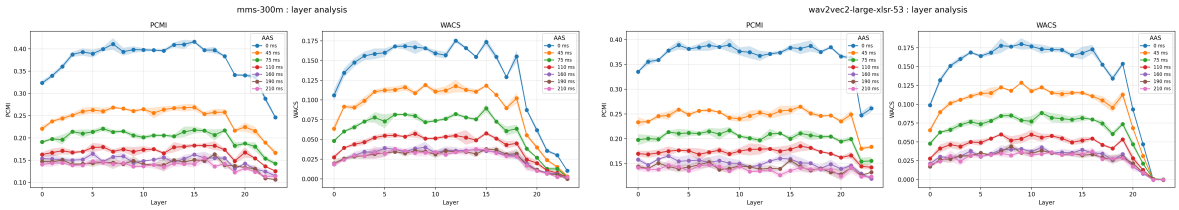

Figure 2: Layer-wise behavior of PCMI and WACS under increasing perturbation severity using MMS (left) and XLSR (right) representations. Middle transformer layers provide the strongest separation between clean and perturbed alignments, while later layers exhibit reduced sensitivity.

perturbations, while WACS remains effective when silence regions are excluded during downstream segment extraction. Moreover, further degenerate merging of phoneme labels cannot artificially inflate PCMI, since PCMI $\rightarrow 0$ as the phoneme-label entropy $H(P) \rightarrow 0$ (see Appendix C). Likewise, if the representation model collapses all word occurrences to identical embeddings, positive and negative DTW similarities become indistinguishable (see §4.2), causing WACS $\rightarrow 0$.

6 Multilingual Evaluation

We analyze PCMI and WACS in a large-scale multilingual setting, which constitutes the primary target application of the proposed metrics. The multilingual experiments reported in this section use layer-15 representations from MMS and XLSR.

6.1 Multilingual Alignment Models

Dataset To analyze the proposed metrics in a multilingual setting, we use the FLEURS dataset (Conneau et al., 2023), which provides speech and transcriptions spanning a diverse set of languages and writing systems. Since the phoneme-based evaluation framework requires a unified cross-lingual representation, all transcripts were converted into International Phonetic Alphabet (IPA) format. To complement this large-scale evaluation, we additionally use the DoReCo corpus (Seifart et al., 2024), a collection of predominantly low-resource and endangered languages with manually annotated word- and phoneme-level timestamps. Unlike FLEURS, DoReCo provides gold alignment annotations, enabling direct comparison against timestamp-based evaluation measures such as AAS. The corpus includes phoneme-level transcriptions in X-SAMPA format, which are retained without further conversion throughout our experiments.

FLEURS		DoReCo			
Family	**#**	**Family**	**#**	**Family**	**#**
Indo-European	41	Austronesian	7	Mayan	1
Niger-Congo	9	Sino-Tibetan	4	Mixe-Zoque	1
Afroasiatic	7	Indo-European	3	Nilo-Saharan	1
Austronesian	5	Niger-Congo	3	Pama-Nyungan	1
Turkic	5	Afroasiatic	2	Pano-Tacanan	1
Dravidian	4	Arawakan	2	Trans-New Guinea	1
Sino-Tibetan	3	Austroasiatic	2	Tungusic	1
Uralic	3	Nakh-Daghestanian	2	Tuu	1
Austroasiatic	2	Turkic	2	Uralic	1
Kra-Dai	2	Algic	1	Yam	1
Japonic	1	Boran	1	Mixed Language	1
Kartvelian	1	Chibchan	1	Isolate	1
Koreanic	1	Gunwinyguan	1		
Mongolic	1	Kartvelian	1		
		Koreanic	1		

Table 2: Language family distributions in the multilingual evaluation datasets — FLEURS (85 languages) and DoReCo (45 languages).

A particular challenge in FLEURS is the presence of numerals, and mixed symbolic forms within transcripts. These are often unsuitable for direct grapheme-to-phoneme conversion because

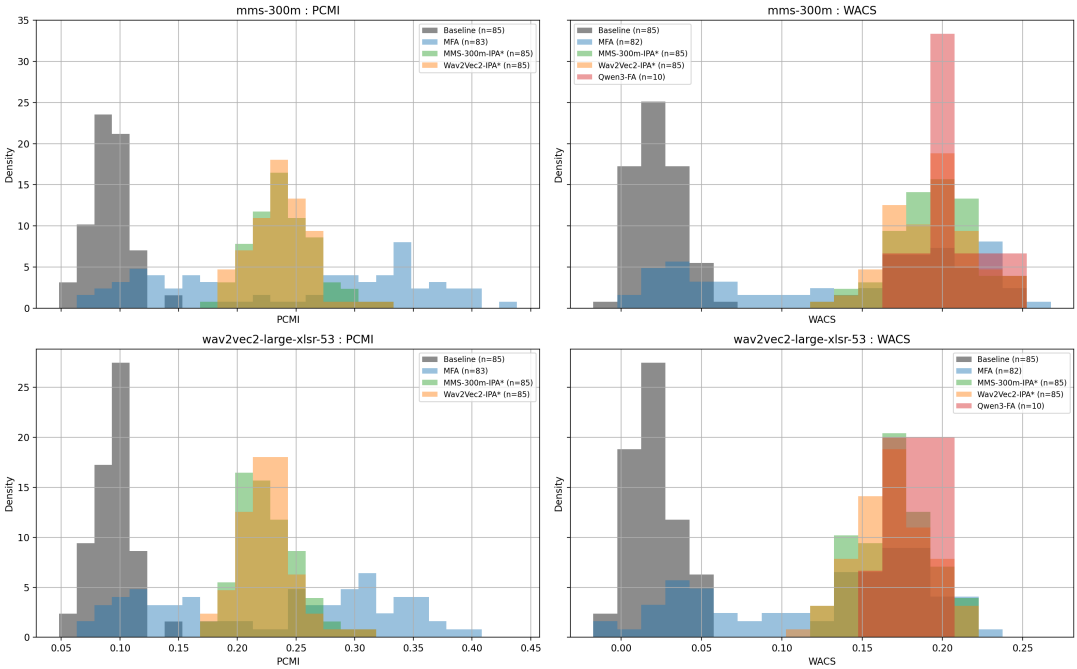

Figure 3: Distribution of PCMI (left) and WACS (right) across multilingual alignments on FLEURS. The top row uses MMS representations while the bottom row uses XLSR representations. MFA exhibits a clear bimodal structure corresponding to successful and failed alignments, while CTC-based aligners produce substantially more stable distributions across languages. * Trained in this work.

their pronunciation depends on linguistic context and language-specific conventions. For example, the token "2019" may be realized as "twenty nineteen", "two thousand nineteen". We therefore normalize numerals into fully written forms prior to phonemization using the `num2words` package[3], which we further extended to support several languages missing from the original implementation. We additionally apply a simple heuristic where values between 1200 and 2050 are interpreted as years, while all remaining are treated as cardinal numbers. Grapheme-to-phoneme conversion for FLEURS data was subsequently performed using Epitran (Mortensen et al., 2018) and XPF (Cohen Priva et al., 2021), producing IPA transcriptions.

The language set spans a broad range of families, summarized in Table 2, while detailed language-wise statistics and metadata are provided in Table 6 (FLEURS) & Table 7 (DoReCo).

Phoneme Models Using these transcriptions, we finetune multilingual phoneme recognition models MMS-300M-IPA and Wav2Vec2-IPA, based on MMS-300M (Pratap et al., 2024) and Wav2Vec2-Phoneme (Xu et al., 2022) respectively, for FLEURS, and MMS-300M-DORECO for DoReCo, based on MMS-300M. All models use language-specific adapters. Additional training details are provided in Appendix D. Alignments are generated using CTC segmentation (Kürzinger et al., 2020).

We compare the resulting alignments on FLEURS against several baselines and alignment systems: (i) MFA (McAuliffe et al., 2017) acoustic models trained independently for each language, (ii) an evenly spaced phoneme baseline obtained by uniformly partitioning non-silent regions (with only utterance-initial and utterance-final silences removed), and (iii) Qwen3-ForcedAligner-0.6B (Qwen3-FA) (Shi et al., 2026), evaluated on the subset of 10 supported languages for which word-level alignments are available.

For evaluation on DoReCo, in addition to the evenly spaced baseline, we include a silence-absorbing post-processed variant of MMS-300M-DORECO (denoted by the suffix '-SIL'). This variant is included because gold alignments are available for DoReCo, allowing AAS to be reported, and AAS is particularly sensitive to silence placement. We do not evaluate MFA on DoReCo, as the corpus provides only approximately two hours of speech per language on average, making language-specific MFA training impractical.

[3] https://github.com/savoirfairelinux/num2words

Results The multilingual distributions of PCMI and WACS on FLEURS are shown in Figure 3. The MFA exhibits pronounced bimodal behavior in both metrics; manual inspection reveals that the lower-scoring modes predominantly correspond to failed alignments, while the higher-scoring modes correspond to successful ones. In contrast, CTC-based aligners produce substantially more stable distributions across languages despite achieving slightly lower peak PCMI values than the successful MFA cluster.

Failure rates further support this observation. MFA alignment training or decoding failed for approximately 12k out of 65k utterances ($\sim$19%), whereas the proposed IPA-based CTC aligners produced only 11 word-level failures overall. The evenly spaced phoneme baseline remains competitive, indicating that coarse temporal consistency alone can yield moderate acoustic agreement. The distributions of metrics on the DoReCo corpus, which additionally provides manually annotated word- and phone-level boundaries, are provided in Appendix Figure 5. Manually annotated gold alignments achieve average scores of approximately 0.33 PCMI and 0.13 WACS, providing practical reference values for high-quality alignments.

The metrics exhibit similar behavior across MMS and XLSR representations, with MMS providing slightly stronger separation overall. Notably, XLSR remains effective despite being pretrained on only 53 languages—with many of the evaluated FLEURS languages and most of DoReCo absent from its training data—whereas MMS pretraining substantially overlaps with the FLEURS language inventory, yet rarely with DoReCo. This suggests that the proposed metrics are reasonably robust across different SSL representation models.

Aggregated multilingual statistics, including cluster purity, vocabulary sizes, pair counts, entropy measures, and score variances, are summarized in Table 15. Across languages and embedding models, normalized cluster entropy consistently remains high (approximately 0.95–0.99), indicating effective utilization of the cluster inventory by the representation space. Normalized phoneme-label entropy also remains relatively high (roughly 0.82–0.95), reflecting substantial phonetic diversity across the evaluated languages. MFA exhibits substantially larger variance in these statistics as well as in PCMI and WACS scores, consistent with its observed bimodal failure behavior. In contrast, CTC-based aligners show considerably tighter distributions.

Embedding	Metric Pair	Pearson's r	p-value
MMS	PCMI vs AAS	-0.7769	$< 0.001^{***}$
	WACS vs AAS	-0.6257	$< 0.001^{***}$
XLSR	PCMI vs AAS	-0.7761	$< 0.001^{***}$
	WACS vs AAS	-0.6686	$< 0.001^{***}$

Table 3: Pearson correlation between reference-free metrics (PCMI, WACS) and alignment quality measured by AAS on the 45-language DoReCo evaluation set. Lower AAS indicates better alignment quality.

Language-wise alignment metrics are reported in Appendix Tables 10, 11, 12, 13 and 14. Language-wise phoneme recognition performance in terms of word and character error rates (WER, CER) is reported in Appendix Tables 8 and 9.

Correlations with AAS The availability of manually annotated word- and phone-level boundaries in DoReCo additionally allows direct comparison of the proposed reference-free metrics against Average Accumulated Shift (AAS), a timestamp-based measure of alignment quality. Table 3 reports Pearson correlations between PCMI/WACS and AAS across 45 languages and four alignment conditions (gold, baseline, MMS-300M-DORECO, and MMS-300M-DORECO-SIL), yielding 180 alignment sets in total.

Both PCMI and WACS exhibit strong negative correlations with AAS across MMS and XLSR representations, indicating that higher metric scores consistently correspond to lower alignment error. PCMI shows the strongest relationship overall ($r \sim -0.78$), while WACS provides complementary evidence with correlations between $r = -0.62$ and -0.67. These results provide large-scale multilingual validation against manually annotated timestamps and further support the effectiveness of the proposed metrics as reference-free indicators of alignment quality.

Phoneme recognition and alignment quality We further analyze the relationship between phoneme recognition performance and alignment quality across languages. As shown in Table 4, character error rate (CER) exhibits consistent moderate negative correlations with alignment metrics, with correlation magnitudes of approximately -0.5 with WACS and $-0.3 \sim -0.5$ with PCMI across embedding and alignment models. This indicates that improvements in phoneme recognition quality are generally associated with improved alignment

WER Correlations

Aligner	Embedding	PCMI	WACS
MMS-300M-IPA*	MMS	-0.161 / -0.137	-0.294 / -0.308
	XLSR	-0.092 / -0.050	-0.355 / -0.363
Wav2Vec2-IPA*	MMS	-0.008 / -0.073	-0.027 / -0.111
	XLSR	-0.001 / -0.108	-0.152 / -0.212
MMS-300M-DORECO*	MMS	-0.200 / -0.225	-0.259 / -0.237
	XLSR	-0.225 / -0.214	-0.261 / -0.230

CER Correlations

Aligner	Embedding	PCMI	WACS
MMS-300M-IPA*	MMS	-0.351 / -0.334	-0.529 / -0.533
	XLSR	-0.263 / -0.226	-0.538 / -0.552
Wav2Vec2-IPA*	MMS	-0.398 / -0.383	-0.438 / -0.441
	XLSR	-0.385 / -0.413	-0.475 / -0.481
MMS-300M-DORECO*	MMS	-0.562 / -0.616	-0.459 / -0.318
	XLSR	-0.547 / -0.552	-0.483 / -0.316

Table 4: Correlation between ASR error rates (WER, CER) and alignment metrics (PCMI, WACS). Values are Pearson/Spearman correlations. * trained in this work

quality, although the relationship is not strict.

WER shows a similar but weaker trend, with moderate negative correlation with WACS for MMS-300M-IPA (~ -0.3), while remaining weak for several configurations. This suggests that character-level evaluation is a more reliable predictor of alignment quality than word-level error rates in this setting.

As reported in Appendix Tables 8 and 9, WER values are generally high across languages, particularly for Wav2Vec2-IPA models and for DoReCo corpus, partly reflecting the absence of explicit language modeling. Despite this, alignment quality remains relatively robust, indicating that forced alignment can greatly compensate for phoneme recognition noise.

