Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens

Nay San[1], Georgios Paraskevopoulos[2], Aryaman Arora[1], Xiluo He[1], Prabhjot Kaur[3], Oliver Adams[4], Dan Jurafsky[1]
[1]Stanford University; [2]Athena Research Center; [3]Wayne State University; [4]Atos zData
nay.san@stanford.edu

Abstract

While massively multilingual speech models like wav2vec 2.0 XLSR-128 can be directly fine-tuned for automatic speech recognition (ASR), downstream performance can still be relatively poor on languages that are under-represented in the pre-training data. Continued pre-training on 70–200 hours of untranscribed speech in these languages can help — but what about languages without that much recorded data? For such cases, we show that supplementing the target language with data from a similar, higher-resource 'donor' language can help. For example, continued pretraining on only 10 hours of low-resource Punjabi supplemented with 60 hours of donor Hindi is almost as good as continued pretraining on 70 hours of Punjabi. By contrast, sourcing data from less similar donors like Bengali does not improve ASR performance. To inform donor language selection, we propose a novel similarity metric based on the sequence distribution of induced acoustic units: the Acoustic Token Distribution Similarity (ATDS). Across a set of typologically different target languages (Punjabi, Galician, Iban, Setswana), we show that the ATDS between the target language and its candidate donors precisely predicts target language ASR performance.

1 Introduction

For developing automatic speech recognition (ASR), 'low resource' languages are typically classified as such based on the availability of transcribed speech. Untranscribed speech, texts, or reliable metadata about the language are often assumed to be easily obtainable. This assumption may not hold true for under-described languages with little digital representation. For such languages, we are interested in two questions: 1) does leveraging untranscribed speech from a similar, higher-resource 'donor' language for pre-trained model adaptation help improve speech recognition in the target language, and 2) how do we select the best donor?

These questions are of interest as ASR system development with little transcribed speech has become viable with multilingual pre-trained transformer models for speech (e.g. wav2vec 2.0 XLSR-128: Babu et al., 2022). Yet, as most languages are under-represented in the pre-training data, directly fine-tuning these models for ASR in the target language can yield lower performance compared to their well-represented counterparts (Conneau et al., 2023). While recent studies have shown the effectiveness of continued pre-training to adapt these models to the target language (Nowakowski et al., 2023; Paraskevopoulos et al., 2024), they involved using 70–200 hours of target language data. For some languages, it may be quite difficult to source this much speech data — even untranscribed.

Thus, in our first set of experiments, we investigated whether supplementing target language data with data from another language could be a viable approach for model adaptation via continued pre-training (CPT). We selected Punjabi as our target language to establish top-line performance when sufficient data *is* available (70 hours, approximating the setup in Paraskevopoulos et al., 2024), along with a limited data baseline (when only 10 hours of Punjabi is available). We compared this baseline to supplementing the 10 hours of Punjabi with 60 hours of data from 8 other Indic languages (Indo-Aryan: Hindi, Urdu, Gujarati, Marathi, Bengali, Odia; Dravidian: Malayalam, Tamil). We fine-tuned each CPT-adapted model using the same 1 hour of transcribed Punjabi speech.

Results indicated that adding data from unrelated Dravidian languages (Malayalam, Tamil) or dissimilar Indo-Aryan languages (Bengali, Odia) yielded no better than baseline performance, 25% word error rate (WER). By contrast, we observed improved WERs from adding more similar languages (Marathi, Urdu, Gujarati, Hindi), with adding Hindi coming close to the 70-hour Punjabi top-line: 23.2% vs. 22.2%, respectively.

Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024), pages 100–112
March 22, 2024 ©2024 Association for Computational Linguistics

In our second set of experiments, we investigated how well measures of similarity between the target and donor languages predicted target language ASR performance. We found that commonly used measures based on external typological databases such as lang2vec (Littell et al., 2017) were not sufficiently fine-grained for our use case and, crucially, also varied with the quality/completeness of the available metadata for a given target language.

To sidestep these issues, we propose the **Acoustic Token Distribution Similarity** (ATDS), which measures the degree of similarity for two untranscribed speech corpora based on frequencies of occurrence of recurring acoustic-phonetic sequences. This measure extends Token Distribution Similarity (Gogoulou et al., 2023), shown to correlate with positive transfer in continued pre-training for text-based language models. To account for the text-/token-less nature of untranscribed speech corpora, we induce them in a bottom-up manner using wav2seq (Wu et al., 2023), a method for inducing pseudo-tokens using pre-trained speech embeddings. We compared the ASR performance from the Indic language experiments to various similarity measures and found that ATDS offered the most accurate ranking. Furthermore, ATDS correctly predicted the best donor language between two options for three non-Indic low-resource languages (Galician, Iban, and Setswana).[1]

In sum, the main contributions of this paper are: 1) a systematic study of pairwise transfer between languages in continued pre-training and its effects on target language ASR performance, and 2) the development, analysis, and first validation of ATDS — a fine-grained, bottom-up measure of acoustic-phonetic similarity to predict this ASR performance. To facilitate reproducibility and further research, we make available all our code, model checkpoints, and experimental artefacts.[2]

2 Background: wav2vec 2.0

In this section, we provide a high-level overview of the wav2vec 2.0 model and highlight specific details about its architecture and training objectives that will be relevant to our later discussions. Developed by Baevski et al. (2020), wav2vec 2.0 is a type of *self-supervised pre-trained transformer model*. In machine learning, supervised learning involves the use of human-generated labels (e.g. transcriptions), which can be time- and cost-intensive to create. In self-supervised pre-training, the goal is to first train the model on a proxy task for which it can derive its own labels, before the resulting model is adapted or 'fine-tuned' to the target task (e.g. ASR). Leveraging self-supervised pre-training has shown remarkable success across a variety of tasks when combined with a transformer-based model (Vaswani et al., 2017), which excels at learning how various units in a sequence co-occur (e.g. words in a sentence). Naturally, this has spurred on much experimentation for leveraging such models for low-resource ASR (e.g. Coto-Solano et al., 2022; Guillaume et al., 2022; Macaire et al., 2022; Bartelds et al., 2023).