Overall, these results suggest a consistent but moderate coupling between phoneme recognition accuracy and alignment quality, with stronger and more stable effects observed for CER compared to WER. Together with the strong negative correlations between PCMI/WACS and AAS on DoReCo (Table 3), these findings support the validity of the proposed metrics as indicators of alignment quality rather than mere proxies for phoneme recognition performance.

6.2 Two phonologically complex languages

The proposed metrics are further evaluated on two endangered East Caucasian languages: Archi (Kibrik et al., 2007) and Kina Rutul (Alekseeva et al., 2024), using manually annotated word-level alignments produced by a linguistically trained annotator, the second author of this work. Speech data and language-specific IPA phoneme recognition models are obtained from Akavarapu et al. (2026). These CTC-based IPA models were trained separately for Archi and Rutul by the original authors and are used directly in our experiments. Each language has only approximately 45–75 minutes of available training audio in the current setup, along with around 7 minutes of test data (100 test utterances in Archi, 90 in Rutul), making conventional MFA training impractical.

Results are shown in Table 5. Despite the phonological complexity and limited supervision, the alignments after trailing-silence removal achieve AAS values roughly between 45–70 ms, which is competitive given the 20 ms frame resolution.

Manual inspection additionally revealed systematic silence absorption in several alignments. We therefore apply the silence-correction postprocessing step after CTC segmentation (denoted CTC-IPA-● - SIL in Table 5). This procedure removes approximately 44 seconds of excess silence across the Archi dataset, reducing AAS from 88.5 ms to 46.3 ms, while producing only minor changes in WACS and PCMI. In contrast, only around 5 seconds of silence are removed for Rutul, leading to comparatively smaller AAS changes. This difference between Rutul and Archi is likely related to the higher speech rate in the Rutul recordings, which consist primarily of spontaneous narratives, compared to the Archi recordings, which are prepared readings of written texts.

These observations are consistent with the perturbation experiments in §5.1, where WACS was found to be relatively insensitive to silence-related errors compared to AAS. Consequently, while the proposed metrics generally track alignment quality well, they do not penalize silence absorption as strongly as the timestamp-based measure AAS.

Interestingly, the evenly spaced baseline performs relatively well on Archi compared to Rutul, particularly under PCMI which may be attributed to the well-articulated nature of the Archi read-out recordings. A similar phenomenon is observed for a small number of languages in DoReCo (Appendix Figure 5). Importantly, this does not indicate a weakness of the proposed metrics, since these cases also exhibit correspondingly low AAS values, suggesting that the baseline alignments are

Language	Aligner	AAS↓	MMS		XLSR	
			PCMI↑	WACS↑	PCMI↑	WACS↑
Archi	Gold	0.0	–	0.225	–	0.220
	CTC-IPA-Archi	88.5	0.306	0.203	0.307	0.194
	CTC-IPA-Archi - SIL	46.3	0.319	0.215	0.330	0.215
	Baseline	85.8	0.204	0.112	0.203	0.102
Rutul	Gold	0.0	–	0.182	–	0.182
	CTC-IPA-Rutul	68.0	0.292	0.172	0.278	0.173
	CTC-IPA-Rutul - SIL	61.9	0.292	0.175	0.277	0.172
	Baseline	180.6	0.126	0.041	0.127	0.040

Table 5: Evaluation against manually annotated alignments for Archi and Rutul. Lower AAS is better, while higher PCMI and WACS are better.

genuinely competitive rather than being incorrectly favored by the metrics.

Systematic errors Beyond silence absorption, expert analysis also identified smaller but systematic alignment errors, such as the merging of the release of the consonant into the following phoneme onsets (e.g. si**k'ʲ**→wuʁat'i, na**t͡s'**→jaraʁ, ja**k**→uwk'uli) and of the last segment of the vowel into the onset of the following consonant (eg. uq'ulʲ**a**→zamanindexa, minowd**ɨ**→ɢuʃimakan). As also suggested by the perturbation results in §5.1, the proposed metrics are less sensitive to such fine-grained systematic mergers. These effects likely contribute to residual differences in AAS that are not explained by silence removal alone.

7 Discussion

Although MFA achieves high PCMI scores when alignment succeeds (Appendix Table 10), it exhibits a substantial failure rate of approximately 19%, in contrast to the near-complete robustness of CTC-based models (§6.1). Moreover, MFA training is often impractical in highly under-resourced settings (§6.2). These limitations suggest that the common practice of reporting alignment performance primarily as deviations from MFA-based alignments is not reliable in low-resource scenarios. The proposed metrics, PCMI and WACS, provide a more stable alternative for such settings.

Furthermore, AAS is highly sensitive to silence absorption effects, whereas PCMI and WACS remain comparatively robust to such perturbations (§5.1, §6.2). This distinction should be viewed as a limitation of AAS rather than a disadvantage of the proposed metrics, particularly in downstream applications such as retrieval, where excessive silence can distort temporal alignment quality. In fact, manual annotations in DoReCo often do not precisely trim short silence intervals, suggesting that fine-grained silence placement is of secondary importance relative to the alignment of linguistic units themselves. Since silence can be handled deterministically through simple postprocessing or segmentation heuristics, sensitivity to silence may introduce unnecessary variance without reflecting true alignment quality.

Similarly, AAS is also expected to be sensitive to differences in speech rate, potentially complicating direct cross-lingual comparisons. In contrast, PCMI and WACS are comparatively rate-invariant by design. Overall, the proposed metrics provide scalable and reliable alternatives for multilingual forced alignment evaluation without requiring manually annotated timestamps.

8 Conclusions

We introduced two reference-free corpus-level metrics for forced alignment evaluation based on SSL-speech representations: Phoneme-Cluster Mutual Information (PCMI) and Word Acoustic Consistency Score (WACS). Across both synthetic perturbations and multilingual evaluation on 85 FLEURS, 45 DoReCo and two phonologically complex languages, the metrics exhibit strong negative correlation with manual timestamp-based alignment quality measures. We further analyzed systematic effects such as silence absorption and phonological mergers, showing complementary behavior between PCMI, WACS and AAS. Unlike approaches relying on tediously manually annotated timestamps or MFA-derived references, the proposed framework enables scalable evaluation in multilingual and low-resource settings. More broadly, these results suggest that SSL-speech representations can support large-scale multilingual forced alignment evaluation and facilitate the development of alignment resources for under-resourced languages.

Limitations

The proposed metrics are intended as scalable surrogate measures of alignment quality rather than direct replacements for manually annotated timestamps. Nevertheless, they provide an efficient mechanism for screening model-generated alignments before manual inspection and correction. Moreover, the proposed measures operate at the corpus level and do not directly provide diagnostic information for individual utterances or local alignment errors.

Additionally, although the perturbation experiments cover both random and systematic alignment errors, certain linguistically fine-grained phenomena—such as systematic phonological mergers or other coarticulatory effects—may not be fully captured by the proposed metrics.

Finally, while we evaluate alignments across 132 languages and multiple alignment systems, the experiments are restricted to datasets with available transcriptions and phoneme conversion pipelines. Languages without reliable grapheme-to-phoneme resources remain comparatively underexplored. We additionally do not ablate the effect of different G2P systems, since for most evaluated languages no viable alternative phonemization resources are available.

Ethics Statement

All datasets used in this work were obtained from publicly available sources and used in accordance with their respective licenses and terms of use. To the best of our knowledge and based on the available dataset documentation, the data do not contain personally identifying information or intentionally offensive content. The manually produced word-level alignment annotations likewise do not present foreseeable ethical concerns, as they consist solely of temporal boundary labels over already publicly available speech recordings. Large language models and AI-assisted coding tools, including ChatGPT and GitHub Copilot, were used to assist with portions of code development, and manuscript refinement. All generated outputs were carefully reviewed and verified by the authors. The core research ideas, experimental design, analyses, and conclusions are the authors' own, although selected refinements suggested by these tools were incorporated where appropriate.

Acknowledgments

Mahesh Akavarapu received funding from Volkswagen Foundation under the Phylomilia project within the Pioneering Projects funding line. We thank the anonymous reviewers, whose feedback significantly improved the paper.

References

Emily Ahn and Eleanor Chodroff. 2022. VoxCommunis: A corpus for cross-linguistic phonetic analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5286–5294, Marseille, France. European Language Resources Association.

V.S.D.S.Mahesh Akavarapu, Michael Daniel, and Gerhard Jäger. 2026. Hard to be heard: Phoneme-level ASR analysis of phonologically complex, low-resource endangered languages. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 3014–3028, San Diego, California, United States. Association for Computational Linguistics.

Anastasia Alekseeva, Nikita Beklemishev, Michael Daniel, Nina Dobrushina, Konstantin Filatov, Anastasia Ivanova, Timur Maisak, and Ivan Osorgin. 2024. Dictionary of kina rutul.

Mathieu Avanzi, Marie-José Béguelin, Gilles Corminboeuf, Federica Diémoz, and Laure Anne Johnsen. 2024. French (swiss) doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Jocelyn Aznar. 2024. Nisvai doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. In *Interspeech*. ISCA.

Natalia Bogomolova, Dmitry Ganenkov, and Nils Norman Schiborr. 2024. Tabasaran doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Niclas Burenhult. 2024. Jahai doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Eleanor Chodroff. 2018. Corpus phonetics tutorial. *arXiv preprint arXiv:1811.05553*.

Alexander Yao Cobbinah. 2024. Baïnounk gubëeher doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Yishai Cohen and Itshak Lapidot. 2021. Speaker clustering quality estimation with logistic regression. *Computer Speech & Language*, 65:101139.

Uriel Cohen Priva, Emily Strand, Shiying Yang, William Mizgerd, Abigail Creighton, Justin Bai, Rebecca Mathew, Allison Shao, Jordan Schuster, and Daniela Wiepert. 2021. *The Cross-linguistic Phonological Frequencies (XPF) Corpus manual*. Accessible online, https://cohenpr-xpf.github.io/XPF/manual/xpf_manual.pdf.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

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

Andrew Cowell. 2024. Arapaho doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Anton de la Fuente and Dan Jurafsky. 2024. A layer-wise analysis of mandarin and english suprasegmentals in ssl speech models. In *Proc. Interspeech 2024*, pages 1290–1294.

Chris Lasse Däbritz, Nina Kudryakova, Eugénie Stapert, and Alexandre Arkhipov. 2024. Dolgan doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Christian Döhler. 2024. Komnzo doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Diana Forker and Nils Norman Schiborr. 2024. Sanzhi dargwa doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Michael Franjieh. 2024. Fanbyak doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Alexandro Garcia-Laguia. 2024. Northern alta doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Jost Gippert. 2024. Svan doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Simon Gonzalez, Catherine Travis, James Grama, Danielle Barth, and Sunkulp Ananthanarayan. 2018. Recursive forced alignment: A test on a minority language. In *Proceedings of the 17th Australasian international conference on speech Science and technology*, volume 145, page 148.

Richard Griscom. 2024. Asimjeeg datooga doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Valentin Gusev, Tiina Klooster, Beáta Wagner-Nagy, and Alexandre Arkhipov. 2024. Kamas doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Tom Güldemann, Martina Ernszt, Sven Siegmund, and Alena Witzlack-Makarevich. 2024. Nllng doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Geoff Haig, Maria Vollmer, and Hanna Thiele. 2024. Northern kurdish (kurmanji) doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus*

(DoReCo) 2.0. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Andrew Harvey. 2024. Gorwaa doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

John-Paul Hosom. 2009. Speaker-independent phoneme alignment using transition-dependent states. *Speech communication*, 51(4):352–368.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Gerhard Jäger. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change*, 3(2):245–291.

Gerhard Jäger. 2015. Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences*, 112(41):12752–12757.

Olga Kazakevich and Elena Klyachko. 2024. Evenki doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Matthew C Kelley, Scott James Perry, and Benjamin V Tucker. 2024. The mason-alberta phonetic segmenter: a forced alignment system based on deep neural networks and interpolation. *Phonetica*, 81(5):451–508.

Aleksandr E. Kibrik, Sandro V. Kodzasov, Irina P. Olovyannikova, Dzhalil S. Samedov, Michael Daniel, Anna Khoroshkina, and Alexandre Arkhipov. 2007. Archi text corpus (1.0).

Soung-U Kim. 2024. Jejuan doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Manfred Krifka. 2024. Daakie doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. Ctc-segmentation of large corpora for german end-to-end speech recognition. In *International Conference on Speech and Computer*, pages 267–278. Springer.

Ilya Loshchilov and Frank Hutter. 2017. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Amit Meghanani and Thomas Hain. 2024. Improving acoustic word embeddings through correspondence training of self-supervised speech representations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1967, St. Julian's, Malta. Association for Computational Linguistics.

Alexis Michaud. 2024. Yongning na doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ulrike Mosel. 2024. Teop doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Muhammad Firoz Mridha, Abu Quwsar Ohi, Muhammad Mostafa Monowar, Md Abdul Hamid, Md Rashedul Islam, and Yutaka Watanobe. 2021. U-vectors: Generating clusterable speaker embedding from unlabeled data. *Applied Sciences*, 11(21):10079.

Mohamed Kamal Omar, Ken Chen, Mark Hasegawa-Johnson, and Yigal Brandman. 2002. An evaluation of using mutual information for selection of acoustic-features representation of phonemes for speech recognition. In *Interspeech*, pages 2129–2132.

Carmel O'Shannessy. 2024a. Light warlpiri doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Carmel O'Shannessy. 2024b. Warlpiri doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave,

editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Pavel Ozerov. 2024. Anal doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics*, 12:372–391.

Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2657–2666, Marseille, France. European Language Resources Association.

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. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Maïa Ponsonnet. 2024. Dalabon doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, and 1 others. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Juan Diego Quesada, Stavros Skopeteas, Carolina Pasamonik, Carolin Brokmann, and Florian Fischer. 2024. Cabécar doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Elena Rastorgueva, Vitaly Lavrukhin, and Boris Ginsburg. 2023. Nemo forced aligner and its application to word alignment for subtitle generation. In *Interspeech*, pages 5257–5258.

Sabine Reiter. 2024. Cashinahua doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Sonja Riesberg. 2024. Yali (apahapsili) doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Hiram Ring. 2024. Pnar doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Françoise Rose. 2024. Mojeño trinitario doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Rotem Rousso, Eyal Cohen, Joseph Keshet, and Eleanor Chodroff. 2024. Tradition or innovation: A comparison of modern asr methods for forced alignment. In *Proc. Interspeech 2024*, pages 1525–1529.

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

Nils Norman Schiborr. 2024. English (southern england) doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Stefan Schnell. 2024. Vera'a doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Frank Seifart. 2024a. Bora doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Frank Seifart. 2024b. Resígaro doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Frank Seifart, Ludger Paschen, and Matthew Stave. 2024. Language documentation reference corpus (DoReCo) 2.0. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).

Xian Shi, Yanni Chen, Shiliang Zhang, and Zhijie Yan. 2022. Achieving timestamp prediction while recognizing with non-autoregressive end-to-end asr model. In *National Conference on Man-Machine Speech Communication*, pages 89–100. Springer.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, and 1 others. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

Stavros Skopeteas. 2024. Yucatec maya doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Stavros Skopeteas, Violeta Moisidi, Nutsa Tsetereli, Johanna Lorenz, and Stefanie Schröter. 2024. Urum doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Javier Tejedor, Michal Fapšo, Igor Szöke, Jan "Honza" Černockỳ, and František Grézl. 2012. Comparison of methods for language-dependent and language-independent query-by-example spoken term detection. *ACM Transactions on Information Systems (TOIS)*, 30(3):1–34.

Amos Teo. 2024. Sümi doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Nick Thieberger. 2024. Nafsan (south efate) doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Martine Vanhove. 2024. Beja doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Alexandra Vydrina. 2024. Kakabe doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Claudia Wegener. 2024. Savosavo doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Søren Wichmann. 2024. Texistepec popoluca doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Alena Witzlack-Makarevich, Saudah Namyalo, Anatol Kiriggwajjo, and Zarina Molochieva. 2024. Ruuli doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and effective zero-shot cross-lingual phoneme recognition. In *Proc. Interspeech 2022*, pages 2113–2117.

Xianming Xu and Bibo Bai. 2024. Sadu doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Jian Zhu, Changbing Yang, Farhan Samir, and Jahurul Islam. 2024. The taste of IPA: Towards open-vocabulary keyword spotting and forced alignment in any language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 750–772, Mexico City, Mexico. Association for Computational Linguistics.

Appendix

A Implementation details

Alignment evaluations were conducted using a single NVIDIA RTX 2080 GPU (11GB VRAM) on an Intel Xeon Gold 6140 2.30GHz (80GB RAM) machine with batch size 2.

For PCMI evaluation, we sample up to 50 utterances per language for phoneme-cluster estimation and limit evaluation to at most 10 000 acoustic frames. Frame sampling is performed without replacement, where phoneme labels are sampled proportionally to the square root of their frequency in order to reduce dominance from highly frequent labels. Clustering is performed using MiniBatch K-means with batch size 1000 using Scikit-learn (Pedregosa et al., 2011).

For WACS evaluation, we sample up to 200 utterances and up to 200 unique word forms per language, although the effective number is additionally constrained by the requirement that each occurrence span at least four embedding frames. To maximize the sampling space for each word form, forms with a larger number of occurrences are preferentially selected by sorting. Further, only word forms appearing at least three times are considered, ensuring that positive similarity estimates are computed from multiple repeated realizations. To avoid dominance by highly frequent word forms, at most 10 positive pairs (from Σ_f^+) are sampled for each word form f. For every sampled word form f, five randomly sampled negative pairs (from Σ_f^-) are additionally considered.

The subsampling strategies used for both PCMI and WACS substantially improve throughput while maintaining stable aggregate estimates with relatively small variances across evaluations. Runtime statistics, vocabulary sizes, and pair statistics are reported in Appendix Table 15. Evaluating the complete multilingual FLEURS benchmark (85 languages) requires approximately 40 minutes per embedding model, corresponding to roughly 30 seconds per language under the defined sampling configuration. The code for reproducing the experiments is available at `https://github.com/mahesh-ak/MFA`.

B WACS Similarity Computation

For WACS, each word occurrence is represented as a sequence of frame-level speech representations $\mathbf{e}_w = \{\mathbf{emb}_1, \ldots, \mathbf{emb}_{|w|}\}$, where $\mathbf{emb}_i \in \mathbb{R}^d$ denotes the embedding extracted from the selected SSL-speech model layer. Given two word occurrences w_1 and w_2, we compute pairwise frame similarities between frames i and j using cosine similarity:

$$s(i,j) = \frac{\mathbf{emb}_i^\top \mathbf{emb}_j}{||\mathbf{emb}_i|| \cdot ||\mathbf{emb}_j||}$$

Since different realizations of the same word may vary in speaking rate and duration, the resulting embedding sequences generally have different lengths. We therefore apply dynamic time warping (DTW) (Sakoe and Chiba, 1978) to find a monotonic alignment path through the frame-similarity matrix. The DTW score is computed as the average cosine similarity along the optimal path:

$$\mathrm{dtw}(\mathbf{e}_{w_1}, \mathbf{e}_{w_2}) = \frac{1}{|\pi^*|} \sum_{(i,j)\in\pi^*} s(i,j),$$

where π^* denotes the optimal DTW alignment path and $|\pi^*|$ its length. This procedure yields a duration-invariant similarity measure between two word realizations while preserving the acoustic structure encoded by the speech representations.

C PCMI Under Label Collapse

A potential concern is whether degenerate collapse of phoneme labels could artificially inflate PCMI through entropy reduction. We show that this cannot occur. Given:

$$\mathrm{PCMI} = \frac{I(P;C)}{\sqrt{H(P)H(C)}},$$

where P denotes phoneme labels, C denotes representation clusters and $H(\cdot)$ denotes Shannon entropy for phoneme probabilities $\{p_i\}$:

$$H(P) = -\sum_i p_i \log p_i$$

Since entropy is non-negative, the condition $H(P \mid C) \geq 0$ yields:

$$I(P;C) = H(P) - H(P \mid C) \leq H(P),$$

we obtain,

$$\mathrm{PCMI} \leq \frac{H(P)}{\sqrt{H(P)H(C)}} = \sqrt{\frac{H(P)}{H(C)}}.$$

Assuming $H(C) > 0$, which holds empirically since representation clusters remain non-degenerate across experiments, we have

$$\text{PCMI} \to 0 \quad \text{as} \quad H(P) \to 0.$$

Therefore, complete collapse of the phoneme inventory cannot artificially inflate PCMI.

We note that this argument guards against degenerate collapse; small-scale systematic merging of acoustically similar phoneme categories may still occur and can affect PCMI as demonstrated in §5.1.

D Phoneme Models Training

We finetune CTC-based phoneme recognition models MMS-300M-IPA/MMS-300M-DORECO and Wav2Vec2-IPA respectively using MMS-300M (Pratap et al., 2024) and Wav2Vec2-Phoneme (Xu et al., 2022) (both of size $\sim$ 300M parameters) on the multilingual IPA transcriptions derived from FLEURS as described in §6.1. Following Pratap et al. (2024), language-specific adapter modules (Houlsby et al., 2019) are used during finetuning.

Training uses an effective batch size of 16, with per-device batch size 2 on two NVIDIA RTX 2080 GPUs (11GB VRAM each), gradient accumulation steps 4, and fp16 precision. Models are optimized using AdamW with learning rate 10^{-5} and cosine learning-rate scheduling with warmup ratio 0.01 (Loshchilov and Hutter, 2017). We first train the backbone models without adapters for 20 000 steps (approximately three epochs), followed by an additional 500 steps per language with adapters enabled and learning-rate restart. ASR performances of the resulting phoneme recognition models are reported in Appendix Table 8 and 9. The training time per model is about 20hrs. MMS-300M-DORECO is finetuned starting from MMS-300M-IPA for 8000 steps before training the language specific adapters. The models with best performances are available at `https://hf.co/mahesh27/mms-300m-ipa-fleurs` and `https://hf.co/mahesh27/mms-300m-xsampa-doreco`

CTC Forced Aligner Forced alignment is performed using CTC segmentation (Kürzinger et al., 2020) at both phoneme and word levels, followed by additional sanity checks and postprocessing to produce valid TextGrid annotations. Alignment inference is performed on a single GPU.

Silence Correction For the silence-absorbing variant (-SIL), we apply a simple energy-based silence trimming procedure. First, an absolute-amplitude envelope is computed and smoothed using a moving-average window of 400 samples (approximately 25 ms at 16 kHz). A global silence threshold is then defined as 5% of the 95th percentile of the smoothed envelope. Within each aligned word interval, leading and trailing regions whose energy falls below this threshold are removed by locating the first and last frames exceeding the threshold. Intervals shorter than a minimum word duration (40ms) are left unchanged. The resulting word-level boundary adjustments are subsequently propagated to the first and last phonemes overlapping each word while enforcing minimum phoneme durations (20ms). This procedure removes extended silent regions absorbed into neighboring words without otherwise modifying the internal phoneme segmentation.

MFA Training For MFA (McAuliffe et al., 2017), we train multilingual acoustic models using the same IPA-transcribed FLEURS training data on an Intel Core i9-14900K CPU system with 64GB RAM. Training each language required approximately 1.5–2 hours. The training configuration broadly follows Ahn and Chodroff (2022), using MFCC features with 10 ms frame shift and successive stages of monophone training, triphone training, Linear Discriminant Analysis (LDA), and two stages of Speaker Adaptation Training (SAT). The corresponding Gaussian mixture sizes are 1000, 10000, 15000, 15000, and 20000 respectively. The triphone, LDA, and SAT stages use 2000, 2500, 2500, and 3000 phonetic decision-tree leaves respectively (Chodroff, 2018). Due to inconsistent speaker metadata in FLEURS, MFA training is performed in single-speaker mode despite the dataset containing balanced male and female speakers.

Dataset splits For DoReCo, the recordings are segmented into manually annotated utterance-level chunks. Only chunks containing at least two words are retained. For each language, the retained chunks are randomly partitioned into training, development, and test sets using a 0.7/0.1/0.2 split. For FLEURS, the original predefined train/dev/test splits are used without modification. All models are trained on the training splits, while all alignment evaluations are performed exclusively on the corresponding test splits. The statistics of datasets are provided in Tables 6 and 7.

E Supplementary Figures

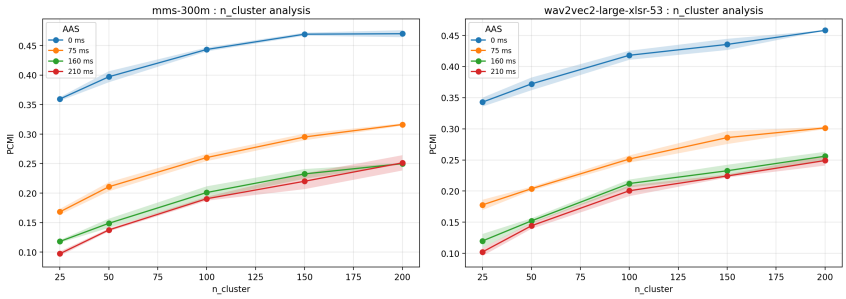

Figure 4: Behavior of PCMI under varying numbers of K-means clusters and increasing perturbation severity using MMS (left) and XLSR (right) representations. The metric resolution, measured as the gap between clean and severely perturbed alignments, remains stable over cluster counts 25–100.

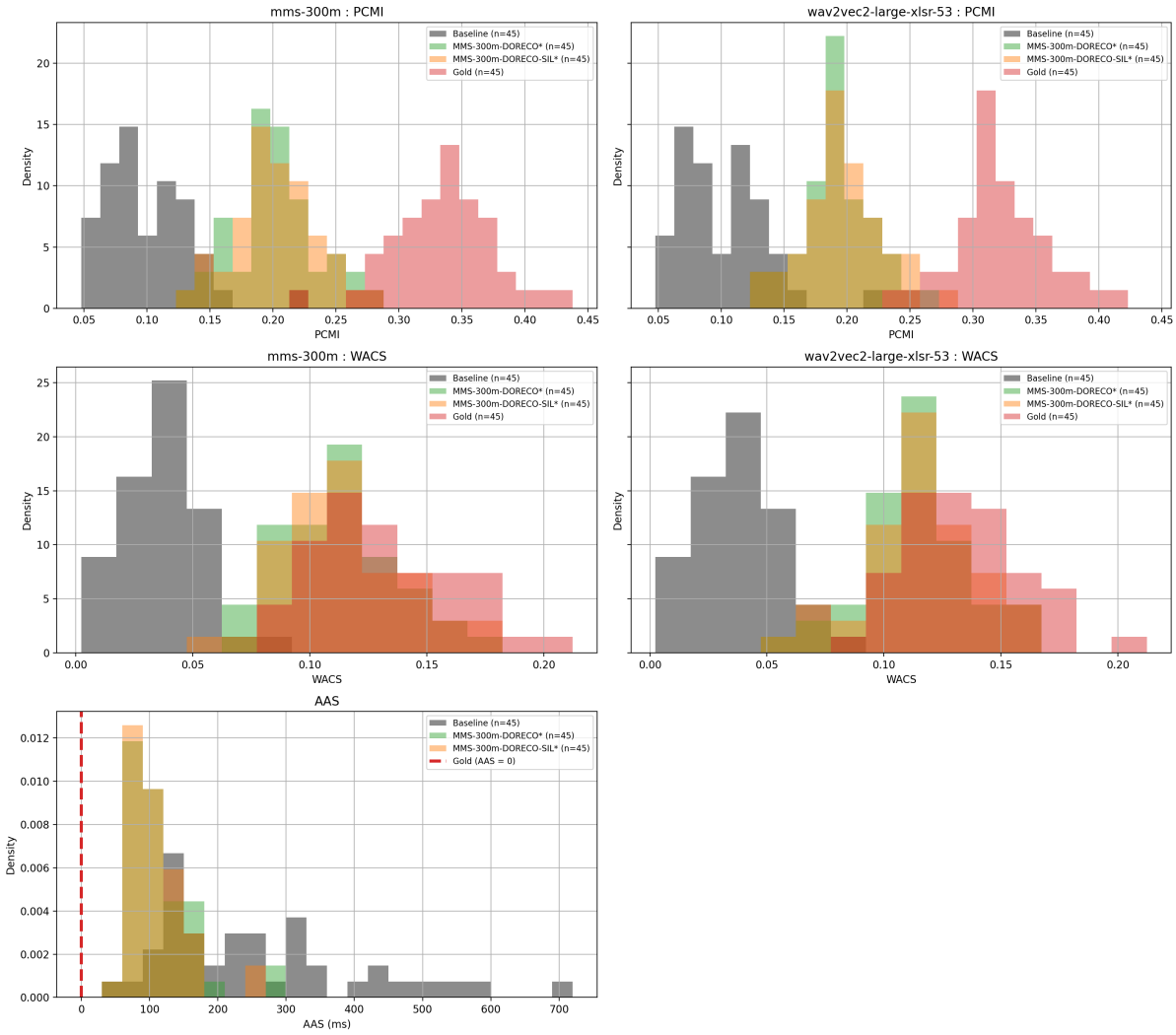

Figure 5: Distribution of PCMI (top row), WACS (middle) and AAS (bottom) across multilingual alignments on DoReCo corpus. The left column uses MMS representations while the right uses XLSR representations.
* Trained in this work.

F Supplementary Tables

Table 6: Languages from FLEURS (Conneau et al., 2023) used in multilingual evaluation.

Language	ISO-1	ISO-3	Family	Subfamily	Division	Train	Dev	Test	G2P
Afrikaans	af	afr	Indo-European	Germanic	West	1032	198	264	epitran
Amharic	am	amh	Afroasiatic	Semitic	South	3163	223	516	epitran
Arabic	ar	ara	Afroasiatic	Semitic	Central	2104	295	428	epitran
Asturian	-	ast	Indo-European	Romance	West	2511	398	946	xpf
Azerbaijani	az	aze	Turkic	Oghuz	-	2665	400	923	epitran
Belarusian	be	bel	Indo-European	Slavic	East	2433	408	967	xpf
Bulgarian	bg	bul	Indo-European	Slavic	South	2973	395	658	xpf
Bengali	bn	ben	Indo-European	Indo-Aryan	East	3006	402	920	epitran
Catalan	ca	cat	Indo-European	Romance	West	2300	404	940	epitran
Cebuano	-	ceb	Austronesian	Philippine	-	3261	225	541	epitran
Central Kurdish	-	ckb	Indo-European	Iranian	West	3040	386	922	epitran
Mandarin Chinese	-	cmn	Sino-Tibetan	Sinitic	-	3246	409	945	epitran
Czech	cs	ces	Indo-European	Slavic	West	2811	305	723	epitran
Welsh	cy	cym	Indo-European	Celtic	-	3427	447	1021	epitran
Danish	da	dan	Indo-European	Germanic	North	2465	395	930	epitran
German	de	deu	Indo-European	Germanic	West	2987	363	862	epitran
Modern Greek (1453-)	el	ell	Indo-European	Hellenic	-	3215	271	650	xpf
English	en	eng	Indo-European	Germanic	West	2602	394	647	epitran
Spanish	es	spa	Indo-European	Romance	West	2796	408	908	epitran
Estonian	et	est	Uralic	Finnic	-	2501	387	893	epitran
Persian	fa	fas	Indo-European	Iranian	West	3101	369	871	epitran
Fulah	ff	ful	Niger-Congo	West-Atlantic	-	3235	273	660	epitran
Finnish	fi	fin	Uralic	Finnic	-	2704	415	918	epitran
French	fr	fra	Indo-European	Romance	West	3193	289	676	epitran
Irish	ga	gle	Indo-European	Celtic	-	2845	369	842	epitran
Galician	gl	glg	Indo-European	Romance	West	2175	395	927	epitran
Hausa	ha	hau	Afroasiatic	Chadic	-	3259	296	621	epitran
Hebrew	he	heb	Afroasiatic	Semitic	Central	3242	328	792	epitran
Hindi	hi	hin	Indo-European	Indo-Aryan	Central	2120	239	418	epitran
Croatian	hr	hrv	Indo-European	Slavic	South	3461	377	914	epitran
Hungarian	hu	hun	Uralic	Ugric	-	3095	407	905	epitran
Armenian	hy	hye	Indo-European	Armenian	-	3053	395	932	xpf
Indonesian	id	ind	Austronesian	Malayic	-	2579	350	687	epitran
Italian	it	ita	Indo-European	Romance	South	3030	391	865	epitran
Japanese	ja	jpn	Japonic	-	-	2292	266	650	epitran
Javanese	jv	jav	Austronesian	Javanesic	-	3051	295	728	epitran
Georgian	ka	kat	Kartvelian	-	-	1491	409	979	epitran
Kazakh	kk	kaz	Turkic	Kipchak	-	3200	369	856	epitran
Khmer	km	khm	Austroasiatic	Khmer	-	1675	326	771	epitran
Kannada	kn	kan	Dravidian	Southern	-	2283	368	838	epitran
Korean	ko	kor	Koreanic	-	-	2307	226	382	epitran
Kyrgyz	ky	kir	Turkic	Kipchak	-	2818	422	977	epitran
Ganda	lg	lug	Niger-Congo	Bantu	Northeast	2478	306	723	epitran
Lao	lo	lao	Kra-Dai	Tai	-	1809	191	405	epitran
Lithuanian	lt	lit	Indo-European	Baltic	-	2937	416	986	epitran
Latvian	lv	lav	Indo-European	Baltic	-	2110	356	851	epitran
Maori	mi	mri	Austronesian	Oceanic	-	3249	429	1008	epitran
Macedonian	mk	mkd	Indo-European	Slavic	South	2337	415	973	xpf
Malayalam	ml	mal	Dravidian	Southern	-	3043	418	958	epitran
Mongolian	mn	mon	Mongolic	-	-	3074	405	949	epitran
Marathi	mr	mar	Indo-European	Indo-Aryan	South	3269	443	1015	epitran
Malay	ms	msa	Austronesian	Malayic	-	2667	324	749	epitran
Maltese	mt	mlt	Afroasiatic	Semitic	Central	2895	404	926	epitran
Burmese	my	mya	Sino-Tibetan	Lolo-Burmese	-	3058	384	880	epitran
Norwegian Bokmål	nb	nob	Indo-European	Germanic	North	3167	163	357	epitran
Nepali	ne	nep	Indo-European	Indo-Aryan	North	3332	305	726	xpf
Dutch	nl	nld	Indo-European	Germanic	West	2918	171	364	epitran
Chichewa	ny	nya	Niger-Congo	Bantu	Nyasa	2694	311	761	epitran
Oromo	om	orm	Afroasiatic	Cushitic	-	1701	19	41	epitran
Odia	or	ori	Indo-European	Indo-Aryan	East	1081	392	883	epitran
Punjabi	pa	pan	Indo-European	Indo-Aryan	Northwest	1923	251	574	epitran
Polish	pl	pol	Indo-European	Slavic	West	2841	338	758	epitran
Pushto	ps	pus	Indo-European	Iranian	East	2513	217	512	epitran
Portuguese	pt	por	Indo-European	Romance	West	2793	386	919	epitran
Romanian	ro	ron	Indo-European	Romance	East	2891	387	883	epitran
Russian	ru	rus	Indo-European	Slavic	East	2562	356	775	epitran
Slovenian	sl	slv	Indo-European	Slavic	South	2512	349	834	epitran
Shona	sn	sna	Niger-Congo	Bantu	Southern	2463	393	925	epitran

Language	ISO-1	ISO-3	Family	Subfamily	Division	Train	Dev	Test	G2P
Somali	so	som	Afroasiatic	Cushitic	-	3149	432	1019	epitran
Swedish	sv	swe	Indo-European	Germanic	North	2385	330	759	epitran
Swahili	sw	swa	Niger-Congo	Bantu	Northeast	3070	211	487	epitran
Tamil	ta	tam	Dravidian	Southern	-	2367	377	591	epitran
Telugu	te	tel	Dravidian	South-Central	-	2302	311	472	xpf
Tajik	tg	tgk	Indo-European	Iranian	West	2298	240	600	epitran
Thai	th	tha	Kra-Dai	Tai	-	2602	439	1021	epitran
Turkish	tr	tur	Turkic	Oghuz	-	2526	338	743	epitran
Ukrainian	uk	ukr	Indo-European	Slavic	East	2810	325	750	epitran
Urdu	ur	urd	Indo-European	Indo-Aryan	Central	2109	267	299	epitran
Uzbek	uz	uzb	Turkic	Karluk	-	2943	363	862	epitran
Vietnamese	vi	vie	Austroasiatic	Viet	-	2994	361	857	epitran
Wolof	wo	wol	Niger-Congo	West-Atlantic	-	2279	169	371	xpf
Xhosa	xh	xho	Niger-Congo	Bantu	Southern	3466	446	1041	epitran
Yoruba	yo	yor	Niger-Congo	Volta-Niger	-	2339	378	831	epitran
Yue Chinese	-	yue	Sino-Tibetan	Sinitic	-	1939	362	819	epitran
Zulu	zu	zul	Niger-Congo	Bantu	Southern	2858	354	854	epitran

Table 7: DoReCo (Seifart et al., 2024) languages used for multilingual validation.

Language	ISO-3	Family	Subfamily	Division	Train	Dev	Test	Source
Anal	anm	Sino-Tibetan	Tibeto-Burman	Kuki-Chin	1588	227	453	Ozerov (2024)
Yali (Apahapsili)	na	Trans-New Guinea	Dani	-	1057	151	302	Riesberg (2024)
Arapaho	arp	Algic	Algonquian	-	782	112	223	Cowell (2024)
Baïnounk Gubëeher	bab	Niger-Congo	West-Atlantic	-	1315	188	376	Cobbinah (2024)
Beja	bej	Afroasiatic	Cushitic	North	3172	453	906	Vanhove (2024)
Bora	boa	Boran	-	-	733	105	209	Seifart (2024a)
Cabécar	cjp	Chibchan	Isthmic	-	801	114	229	Quesada et al. (2024)
Cashinahua	cbs	Pano-Tacanan	Panoan	-	1185	169	339	Reiter (2024)
Dolgan	dlg	Turkic	Siberian	-	797	114	228	Däbritz et al. (2024)
Evenki	evn	Tungusic	Ewenic	-	1310	187	374	Kazakevich and Klyachko (2024)
Gorwaa	gow	Afroasiatic	Cushitic	South	996	142	285	Harvey (2024)
Jahai	jhi	Austroasiatic	Aslian	-	836	119	239	Burenhult (2024)
Jejuan	jje	Koreanic	-	-	307	44	88	Kim (2024)
Kakabe	kke	Niger-Congo	Mande	-	668	95	191	Vydrina (2024)
Kamas	xas	Uralic	Samoyedic	-	1270	182	363	Gusev et al. (2024)
Komnzo	tci	Yam	-	-	1526	218	436	Döhler (2024)
Light Warlpiri	na	Mixed Language	-	-	694	99	198	O'Shannessy (2024a)
Dalabon	ngk	Gunwinyguan	-	-	291	42	83	Ponsonnet (2024)
Nisvai	none	Austronesian	Oceanic	Southern	1185	169	339	Aznar (2024)
Nllng	ngh	Tuu	-	-	1091	156	311	Güldemann et al. (2024)
Northern Kurdish (Kurmanji)	kmr	Indo-European	Iranian	Western	584	83	167	Haig et al. (2024)
Northern Alta	aqn	Austronesian	Phillippine	-	1323	189	378	Garcia-Laguia (2024)
Fanbyak	fnb	Austronesian	Oceanic	Southern	1060	151	303	Franjieh (2024)
Pnar	pbv	Austroasiatic	Khasic	-	219	31	63	Ring (2024)
Daakie	ptv	Austronesian	Oceanic	Southern	921	132	263	Krifka (2024)
Resígaro	rgr	Arawakan	-	-	944	135	269	Seifart (2024b)
Ruuli	ruc	Niger-Congo	Bantu	Northeast	818	117	233	Witzlack-Makarevich et al. (2024)
Sadu	na	Sino-Tibetan	Tibeto-Burman	Lolo-Burmese	1035	148	296	Xu and Bai (2024)
Sanzhi Dargwa	na	Nakh-Daghestanian	Dargin	-	314	45	90	Forker and Schiborr (2024)
Savosavo	svs	Isolate	-	-	625	89	179	Wegener (2024)
Nafsan (South Efate)	erk	Austronesian	Oceanic	Southern	543	78	155	Thieberger (2024)
English (Southern England)	na	Indo-European	Germanic	West	431	62	123	Schiborr (2024)
French (Swiss)	fra	Indo-European	Romance	West	1222	174	349	Avanzi et al. (2024)
Sümi	nsm	Sino-Tibetan	Angami-Pochuri	-	809	116	231	Teo (2024)
Svan	sva	Kartvelian	-	-	498	71	142	Gippert (2024)
Tabasaran	tab	Nakh-Daghestanian	Lezgic	-	440	63	125	Bogomolova et al. (2024)
Teop	tio	Austronesian	Oceanic	Western	1325	189	379	Mosel (2024)
Texistepec Popoluca	poq	Mixe-Zoque	Zoquean	-	1648	236	471	Wichmann (2024)
Mojeño Trinitario	trn	Arawakan	-	-	626	90	179	Rose (2024)
Asimjeeg Datooga	na	Nilo-Saharan	Nilotic	-	1138	163	325	Griscom (2024)
Urum	uum	Turkic	Kipchak	-	889	127	254	Skopeteas et al. (2024)
Vera'a	vra	Austronesian	Oceanic	Southern	943	135	269	Schnell (2024)
Warlpiri	wbp	Pama-Nyungan	Ngumpic-Yapa	-	1173	168	335	O'Shannessy (2024b)
Yongning Na	nru	Sino-Tibetan	Tibeto-Burman	Lolo-Burmese	465	66	133	Michaud (2024)
Yucatec Maya	yua	Mayan	-	-	788	112	225	Skopeteas (2024)

Table 8: Language-wise multilingual ASR performance on FLEURS. N denotes sizes of test splits. MMS-IPA* and W2V2-IPA* are same as MMS-300M-IPA* and Wav2Vec2-IPA* respectively. * Trained in this work.

Language	N	WER ↓		CER ↓	
		MMS-IPA*	W2V2-IPA*	MMS-IPA*	W2V2-IPA*
Afrikaans	264	**0.576**	0.852	**0.190**	0.323
Amharic	516	**0.638**	0.866	**0.124**	0.208
Arabic	428	**0.526**	0.927	**0.141**	0.308

Language	N	WER ↓		CER ↓	
		MMS-IPA*	W2V2-IPA*	MMS-IPA*	W2V2-IPA*
Asturian	946	**0.370**	0.694	**0.086**	0.168
Azerbaijani	923	**0.509**	0.839	**0.113**	0.223
Belarusian	967	**0.378**	0.805	**0.079**	0.176
Bulgarian	658	**0.330**	0.794	**0.077**	0.207
Bengali	920	**0.649**	0.849	**0.149**	0.248
Catalan	940	**0.340**	0.711	**0.085**	0.199
Cebuano	541	**0.261**	0.530	**0.099**	0.161
Central Kurdish	922	**0.607**	0.842	**0.138**	0.214
Mandarin Chinese	945	**0.765**	0.915	**0.148**	0.232
Czech	723	**0.369**	0.867	**0.086**	0.245
Welsh	1021	**0.631**	0.751	**0.216**	0.264
Danish	930	**0.630**	0.893	**0.209**	0.353
German	862	**0.427**	0.780	**0.104**	0.228
Modern Greek (1453-)	650	**0.269**	0.851	**0.063**	0.205
English	647	0.293	**0.235**	0.086	**0.070**
Spanish	908	**0.152**	0.586	**0.040**	0.135
Estonian	893	**0.358**	0.766	**0.063**	0.145
Persian	871	**0.362**	0.640	**0.091**	0.171
Fulah	660	**0.587**	0.752	**0.186**	0.222
Finnish	918	**0.322**	0.723	**0.057**	0.139
French	676	**0.559**	0.794	**0.156**	0.290
Irish	842	**0.800**	0.874	**0.350**	0.392
Galician	927	**0.277**	0.689	**0.066**	0.167
Hausa	621	**0.493**	0.753	**0.159**	0.233
Hebrew	792	**0.721**	0.910	**0.217**	0.299
Hindi	418	**0.491**	0.746	**0.134**	0.225
Croatian	914	**0.302**	0.721	**0.066**	0.164
Hungarian	905	**0.429**	0.841	**0.096**	0.247
Armenian	932	**0.413**	0.773	**0.077**	0.160
Indonesian	687	**0.325**	0.654	**0.068**	0.151
Italian	865	**0.208**	0.677	**0.049**	0.147
Japanese	650	**0.412**	0.745	**0.126**	0.234
Javanese	728	**0.398**	0.714	**0.098**	0.186
Georgian	979	**0.632**	0.905	**0.110**	0.203
Kazakh	856	**0.337**	0.685	**0.069**	0.148
Khmer	771	**0.606**	0.903	**0.219**	0.342
Kannada	838	**0.517**	0.796	**0.090**	0.159
Korean	382	**0.702**	0.838	**0.130**	0.188
Kyrgyz	977	**0.538**	0.773	**0.092**	0.160
Ganda	723	**0.516**	0.872	**0.128**	0.227
Lao	405	**0.676**	0.834	**0.199**	0.272
Lithuanian	986	**0.472**	0.904	**0.108**	0.259
Latvian	851	**0.395**	0.810	**0.073**	0.187
Maori	1008	**0.399**	0.585	**0.173**	0.222
Macedonian	973	**0.243**	0.629	**0.047**	0.111
Malayalam	958	**0.644**	0.827	**0.116**	0.176
Mongolian	949	**0.703**	0.915	**0.176**	0.292
Marathi	1015	**0.663**	0.839	**0.163**	0.236
Malay	749	**0.354**	0.668	**0.081**	0.161
Maltese	926	**0.377**	0.878	**0.084**	0.233
Burmese	880	**0.685**	0.817	**0.231**	0.314
Norwegian Bokmål	357	**0.478**	0.784	**0.130**	0.234
Nepali	726	**0.500**	0.805	**0.114**	0.204
Dutch	364	**0.439**	0.803	**0.129**	0.282
Chichewa	761	**0.608**	0.887	**0.168**	0.256
Oromo	41	**0.790**	0.898	**0.213**	0.277
Odia	883	**0.687**	0.859	**0.165**	0.257
Punjabi	574	**0.699**	0.834	**0.204**	0.290
Polish	758	**0.460**	0.873	**0.084**	0.226
Pushto	512	**0.615**	0.816	**0.239**	0.332
Portuguese	919	**0.275**	0.787	**0.076**	0.245
Romanian	883	**0.323**	0.783	**0.077**	0.212
Russian	775	**0.533**	0.896	**0.102**	0.239
Slovenian	834	**0.337**	0.799	**0.083**	0.198
Shona	925	**0.537**	0.814	**0.124**	0.206
Somali	1019	**0.589**	0.851	**0.184**	0.281
Swedish	759	**0.476**	0.888	**0.134**	0.294
Swahili	487	**0.289**	0.601	**0.067**	0.135
Tamil	591	**0.713**	0.877	**0.170**	0.249
Telugu	472	**0.571**	0.797	**0.113**	0.177
Tajik	600	**0.361**	0.645	**0.085**	0.143
Thai	1021	**0.782**	0.937	**0.230**	0.338

Language	N	WER ↓		CER ↓	
		MMS-IPA*	W2V2-IPA*	MMS-IPA*	W2V2-IPA*
Turkish	743	**0.439**	0.726	**0.085**	0.160
Ukrainian	750	**0.448**	0.834	**0.084**	0.198
Urdu	299	**0.594**	0.811	**0.225**	0.315
Uzbek	862	**0.540**	0.820	**0.123**	0.231
Vietnamese	857	**0.692**	0.962	**0.193**	0.302
Wolof	371	**0.602**	0.771	**0.213**	0.281
Xhosa	1041	**0.638**	0.840	**0.165**	0.244
Yoruba	831	**0.868**	0.917	**0.348**	0.400
Yue Chinese	819	**0.906**	0.985	**0.257**	0.351
Zulu	854	**0.535**	0.733	**0.156**	0.218

Table 9: Language-wise ASR performance on DoReCo. Scores are reported for MMS-300M-DORECO*. N denotes sizes of test splits

Language	N	WER ↓	CER ↓
Anal	453	0.979	0.454
Arapaho	223	1.007	0.630
Asimjeeg Datooga	325	0.989	0.356
Baïnounk Gubëeher	376	0.957	0.366
Beja	906	0.975	0.322
Bora	209	0.998	0.399
Cabécar	229	0.991	0.460
Cashinahua	339	0.991	0.342
Daakie	263	0.894	0.375
Dalabon	83	0.968	0.313
Dolgan	228	0.909	0.357
English (Southern England)	123	0.927	0.493
Evenki	374	0.977	0.405
Fanbyak	303	0.830	0.291
French (Swiss)	349	0.743	0.308
Gorwaa	285	0.930	0.386
Jahai	239	0.856	0.322
Jejuan	88	0.974	0.443
Kakabe	191	0.805	0.260
Kamas	363	0.897	0.323
Komnzo	436	0.969	0.344
Light Warlpiri	198	0.901	0.405
Mojeño Trinitario	179	0.744	0.208
Nafsan (South Efate)	155	0.864	0.316
Nisvai	339	0.928	0.309
Northern Alta	378	0.815	0.312
Northern Kurdish (Kurmanji)	167	0.991	0.328
Nllng	311	0.999	0.645
Pnar	63	0.810	0.368
Resígaro	269	0.954	0.273
Ruuli	233	0.979	0.321
Sadu	296	0.956	0.428
Sanzhi Dargwa	90	0.949	0.394
Savosavo	179	0.761	0.210
Svan	142	0.968	0.334
Sümi	231	0.977	0.470
Tabasaran	125	0.883	0.331
Teop	379	0.907	0.295
Texistepec Popoluca	471	1.002	0.624
Urum	254	0.919	0.391
Vera'a	269	0.855	0.352
Warlpiri	335	0.794	0.198
Yali (Apahapsili)	302	0.936	0.415
Yongning Na	133	0.990	0.406
Yucatec Maya	225	0.934	0.371

Table 10: Language-wise multilingual PCMI scores on FLEURS. Higher values are better. MMS-IPA* and W2V2-IPA* are same as MMS-300M-IPA* and Wav2Vec2-IPA* respectively. * Trained in this work.

Language	MMS				XLSR			
	Base	MFA	MMS-IPA*	W2V2-IPA*	Base	MFA	MMS-IPA*	W2V2-IPA*
Afrikaans	0.093	0.112	0.221	**0.226**	0.101	0.105	**0.218**	0.215
Amharic	0.103	0.122	**0.258**	0.251	0.108	0.120	**0.249**	0.234
Arabic	0.114	**0.316**	0.254	0.254	0.107	**0.281**	0.244	0.232
Asturian	0.091	0.117	**0.223**	0.219	0.096	0.106	0.199	**0.203**
Azerbaijani	0.071	0.124	0.228	**0.243**	0.076	0.126	**0.227**	0.220
Belarusian	0.075	**0.352**	0.247	0.231	0.078	**0.335**	0.237	0.233
Bulgarian	0.099	0.082	0.212	**0.213**	0.097	0.080	0.184	**0.193**
Bengali	0.081	**0.273**	0.241	0.224	0.085	**0.250**	0.224	0.225
Catalan	0.100	**0.314**	0.229	0.229	0.103	**0.289**	0.210	0.219
Cebuano	0.066	0.099	0.219	**0.235**	0.068	0.107	0.210	**0.216**
Central Kurdish	0.071	0.221	**0.261**	0.251	0.077	0.211	**0.242**	0.236
Mandarin Chinese	0.109	**0.378**	0.236	0.236	0.111	**0.346**	0.224	0.224
Czech	0.092	**0.292**	0.270	0.273	0.095	**0.270**	0.252	0.241
Welsh	0.086	0.143	0.232	**0.241**	0.089	0.145	0.227	**0.230**
Danish	0.097	0.120	0.190	**0.197**	0.108	0.119	**0.191**	0.187
German	0.081	**0.295**	0.215	0.214	0.084	**0.291**	0.211	0.208
Modern Greek (1453-)	0.089	0.097	**0.252**	0.245	0.093	0.093	**0.218**	0.214
English	0.075	**0.342**	0.208	0.212	0.076	**0.317**	0.201	0.204
Spanish	0.081	0.118	0.243	**0.253**	0.083	0.115	0.218	**0.232**
Estonian	0.092	**0.274**	0.266	0.263	0.091	0.258	**0.262**	0.245

Language	MMS				XLSR			
	Base	MFA	MMS-IPA*	W2V2-IPA*	Base	MFA	MMS-IPA*	W2V2-IPA*
Persian	0.072	**0.284**	0.241	0.232	0.077	**0.271**	0.230	0.203
Fulah	0.093	**0.371**	0.221	0.213	0.096	**0.352**	0.204	0.209
Finnish	0.113	0.291	**0.309**	0.306	0.113	0.274	**0.286**	0.278
French	0.115	0.109	0.222	**0.230**	0.117	0.109	0.204	**0.216**
Irish	0.089	0.174	**0.202**	0.198	0.096	0.170	**0.214**	0.200
Galician	0.109	0.089	0.233	**0.233**	0.109	0.087	0.222	**0.228**
Hausa	0.086	**0.432**	0.245	0.238	0.091	**0.397**	0.234	0.228
Hebrew	0.074	0.180	**0.229**	0.218	0.074	0.164	**0.211**	0.206
Hindi	0.095	**0.348**	0.244	0.241	0.100	**0.332**	0.227	0.218
Croatian	0.076	**0.313**	0.226	0.225	0.066	**0.298**	0.210	0.213
Hungarian	0.081	-	0.259	**0.260**	0.088	-	**0.237**	0.236
Armenian	0.075	**0.327**	0.249	0.252	0.079	**0.286**	0.239	0.239
Indonesian	0.069	**0.292**	0.243	0.236	0.073	**0.249**	0.209	0.218
Italian	0.097	0.160	**0.289**	0.280	0.102	0.145	0.264	**0.268**
Japanese	0.089	0.136	**0.273**	0.263	0.098	0.139	**0.243**	0.238
Javanese	0.081	**0.255**	0.226	0.234	0.085	**0.237**	0.207	0.228
Georgian	0.096	0.070	**0.260**	0.246	0.092	0.067	**0.228**	0.223
Kazakh	0.098	**0.392**	0.237	0.239	0.100	**0.363**	0.229	0.240
Khmer	0.103	0.156	0.202	**0.208**	0.107	0.164	**0.205**	0.202
Kannada	0.099	0.154	**0.257**	0.244	0.097	0.147	**0.249**	0.246
Korean	0.105	0.167	0.263	**0.271**	0.107	0.161	**0.272**	0.262
Kyrgyz	0.103	**0.348**	0.252	0.261	0.099	**0.335**	0.238	0.250
Ganda	0.141	0.164	**0.329**	0.322	0.140	0.155	0.305	**0.309**
Lao	0.079	**0.392**	0.226	0.224	0.082	**0.342**	0.218	0.207
Lithuanian	0.096	**0.332**	0.233	0.240	0.099	**0.313**	0.222	0.229
Latvian	0.084	0.277	**0.278**	0.260	0.090	0.254	**0.262**	0.248
Maori	0.077	**0.377**	0.189	0.189	0.080	**0.355**	0.180	0.181
Macedonian	0.085	**0.343**	0.249	0.258	0.088	**0.342**	0.241	0.242
Malayalam	0.093	0.150	**0.273**	0.260	0.094	0.137	**0.265**	0.255
Mongolian	0.099	0.103	0.220	**0.222**	0.105	0.107	0.202	**0.216**
Marathi	0.092	0.101	**0.259**	0.253	0.095	0.103	0.251	**0.255**
Malay	0.048	0.081	0.214	**0.224**	0.053	0.079	0.207	**0.210**
Maltese	0.105	-	**0.301**	0.293	0.109	-	**0.291**	0.290
Burmese	0.106	0.150	**0.205**	0.205	0.109	0.133	**0.210**	0.202
Norwegian Bokmål	0.148	0.229	0.244	**0.245**	0.148	0.222	**0.250**	0.249
Nepali	0.112	**0.352**	0.214	0.205	0.113	**0.322**	0.202	0.198
Dutch	0.099	0.195	0.204	**0.207**	0.101	0.180	**0.200**	0.196
Chichewa	0.081	**0.343**	0.239	0.239	0.089	**0.305**	0.228	0.221
Oromo	0.095	**0.346**	0.277	0.256	0.098	**0.310**	0.252	0.235
Odia	0.092	**0.400**	0.213	0.217	0.097	**0.389**	0.193	0.205
Punjabi	0.110	0.172	**0.228**	0.219	0.104	0.186	**0.219**	0.197
Polish	0.100	**0.301**	0.204	0.223	0.097	**0.264**	0.201	0.216
Pushto	0.079	0.076	**0.191**	0.186	0.082	0.077	**0.190**	0.180
Portuguese	0.089	**0.273**	0.237	0.240	0.096	**0.250**	0.218	0.215
Romanian	0.054	0.124	**0.222**	0.220	0.060	0.125	0.203	**0.217**
Russian	0.094	0.171	**0.258**	0.256	0.097	0.168	0.228	**0.239**
Slovenian	0.061	**0.315**	0.194	0.202	0.060	**0.287**	0.191	0.195
Shona	0.098	**0.336**	0.265	0.260	0.096	**0.294**	0.255	0.257
Somali	0.072	**0.332**	0.183	0.195	0.077	**0.311**	0.187	0.199
Swedish	0.057	**0.339**	0.243	0.231	0.065	**0.312**	0.230	0.218
Swahili	0.084	**0.375**	0.262	0.252	0.085	**0.357**	0.254	0.236
Tamil	0.086	0.217	0.229	**0.230**	0.089	0.198	**0.213**	0.209
Telugu	0.106	**0.324**	0.248	0.265	0.106	**0.313**	0.236	0.234
Tajik	0.109	**0.401**	0.292	0.263	0.109	**0.371**	0.275	0.265
Thai	0.081	**0.276**	0.201	0.185	0.084	**0.248**	0.180	0.180
Turkish	0.083	0.125	**0.235**	0.232	0.084	0.119	**0.230**	0.225
Ukrainian	0.097	**0.338**	0.234	0.239	0.094	**0.326**	0.217	0.232
Urdu	0.083	**0.340**	0.240	0.225	0.086	**0.303**	0.211	0.203
Uzbek	0.084	**0.287**	0.239	0.251	0.086	**0.271**	0.215	0.228
Vietnamese	0.063	**0.382**	0.205	0.198	0.066	**0.358**	0.191	0.202
Wolof	0.111	0.126	**0.251**	0.244	0.115	0.121	**0.242**	0.236
Xhosa	0.095	**0.329**	0.275	0.267	0.100	**0.295**	0.256	0.241
Yoruba	0.105	**0.394**	0.232	0.231	0.113	**0.353**	0.224	0.226
Yue Chinese	0.095	**0.338**	0.235	0.244	0.096	**0.313**	0.223	0.227
Zulu	0.096	0.201	**0.255**	0.238	0.099	0.194	**0.252**	0.234

Table 11: Language-wise multilingual WACS scores on FLEURS. Higher values are better. MMS-IPA* and W2V2-IPA* are same as MMS-300M-IPA* and Wav2Vec2-IPA* respectively. * Trained in this work.

Language	MMS					XLSR				
	Base	MFA	MMS-IPA*	W2V2-IPA*	Qwen3-FA	Base	MFA	MMS-IPA*	W2V2-IPA*	Qwen3-FA
Afrikaans	0.008	0.041	**0.181**	0.173	-	0.008	0.025	**0.161**	0.159	-
Amharic	0.056	0.044	0.219	**0.225**	-	0.053	0.044	**0.198**	0.198	-
Arabic	0.041	**0.208**	0.207	0.206	-	0.034	**0.185**	0.177	0.178	-
Asturian	0.032	0.052	**0.186**	0.182	-	0.023	0.053	0.172	**0.172**	-
Azerbaijani	0.026	0.031	0.220	**0.220**	-	0.032	0.033	**0.195**	0.192	-
Belarusian	0.005	**0.231**	0.195	0.190	-	0.004	**0.200**	0.168	0.164	-
Bulgarian	0.039	0.024	**0.209**	0.203	-	0.031	0.027	**0.176**	0.169	-
Bengali	0.026	0.192	**0.204**	0.196	-	0.024	0.164	**0.174**	0.170	-
Catalan	0.009	**0.199**	0.198	0.198	-	0.010	0.182	**0.182**	0.178	-
Cebuano	0.028	0.054	0.188	**0.190**	-	0.022	0.051	0.167	**0.168**	-
Central Kurdish	0.007	-	**0.181**	0.180	-	0.019	-	**0.164**	0.161	-
Mandarin Chinese	0.022	**0.231**	0.213	0.206	0.194	0.024	**0.191**	0.175	0.173	0.161
Czech	0.018	0.185	0.206	**0.210**	-	0.015	0.159	0.175	**0.176**	-
Welsh	0.011	0.032	0.170	**0.171**	-	0.011	0.044	**0.160**	0.159	-
Danish	0.012	0.032	**0.178**	0.167	-	0.009	0.042	**0.159**	0.152	-
German	0.017	0.184	0.184	0.182	**0.187**	0.016	0.183	0.181	0.172	**0.192**
Modern Greek (1453-)	0.019	0.022	**0.196**	0.194	-	0.015	0.024	0.163	**0.164**	-
English	0.015	0.135	0.172	**0.174**	0.172	0.013	0.137	0.165	**0.170**	0.169
Spanish	0.013	0.108	**0.217**	0.210	0.210	0.015	0.101	0.200	**0.202**	0.196
Estonian	0.051	0.233	0.252	**0.252**	-	0.044	0.209	**0.216**	0.216	-
Persian	0.014	0.204	**0.218**	0.217	-	0.007	0.172	**0.189**	0.186	-
Fulah	0.004	**0.170**	0.150	0.148	-	0.005	**0.154**	0.138	0.133	-
Finnish	0.046	0.177	0.232	**0.238**	-	0.044	0.143	**0.204**	0.202	-
French	0.024	0.029	0.192	0.194	**0.205**	0.021	0.032	0.168	0.172	**0.178**
Irish	0.005	0.069	**0.147**	0.137	-	-0.002	0.070	**0.139**	0.135	-
Galician	0.023	0.029	**0.191**	0.190	-	0.020	0.038	0.173	**0.175**	-
Hausa	0.009	**0.238**	0.171	0.170	-	0.005	**0.210**	0.147	0.149	-
Hebrew	0.029	0.131	0.153	**0.157**	-	0.026	0.116	**0.143**	0.139	-
Hindi	0.017	**0.192**	0.190	0.186	-	0.016	**0.166**	0.158	0.159	-
Croatian	0.028	**0.219**	0.209	0.206	-	0.028	**0.188**	0.175	0.175	-
Hungarian	0.014	-	**0.218**	0.209	-	0.014	-	**0.187**	0.180	-
Armenian	0.025	**0.236**	0.224	0.224	-	0.025	**0.210**	0.199	0.198	-
Indonesian	0.021	0.176	**0.183**	0.182	-	0.024	0.139	**0.145**	0.145	-
Italian	0.024	0.086	0.214	**0.215**	0.204	0.012	0.084	**0.195**	0.194	0.185
Japanese	0.013	0.049	0.202	**0.205**	0.197	0.011	0.048	0.170	0.171	**0.173**
Javanese	0.020	0.163	**0.185**	0.185	-	0.014	0.138	**0.158**	0.157	-
Georgian	0.036	0.013	0.212	**0.214**	-	0.033	-0.015	**0.182**	0.180	-
Kazakh	0.028	**0.233**	0.204	0.205	-	0.022	**0.205**	0.181	0.182	-
Khmer	-0.011	-0.001	0.153	**0.153**	-	-0.012	-0.017	**0.129**	0.126	-
Kannada	0.023	0.091	0.195	**0.198**	-	0.027	0.085	**0.178**	0.173	-
Korean	0.051	0.128	0.243	0.238	**0.244**	0.045	0.119	**0.210**	0.204	0.207
Kyrgyz	0.034	**0.238**	0.209	0.212	-	0.034	**0.215**	0.187	0.197	-
Ganda	0.052	0.152	**0.236**	0.233	-	0.047	0.144	0.207	**0.215**	-
Lao	0.001	**0.189**	0.159	0.153	-	0.007	**0.159**	0.127	0.120	-
Lithuanian	0.040	**0.223**	0.196	0.200	-	0.044	**0.188**	0.170	0.166	-
Latvian	0.041	0.212	**0.243**	0.234	-	0.030	0.176	**0.210**	0.199	-
Maori	-0.001	**0.166**	0.128	0.124	-	-0.004	**0.151**	0.122	0.114	-
Macedonian	0.030	**0.225**	0.223	0.224	-	0.028	**0.202**	0.199	0.196	-
Malayalam	0.035	0.070	**0.221**	0.220	-	0.032	0.068	**0.191**	0.191	-
Mongolian	0.021	0.029	**0.207**	0.204	-	0.023	0.030	**0.172**	0.171	-
Marathi	0.022	0.101	**0.200**	0.198	-	0.021	0.100	**0.178**	0.178	-
Malay	0.007	0.009	**0.172**	0.171	-	-0.000	0.013	**0.148**	0.145	-
Maltese	0.023	-	0.195	**0.200**	-	0.014	-	0.174	**0.174**	-
Burmese	0.014	0.069	0.179	**0.179**	-	0.012	0.060	0.157	**0.158**	-
Norwegian Bokmål	0.058	0.174	0.239	**0.240**	-	0.056	0.157	0.212	**0.214**	-
Nepali	0.033	**0.220**	0.188	0.195	-	0.028	**0.187**	0.169	0.166	-
Dutch	0.034	0.154	0.199	**0.199**	-	0.031	0.145	**0.186**	0.185	-
Chichewa	0.006	**0.151**	0.146	0.149	-	0.010	**0.134**	0.123	0.124	-
Oromo	0.029	**0.197**	0.165	0.161	-	0.013	**0.184**	0.166	0.156	-
Odia	0.019	**0.205**	0.173	0.169	-	0.025	**0.220**	0.137	0.139	-
Punjabi	0.011	0.085	**0.183**	0.169	-	0.000	0.088	**0.143**	0.133	-
Polish	0.051	0.191	**0.208**	0.206	-	0.047	0.156	0.176	**0.176**	-
Pushto	0.014	0.060	**0.175**	0.174	-	0.013	0.054	**0.152**	0.148	-
Portuguese	0.019	0.189	0.211	0.204	**0.231**	0.013	0.167	0.185	0.180	**0.203**
Romanian	0.008	0.021	**0.190**	0.185	-	0.009	0.041	**0.162**	0.161	-
Russian	0.028	0.119	**0.206**	0.203	0.206	0.024	0.105	**0.175**	0.167	0.173
Slovenian	0.026	**0.205**	0.200	0.196	-	0.025	0.180	**0.180**	0.179	-
Shona	0.011	0.166	**0.182**	0.177	-	0.003	0.135	0.147	**0.152**	-
Somali	0.013	**0.195**	0.174	0.169	-	0.012	**0.161**	0.142	0.140	-
Swedish	0.011	0.179	0.183	**0.183**	-	0.012	0.167	**0.170**	0.164	-
Swahili	0.008	**0.207**	0.194	0.194	-	0.005	**0.177**	0.163	0.162	-
Tamil	0.020	0.142	**0.172**	0.169	-	0.018	0.123	**0.145**	0.144	-
Telugu	0.029	**0.174**	0.174	0.172	-	0.030	**0.159**	0.157	0.151	-
Tajik	0.033	**0.258**	0.226	0.231	-	0.025	**0.227**	0.196	0.201	-
Thai	0.002	**0.214**	0.198	0.190	-	0.002	**0.172**	0.161	0.150	-
Turkish	0.036	0.021	**0.247**	0.245	-	0.032	0.034	**0.219**	0.217	-
Ukrainian	0.029	**0.218**	0.210	0.212	-	0.024	**0.185**	0.179	0.183	-
Urdu	0.022	**0.235**	0.214	0.210	-	0.027	**0.201**	0.183	0.183	-
Uzbek	0.035	**0.197**	0.194	0.193	-	0.032	**0.167**	0.164	0.161	-
Vietnamese	0.003	**0.223**	0.178	0.170	-	0.010	**0.175**	0.139	0.132	-
Wolof	0.008	0.019	**0.144**	0.139	-	0.012	0.011	0.135	**0.135**	-
Xhosa	0.044	**0.227**	0.205	0.199	-	0.038	**0.187**	0.175	0.169	-
Yoruba	0.006	**0.197**	0.177	0.170	-	0.009	**0.175**	0.155	0.153	-
Yue Chinese	0.019	**0.246**	0.217	0.218	-	0.021	**0.199**	0.178	0.177	-
Zulu	0.019	0.116	**0.169**	0.166	-	0.018	0.106	0.145	**0.148**	-

Table 12: Language-wise multilingual PCMI scores on DoReCo. Higher values are better. * Trained in this work.

Language	MMS				XLSR			
	Base	MMS-DORECO*	MMS-DORECO*-SIL	Gold	Base	MMS-DORECO*	MMS-DORECO*-SIL	Gold
Anal	0.133	0.209	0.206	**0.330**	0.122	0.197	0.195	**0.316**
Yali (Apahapsili)	0.127	0.196	0.193	**0.263**	0.125	0.184	0.184	**0.248**
Arapaho	0.075	0.192	0.210	**0.287**	0.072	0.193	0.204	**0.292**
Baïnounk Gubëeher	0.097	0.185	0.194	**0.301**	0.098	0.182	0.193	**0.308**
Beja	0.228	0.284	0.279	**0.386**	0.217	0.270	0.278	**0.363**
Bora	0.061	0.193	0.186	**0.325**	0.069	0.191	0.193	**0.317**
Cabécar	0.073	0.135	0.138	**0.338**	0.070	0.137	0.125	**0.333**
Cashinahua	0.084	0.138	0.139	**0.335**	0.083	0.138	0.138	**0.313**
Dolgan	0.091	0.208	0.215	**0.335**	0.097	0.191	0.199	**0.303**
Evenki	0.088	0.196	0.212	**0.283**	0.080	0.191	0.198	**0.273**
Gorwaa	0.146	0.187	0.182	**0.351**	0.140	0.197	0.186	**0.342**
Jahai	0.102	0.201	0.201	**0.323**	0.110	0.201	0.201	**0.320**
Jejuan	0.132	0.183	0.188	**0.302**	0.131	0.178	0.175	**0.297**
Kakabe	0.081	0.202	0.197	**0.418**	0.080	0.190	0.195	**0.401**
Kamas	0.114	0.260	0.257	**0.377**	0.109	0.248	0.244	**0.369**
Komnzo	0.140	0.248	0.236	**0.337**	0.143	0.227	0.231	**0.313**
Light Warlpiri	0.105	0.169	0.170	**0.309**	0.109	0.183	0.174	**0.306**
Dalabon	0.125	0.248	0.236	**0.323**	0.131	0.239	0.248	**0.323**
Nisvai	0.095	0.210	0.210	**0.327**	0.099	0.186	0.192	**0.300**
Nllng	0.155	0.152	0.151	**0.315**	0.160	0.149	0.146	**0.314**
Northern Kurdish (Kurmanji)	0.062	0.196	0.192	**0.335**	0.061	0.181	0.186	**0.316**
Northern Alta	0.119	0.218	0.223	**0.301**	0.120	0.205	0.208	**0.268**
Fanbyak	0.121	0.219	0.223	**0.364**	0.114	0.202	0.207	**0.346**
Pnar	0.051	0.161	0.173	**0.352**	0.051	0.164	0.165	**0.325**
Daakie	0.093	0.188	0.175	**0.316**	0.093	0.169	0.167	**0.305**
Resígaro	0.064	0.232	0.241	**0.363**	0.073	0.224	0.234	**0.354**
Ruuli	0.120	0.228	0.214	**0.337**	0.124	0.221	0.223	**0.317**
Sadu	0.127	0.202	0.194	**0.355**	0.132	0.191	0.189	**0.358**
Sanzhi Dargwa	0.065	0.202	0.213	**0.313**	0.069	0.190	0.193	**0.297**
Savosavo	0.060	0.218	0.228	**0.430**	0.062	0.217	0.213	**0.410**
Nafsan (South Efate)	0.048	0.167	0.176	**0.355**	0.048	0.163	0.177	**0.325**
English (Southern England)	0.077	0.165	0.164	**0.306**	0.078	0.168	0.171	**0.302**
French (Swiss)	0.145	0.192	0.185	**0.338**	0.148	0.189	0.176	**0.327**
Sümi	0.119	0.187	0.183	**0.288**	0.119	0.176	0.181	**0.275**
Svan	0.079	0.221	0.221	**0.352**	0.074	0.205	0.216	**0.325**
Tabasaran	0.079	0.188	0.204	**0.336**	0.079	0.183	0.197	**0.334**
Teop	0.117	0.233	0.245	**0.402**	0.109	0.223	0.226	**0.393**
Texistepec Popoluca	0.121	0.160	0.162	**0.221**	0.127	0.153	0.154	**0.235**
Mojeño Trinitario	0.073	0.244	0.255	**0.384**	0.074	0.241	0.246	**0.378**
Asimjeeg Datooga	0.123	0.224	0.230	**0.318**	0.120	0.210	0.217	**0.310**
Urum	0.070	0.203	0.212	**0.342**	0.069	0.201	0.200	**0.326**
Vera'a	0.080	0.199	0.187	**0.376**	0.080	0.189	0.187	**0.354**
Warlpiri	0.084	0.267	0.265	**0.355**	0.087	0.239	0.236	**0.344**
Yongning Na	0.084	0.204	0.212	**0.368**	0.082	0.195	0.198	**0.352**
Yucatec Maya	0.076	0.156	0.150	**0.274**	0.073	0.150	0.148	**0.262**

Table 13: Language-wise multilingual WACS scores on DoReCo. Higher values are better. * Trained in this work.

Language	MMS				XLSR			
	Base	MMS-DORECO*	MMS-DORECO*-SIL	Gold	Base	MMS-DORECO*	MMS-DORECO*-SIL	Gold
Anal	0.042	0.091	0.097	**0.104**	0.046	0.096	0.095	**0.105**
Yali (Apahapsili)	0.049	0.084	**0.085**	0.083	0.053	0.095	0.101	**0.111**
Arapaho	0.071	0.100	0.116	**0.136**	0.072	0.115	0.123	**0.151**
Baïnounk Gubëeher	0.037	0.092	0.102	**0.114**	0.038	0.093	0.097	**0.114**
Beja	0.084	0.104	0.092	**0.111**	0.083	0.118	0.119	**0.127**
Bora	0.043	0.118	0.117	**0.145**	0.044	0.120	0.119	**0.148**
Cabécar	0.044	0.125	0.127	**0.157**	0.041	0.119	0.121	**0.147**
Cashinahua	0.055	0.117	0.116	**0.153**	0.047	0.117	0.126	**0.153**
Dolgan	0.054	0.142	0.149	**0.153**	0.053	**0.146**	0.139	0.144
Evenki	0.031	0.091	0.090	**0.111**	0.037	0.082	0.086	**0.127**
Gorwaa	0.051	0.115	0.111	**0.125**	0.055	0.116	0.113	**0.127**
Jahai	0.039	0.113	0.109	**0.137**	0.046	0.111	0.107	**0.137**
Jejuan	0.045	0.097	0.091	**0.103**	0.064	0.116	0.124	**0.127**
Kakabe	0.037	0.140	0.141	**0.171**	0.028	0.137	0.140	**0.170**
Kamas	0.054	0.141	0.143	**0.178**	0.053	0.143	0.150	**0.177**
Komnzo	0.048	0.100	0.101	**0.112**	0.052	0.106	0.107	**0.121**
Light Warlpiri	0.021	0.086	0.083	**0.101**	0.022	0.091	0.086	**0.098**
Dalabon	0.056	**0.105**	0.099	0.085	0.075	0.100	**0.111**	0.103
Nisvai	0.053	0.114	0.116	**0.124**	0.056	0.112	0.120	**0.125**
Nllng	0.043	0.068	0.069	**0.103**	0.047	0.068	0.072	**0.112**
Northern Kurdish (Kurmanji)	0.022	0.133	0.140	**0.140**	0.027	0.129	0.129	**0.142**
Northern Alta	0.044	0.105	0.104	**0.111**	0.042	0.096	0.097	**0.104**
Fanbyak	0.038	0.105	0.104	**0.110**	0.044	**0.119**	0.114	0.118
Pnar	0.010	0.134	0.133	**0.153**	0.005	0.128	0.129	**0.153**
Daakie	0.024	0.089	0.095	**0.099**	0.025	0.101	0.099	**0.112**
Resígaro	0.024	0.127	0.141	**0.156**	0.022	0.134	0.157	**0.162**
Ruuli	0.056	**0.115**	0.105	0.112	0.045	**0.121**	0.120	0.116
Sadu	0.042	0.107	0.113	**0.127**	0.054	0.106	0.109	**0.129**
Sanzhi Dargwa	0.031	0.169	0.171	**0.178**	0.011	0.142	0.141	**0.153**
Savosavo	0.021	0.132	0.137	**0.150**	0.018	0.128	0.134	**0.143**
Nafsan (South Efate)	0.002	0.109	0.115	**0.143**	0.004	0.104	0.116	**0.142**
English (Southern England)	0.014	0.116	0.115	**0.121**	0.007	0.125	0.126	**0.139**
French (Swiss)	0.017	0.078	0.077	**0.104**	0.018	0.085	0.077	**0.101**

Language	MMS				XLSR			
	Base	MMS-DORECO*	MMS-DORECO*-SIL	Gold	Base	MMS-DORECO*	MMS-DORECO*-SIL	Gold
Sümi	0.042	0.077	0.079	**0.108**	0.043	0.073	0.073	**0.113**
Svan	0.013	0.108	0.127	**0.152**	0.012	0.109	0.115	**0.131**
Tabasaran	0.042	0.154	0.168	**0.179**	0.036	0.156	0.149	**0.168**
Teop	0.031	0.092	0.094	**0.103**	0.044	0.106	0.101	**0.121**
Texistepec Popoluca	0.016	0.064	0.057	**0.090**	0.015	0.057	0.057	**0.092**
Mojeño Trinitario	0.034	0.167	0.164	**0.183**	0.031	0.160	0.167	**0.176**
Asimjeeg Datooga	0.041	0.108	0.105	**0.127**	0.036	0.112	0.109	**0.133**
Urum	0.025	0.128	0.135	**0.173**	0.023	0.119	0.128	**0.153**
Vera'a	0.026	0.114	0.113	**0.134**	0.020	0.124	0.121	**0.145**
Warlpiri	0.045	0.109	0.111	**0.121**	0.049	**0.118**	0.116	0.107
Yongning Na	0.039	0.151	0.161	**0.199**	0.037	0.156	0.166	**0.201**
Yucatec Maya	0.032	0.112	0.113	**0.129**	0.028	0.108	0.115	**0.135**

Table 14: Language-wise multilingual AAS scores on DoReCo. Lower values are better. * Trained in this work.

Language	Base	MMS-DORECO*	MMS-DORECO*-SIL	Gold
Anal	143.4	81.8	79.4	**0.0**
Yali (Apahapsili)	91.6	63.6	62.8	**0.0**
Arapaho	529.8	298.2	250.1	**0.0**
Baïnounk Gubëeher	172.3	101.9	92.4	**0.0**
Beja	57.7	52.8	53.1	**0.0**
Bora	426.1	178.0	171.2	**0.0**
Cabécar	230.7	111.7	113.5	**0.0**
Cashinahua	308.7	122.7	107.9	**0.0**
Dolgan	257.8	108.5	99.0	**0.0**
Evenki	418.9	189.0	171.4	**0.0**
Gorwaa	137.0	76.1	72.2	**0.0**
Jahai	241.5	118.0	111.8	**0.0**
Jejuan	167.7	116.1	115.0	**0.0**
Kakabe	301.8	104.1	91.8	**0.0**
Kamas	317.6	166.1	155.7	**0.0**
Komnzo	86.6	60.6	61.2	**0.0**
Light Warlpiri	183.9	88.0	87.5	**0.0**
Dalabon	147.8	111.8	112.4	**0.0**
Nisvai	124.0	67.6	68.8	**0.0**
Nllng	155.8	123.8	125.5	**0.0**
Northern Kurdish (Kurmanji)	231.0	70.1	62.1	**0.0**
Northern Alta	158.8	106.2	106.1	**0.0**
Fanbyak	110.5	72.7	74.5	**0.0**
Pnar	567.0	160.3	148.7	**0.0**
Daakie	215.6	83.6	81.4	**0.0**
Resígaro	582.3	168.0	134.7	**0.0**
Ruuli	133.3	77.2	71.1	**0.0**
Sadu	121.7	82.6	86.3	**0.0**
Sanzhi Dargwa	480.9	154.7	143.0	**0.0**
Savosavo	321.5	82.1	70.5	**0.0**
Nafsan (South Efate)	707.5	274.6	246.4	**0.0**
English (Southern England)	350.2	141.3	138.2	**0.0**
French (Swiss)	142.6	77.1	78.3	**0.0**
Sümi	109.5	80.7	80.9	**0.0**
Svan	302.4	104.0	93.9	**0.0**
Tabasaran	275.8	92.8	82.2	**0.0**
Teop	143.1	74.3	69.5	**0.0**
Texistepec Popoluca	224.0	141.7	133.2	**0.0**
Mojeño Trinitario	422.0	138.6	128.6	**0.0**
Asimjeeg Datooga	144.3	91.0	93.7	**0.0**
Urum	456.7	171.7	154.5	**0.0**
Vera'a	202.1	68.5	67.6	**0.0**
Warlpiri	242.8	113.2	100.9	**0.0**
Yongning Na	331.4	133.7	121.8	**0.0**
Yucatec Maya	253.0	97.1	91.4	**0.0**

Table 15: Multilingual statistics across aligners and embedding models. * Trained in this work.

Aligner	Metric	MMS				XLSR			
		Mean	Std	Min	Max	Mean	Std	Min	Max
Baseline (FLEURS)	PCMI	0.091	0.017	0.048	0.148	0.093	0.016	0.053	0.148
	Mean cluster purity	0.135	0.022	0.083	0.183	0.138	0.026	0.081	0.194
	Std cluster purity	0.140	0.041	0.036	0.226	0.146	0.049	0.047	0.270
	Cluster entropy	3.834	0.024	3.750	3.875	3.853	0.014	3.811	3.887
	Norm. cluster entropy	0.980	0.006	0.959	0.991	0.985	0.004	0.974	0.994
	Label entropy	3.501	0.206	3.047	3.946	3.501	0.206	3.047	3.946
	Norm. label entropy	0.893	0.024	0.839	0.963	0.893	0.024	0.839	0.963
	Clusters used	50.000	0.000	50.000	50.000	50.000	0.000	50.000	50.000
	Labels	51.635	11.029	31.000	81.000	51.635	11.029	31.000	81.000
	Frames	9589.118	137.969	9158.000	9921.000	9589.118	137.969	9158.000	9921.000
	WACS	0.022	0.014	-0.011	0.058	0.020	0.013	-0.012	0.056
	Positive DTW	0.402	0.023	0.337	0.455	0.435	0.025	0.369	0.487
	Negative DTW	0.379	0.019	0.338	0.420	0.415	0.017	0.373	0.461

Aligner	Metric	MMS				XLSR			
		Mean	Std	Min	Max	Mean	Std	Min	Max
	Positive std	0.112	0.014	0.087	0.160	0.117	0.012	0.095	0.155
	Negative std	0.078	0.009	0.063	0.110	0.086	0.009	0.074	0.128
	Positive min	0.152	0.032	0.069	0.236	0.139	0.033	0.063	0.199
	Negative min	0.167	0.034	0.096	0.260	0.156	0.034	0.060	0.224
	Positive max	0.875	0.037	0.769	0.969	0.884	0.036	0.782	0.969
	Negative max	0.758	0.077	0.605	0.912	0.768	0.075	0.619	0.947
	Positive pairs	1507.247	279.933	208.000	2000.000	1507.247	279.933	208.000	2000.000
	Negative pairs	980.835	89.681	184.000	1000.000	981.035	89.435	186.000	998.000
	Vocabulary size	197.200	17.896	38.000	200.000	197.200	17.896	38.000	200.000
	Cohen's d	0.219	0.128	-0.116	0.487	0.189	0.118	-0.112	0.467
	SNR	0.116	0.069	-0.060	0.267	0.099	0.062	-0.058	0.252
MFA (FLEURS)	PCMI	0.247	0.105	0.070	0.432	0.231	0.095	0.067	0.397
	Mean cluster purity	0.405	0.135	0.209	0.972	0.402	0.137	0.219	0.967
	Std cluster purity	0.195	0.043	0.078	0.323	0.204	0.047	0.080	0.329
	Cluster entropy	3.841	0.026	3.726	3.876	3.848	0.023	3.698	3.878
	Norm. cluster entropy	0.982	0.007	0.952	0.991	0.984	0.006	0.945	0.991
	Label entropy	2.794	0.605	0.681	3.532	2.794	0.605	0.681	3.532
	Norm. label entropy	0.746	0.152	0.235	0.913	0.746	0.152	0.235	0.913
	Clusters used	50.000	0.000	50.000	50.000	50.000	0.000	50.000	50.000
	Labels	42.952	9.809	18.000	71.000	42.952	9.809	18.000	71.000
	Frames	8234.723	1798.210	299.000	9781.000	8234.723	1798.210	299.000	9781.000
	WACS	0.143	0.077	-0.001	0.258	0.127	0.066	-0.017	0.227
	Positive DTW	0.529	0.086	0.379	0.653	0.556	0.083	0.407	0.682
	Negative DTW	0.386	0.024	0.339	0.473	0.429	0.026	0.373	0.505
	Positive std	0.119	0.018	0.088	0.180	0.117	0.018	0.085	0.167
	Negative std	0.078	0.014	0.060	0.131	0.085	0.018	0.069	0.177
	Positive min	0.197	0.068	0.051	0.516	0.204	0.080	0.061	0.513
	Negative min	0.167	0.044	0.050	0.338	0.172	0.054	0.030	0.328
	Positive max	0.881	0.041	0.689	0.954	0.892	0.038	0.712	0.958
	Negative max	0.679	0.070	0.540	0.880	0.713	0.061	0.575	0.887
	Positive pairs	978.341	550.726	27.000	1932.000	978.341	550.726	27.000	1932.000
	Negative pairs	687.073	322.087	12.000	998.000	686.793	321.992	12.000	999.000
	Vocabulary size	138.366	64.396	4.000	200.000	138.366	64.396	4.000	200.000
	Cohen's d	1.444	0.852	-0.005	2.875	1.294	0.746	-0.113	2.709
	SNR	0.766	0.452	-0.003	1.487	0.674	0.388	-0.056	1.377
MMS-300M-IPA* (FLEURS)	PCMI	0.239	0.028	0.183	0.329	0.226	0.026	0.180	0.305
	Mean cluster purity	0.257	0.037	0.168	0.344	0.248	0.032	0.174	0.328
	Std cluster purity	0.163	0.031	0.085	0.246	0.177	0.034	0.092	0.251
	Cluster entropy	3.846	0.017	3.780	3.875	3.856	0.012	3.827	3.881
	Norm. cluster entropy	0.983	0.004	0.966	0.990	0.986	0.003	0.978	0.992
	Label entropy	3.484	0.205	3.010	3.929	3.484	0.205	3.010	3.929
	Norm. label entropy	0.889	0.023	0.834	0.949	0.889	0.023	0.834	0.949
	Clusters used	50.000	0.000	50.000	50.000	50.000	0.000	50.000	50.000
	Labels	51.529	10.981	31.000	81.000	51.529	10.981	31.000	81.000
	Frames	9561.271	143.986	9100.000	9856.000	9561.271	143.986	9100.000	9856.000
	WACS	0.195	0.025	0.128	0.252	0.170	0.022	0.122	0.219
	Positive DTW	0.587	0.028	0.492	0.655	0.607	0.027	0.515	0.668
	Negative DTW	0.392	0.018	0.353	0.434	0.436	0.018	0.393	0.489
	Positive std	0.118	0.012	0.090	0.161	0.118	0.011	0.095	0.149
	Negative std	0.072	0.004	0.062	0.086	0.078	0.004	0.072	0.102
	Positive min	0.206	0.047	0.065	0.301	0.215	0.049	0.079	0.307
	Negative min	0.177	0.033	0.102	0.264	0.196	0.036	0.098	0.280
	Positive max	0.898	0.022	0.826	0.945	0.909	0.019	0.877	0.953
	Negative max	0.685	0.045	0.588	0.824	0.717	0.039	0.633	0.832
	Positive pairs	1506.424	279.898	208.000	2000.000	1506.424	279.898	208.000	2000.000
	Negative pairs	980.741	89.333	186.000	1000.000	981.106	89.559	184.000	1000.000
	Vocabulary size	197.188	17.897	38.000	200.000	197.188	17.897	38.000	200.000
	Cohen's d	1.923	0.290	1.091	2.702	1.647	0.247	1.026	2.345
	SNR	1.032	0.146	0.614	1.405	0.871	0.124	0.561	1.209
Wav2Vec2-IPA* (FLEURS)	PCMI	0.237	0.025	0.185	0.322	0.225	0.023	0.180	0.309
	Mean cluster purity	0.255	0.035	0.174	0.355	0.246	0.033	0.182	0.343
	Std cluster purity	0.161	0.028	0.094	0.220	0.174	0.032	0.110	0.265
	Cluster entropy	3.849	0.015	3.795	3.882	3.855	0.014	3.812	3.886
	Norm. cluster entropy	0.984	0.004	0.970	0.992	0.985	0.004	0.975	0.993
	Label entropy	3.484	0.203	3.023	3.924	3.484	0.203	3.023	3.924
	Norm. label entropy	0.889	0.023	0.823	0.946	0.889	0.023	0.823	0.946
	Clusters used	50.000	0.000	50.000	50.000	50.000	0.000	50.000	50.000
	Labels	51.565	10.967	31.000	81.000	51.565	10.967	31.000	81.000
	Frames	9559.365	144.804	9081.000	9847.000	9559.365	144.804	9081.000	9847.000
	WACS	0.193	0.026	0.124	0.252	0.168	0.023	0.114	0.217
	Positive DTW	0.585	0.029	0.484	0.650	0.604	0.028	0.507	0.663
	Negative DTW	0.392	0.019	0.355	0.437	0.436	0.018	0.392	0.486
	Positive std	0.119	0.012	0.090	0.158	0.118	0.011	0.094	0.149
	Negative std	0.071	0.004	0.064	0.087	0.078	0.005	0.069	0.104
	Positive min	0.211	0.043	0.098	0.306	0.215	0.050	0.079	0.309
	Negative min	0.182	0.033	0.100	0.268	0.197	0.037	0.050	0.264
	Positive max	0.896	0.021	0.845	0.941	0.907	0.019	0.867	0.949
	Negative max	0.680	0.038	0.606	0.791	0.717	0.042	0.637	0.840
	Positive pairs	1505.741	279.817	208.000	2000.000	1505.741	279.817	208.000	2000.000
	Negative pairs	980.882	89.777	183.000	999.000	980.882	89.230	189.000	1000.000
	Vocabulary size	197.176	17.916	38.000	200.000	197.176	17.916	38.000	200.000
	Cohen's d	1.900	0.295	1.054	2.687	1.627	0.249	0.976	2.280

Aligner	Metric	MMS				XLSR			
		Mean	Std	Min	Max	Mean	Std	Min	Max
	SNR	1.022	0.149	0.592	1.396	0.861	0.126	0.540	1.168
Qwen3-FA (FLEURS)	PCMI	–	–	–	–	–	–	–	–
	Mean cluster purity	–	–	–	–	–	–	–	–
	Std cluster purity	–	–	–	–	–	–	–	–
	Cluster entropy	–	–	–	–	–	–	–	–
	Norm. cluster entropy	–	–	–	–	–	–	–	–
	Label entropy	–	–	–	–	–	–	–	–
	Norm. label entropy	–	–	–	–	–	–	–	–
	Clusters used	–	–	–	–	–	–	–	–
	Labels	–	–	–	–	–	–	–	–
	Frames	–	–	–	–	–	–	–	–
	WACS	0.205	0.020	0.172	0.244	0.184	0.015	0.161	0.207
	Positive DTW	0.594	0.022	0.571	0.648	0.613	0.022	0.581	0.661
	Negative DTW	0.389	0.016	0.369	0.419	0.430	0.017	0.404	0.462
	Positive std	0.104	0.009	0.092	0.127	0.105	0.011	0.088	0.132
	Negative std	0.072	0.005	0.067	0.081	0.078	0.005	0.073	0.085
	Positive min	0.232	0.063	0.081	0.335	0.255	0.083	0.042	0.378
	Negative min	0.171	0.040	0.070	0.222	0.161	0.048	0.044	0.232
	Positive max	0.905	0.009	0.893	0.919	0.916	0.015	0.893	0.940
	Negative max	0.664	0.048	0.603	0.778	0.697	0.034	0.667	0.757
	Positive pairs	1586.700	174.696	1353.000	2000.000	1586.700	174.696	1353.000	2000.000
	Negative pairs	994.600	2.691	990.000	998.000	995.700	2.326	992.000	998.000
	Vocabulary size	200.000	0.000	200.000	200.000	200.000	0.000	200.000	200.000
	Cohen's d	2.227	0.287	1.865	2.875	1.945	0.241	1.617	2.507
	SNR	1.175	0.144	1.007	1.489	1.013	0.122	0.867	1.286
Baseline (DoReCo)	PCMI	0.100	0.034	0.048	0.228	0.100	0.033	0.048	0.217
	Mean cluster purity	0.186	0.045	0.104	0.274	0.184	0.046	0.098	0.277
	Std cluster purity	0.087	0.025	0.033	0.139	0.083	0.034	0.021	0.204
	Cluster entropy	3.823	0.024	3.750	3.876	3.833	0.021	3.779	3.874
	Norm. cluster entropy	0.977	0.006	0.959	0.991	0.980	0.005	0.966	0.990
	Label entropy	3.084	0.197	2.617	3.451	3.084	0.197	2.617	3.451
	Norm. label entropy	0.885	0.031	0.815	0.954	0.885	0.031	0.815	0.954
	Clusters used	50.000	0.000	50.000	50.000	50.000	0.000	50.000	50.000
	Labels	33.111	5.618	23.000	45.000	33.111	5.618	23.000	45.000
	Frames	6790.889	2156.754	2782.000	9804.000	6790.889	2156.754	2782.000	9804.000
	WACS	0.038	0.016	0.002	0.084	0.038	0.018	0.004	0.083
	Positive DTW	0.557	0.068	0.393	0.705	0.559	0.066	0.398	0.699
	Negative DTW	0.519	0.059	0.383	0.637	0.521	0.054	0.394	0.625
	Positive std	0.098	0.011	0.077	0.118	0.105	0.012	0.082	0.131
	Negative std	0.072	0.008	0.055	0.090	0.077	0.010	0.054	0.100
	Positive min	0.306	0.080	0.107	0.466	0.278	0.086	0.124	0.468
	Negative min	0.316	0.081	0.130	0.498	0.291	0.077	0.119	0.439
	Positive max	0.848	0.038	0.758	0.906	0.855	0.039	0.752	0.917
	Negative max	0.741	0.045	0.607	0.831	0.753	0.040	0.641	0.844
	Positive pairs	592.867	291.575	103.000	1225.000	592.867	291.575	103.000	1225.000
	Negative pairs	421.933	195.032	76.000	889.000	422.111	196.041	71.000	891.000
	Vocabulary size	85.467	39.028	16.000	179.000	85.467	39.028	16.000	179.000
	Cohen's d	0.430	0.196	0.025	1.130	0.410	0.211	0.039	0.998
	SNR	0.223	0.102	0.013	0.579	0.213	0.112	0.020	0.538
MMS-300M-DoReCo*	PCMI	0.202	0.032	0.135	0.284	0.194	0.028	0.137	0.270
	Mean cluster purity	0.326	0.052	0.212	0.440	0.316	0.048	0.199	0.411
	Std cluster purity	0.166	0.040	0.075	0.251	0.156	0.035	0.092	0.254
	Cluster entropy	3.823	0.025	3.758	3.869	3.835	0.020	3.790	3.868
	Norm. cluster entropy	0.977	0.006	0.961	0.989	0.980	0.005	0.969	0.989
	Label entropy	2.922	0.206	2.586	3.410	2.922	0.206	2.586	3.410
	Norm. label entropy	0.841	0.042	0.729	0.920	0.841	0.042	0.729	0.920
	Clusters used	50.000	0.000	50.000	50.000	50.000	0.000	50.000	50.000
	Labels	32.733	5.535	23.000	45.000	32.733	5.535	23.000	45.000
	Frames	6724.644	2092.506	2782.000	9804.000	6724.644	2092.506	2782.000	9804.000
	WACS	0.112	0.024	0.064	0.169	0.114	0.022	0.057	0.160
	Positive DTW	0.639	0.045	0.518	0.740	0.647	0.042	0.534	0.741
	Negative DTW	0.527	0.058	0.384	0.636	0.533	0.052	0.406	0.635
	Positive std	0.090	0.010	0.063	0.111	0.095	0.011	0.070	0.118
	Negative std	0.070	0.008	0.044	0.084	0.074	0.007	0.053	0.091
	Positive min	0.374	0.069	0.216	0.546	0.360	0.077	0.189	0.528
	Negative min	0.329	0.077	0.178	0.523	0.312	0.080	0.109	0.457
	Positive max	0.872	0.024	0.816	0.912	0.886	0.026	0.812	0.929
	Negative max	0.724	0.043	0.596	0.798	0.746	0.049	0.621	0.844
	Positive pairs	588.733	290.884	103.000	1218.000	588.733	290.884	103.000	1218.000
	Negative pairs	419.622	195.226	75.000	891.000	419.178	194.474	75.000	887.000
	Vocabulary size	84.889	38.995	16.000	179.000	84.889	38.995	16.000	179.000
	Cohen's d	1.368	0.260	0.827	2.031	1.319	0.240	0.718	1.847
	SNR	0.704	0.135	0.426	1.058	0.678	0.123	0.366	0.951
MMS-300M-DoReCo-SIL*	PCMI	0.203	0.032	0.138	0.279	0.196	0.030	0.125	0.278
	Mean cluster purity	0.345	0.055	0.223	0.466	0.336	0.052	0.231	0.418
	Std cluster purity	0.187	0.040	0.089	0.259	0.176	0.037	0.102	0.265
	Cluster entropy	3.821	0.024	3.772	3.868	3.833	0.021	3.788	3.879
	Norm. cluster entropy	0.977	0.006	0.964	0.989	0.980	0.005	0.968	0.992
	Label entropy	2.876	0.222	2.476	3.396	2.876	0.222	2.476	3.396
	Norm. label entropy	0.828	0.048	0.706	0.910	0.828	0.048	0.706	0.910
	Clusters used	50.000	0.000	50.000	50.000	50.000	0.000	50.000	50.000
	Labels	32.711	5.528	23.000	45.000	32.711	5.528	23.000	45.000

Aligner	Metric	MMS				XLSR			
		Mean	**Std**	**Min**	**Max**	**Mean**	**Std**	**Min**	**Max**
	Frames	6701.978	2072.661	2782.000	9804.000	6701.978	2072.661	2782.000	9804.000
	WACS	0.114	0.026	0.057	0.171	0.116	0.023	0.057	0.167
	Positive DTW	0.641	0.044	0.524	0.738	0.649	0.041	0.542	0.739
	Negative DTW	0.527	0.059	0.390	0.647	0.533	0.050	0.413	0.630
	Positive std	0.089	0.010	0.062	0.110	0.093	0.010	0.070	0.117
	Negative std	0.070	0.008	0.044	0.086	0.072	0.006	0.055	0.086
	Positive min	0.374	0.068	0.216	0.546	0.365	0.074	0.145	0.528
	Negative min	0.324	0.078	0.175	0.536	0.319	0.075	0.131	0.470
	Positive max	0.872	0.025	0.816	0.918	0.886	0.027	0.812	0.950
	Negative max	0.736	0.040	0.642	0.817	0.741	0.044	0.620	0.838
	Positive pairs	587.067	290.491	103.000	1215.000	587.067	290.491	103.000	1215.000
	Negative pairs	418.267	194.548	72.000	891.000	418.111	195.244	72.000	892.000
	Vocabulary size	84.644	38.892	16.000	179.000	84.644	38.892	16.000	179.000
	Cohen's d	1.394	0.269	0.766	1.989	1.365	0.254	0.747	2.024
	SNR	0.715	0.139	0.396	1.035	0.702	0.130	0.383	1.035