As illustrated below in Figure 1, the wav2vec 2.0 architecture consists of three parts: 1) a convolutional feature extractor that extracts learnable features from strided frames in the audio input, 2) a quantiser which clusters the audio features into into a set of discrete code vectors (q_1, ..., q_5), and 3) a transformer attention network that learns context-enriched representations (h_1, ..., h_5). For self-supervised pre-training, the model is optimised using a joint contrastive loss and a diversity loss. The contrastive loss requires the model to use the neighbouring context to distinguish each masked frame amongst a set of negative distractors.[3] The diversity loss requires the model to make equal use of all code vectors, preventing the model from relying on a small subset which can lead to trivial/sub-optimal solutions.

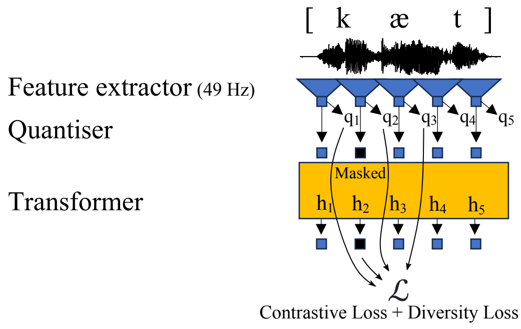


Figure 1: Illustration of the wav2vec 2.0 architecture. Adapted from Baevski et al. (2020).

We highlight here two important details relevant for our later discussions on acoustic tokens. The first is that the representations learned by the trans-

[1] Appendix B provides an overview of the languages.

[2] https://github.com/fauxneticien/w2v2-cpt-transfer

[3] A speech variant of a Cloze test: e.g. given "*The big* ___ *chased the small rat.*", select the correct answer from {*cat*, *rat*, *small*}.

former network are particularly useful for speech applications requiring fine-grained comparisons of acoustic-phonetic content, e.g. speech information retrieval (San et al., 2021), second language pronunciation scoring (Bartelds et al., 2022; Richter and Guðnason, 2023), spoken dialect classification (Bartelds and Wieling, 2022; Guillaume et al., 2023) — particularly those learned by the middle layers of the transformer network (e.g. layers 12–16 of a 24-layer network).

The second detail is that these representations are encoded as vectors in a high-dimensional latent space and emitted at a rate of 49 Hz. Combined with the diversity loss that requires exploration of this space, there is a many-to-many relationship between phonetic categories and these vectors — both in the latent space and in time. For example, as illustrated in Figure 1, a single speech sound such as [æ] may last for three time steps (yielding h_2, h_3, h_4). The first two vectors (h_2, h_3) may be very close in the latent space, as they map to the start of [æ] sounds while the third h_4 may map to [æ] sounds preceding [t] and is thus located in a different part of the latent space. Thus to derive phone-like tokens from these representations, we must group them in the latent space (e.g. using k-means clustering) and then again in time according to how the grouped units themselves routinely co-occur (e.g. using subword modelling).

In this way, the goal of tokenisation for both text and speech is driven by the need to derive units of a practical granularity based on the nature of the input: from coarser-grained words to finer-grained sub-words in text and, inversely, finer-grained sub-phones to coarser-grained phones in speech. We return to these two details in our development of the Acoustic Token Distribution Similarity measure for comparing the acoustic-phonetic similarity of two untranscribed speech corpora.

3 A systematic study of pairwise transfer

3.1 Motivation

Since the release of the original English wav2vec 2.0 model pre-trained on the 960 hour LibriSpeech corpus (Panayotov et al., 2015), additional massively multi-lingual variants have also been developed: XLSR-53, pre-trained on 56k hours from 53 languages (Conneau et al., 2021); XLSR-128, pre-trained on 436k hours from 128 languages (Babu et al., 2022); and MMS, pre-trained on 491k hours from 1,406 languages (Pratap et al., 2023). In each case, the vast majority of the pre-training data is drawn from European language sources.

Given the under-represented nature of most languages in these wav2vec 2.0 models, several studies have investigated continued pre-training (CPT) to adapt them for target languages (e.g. Javed et al., 2022; DeHaven and Billa, 2022; Nowakowski et al., 2023; Bartelds et al., 2023; Paraskevopoulos et al., 2024). Nowakowski et al. (2023) adapted the XLSR-53 model using 200 hours of Ainu. Using the same 40 minutes of transcribed Ainu for ASR fine-tuning, they found that the adapted model resulted in a 8.8% absolute word error rate (WER) decrease over the off-the-shelf XLSR-53 model. For many low resource languages, however, obtaining 200 hours of speech data may not be feasible.

In their study of unsupervised domain adaptation for Greek, Paraskevopoulos et al. (2024) found adapting the XLSR-53 model via CPT using a small 12-hour dataset of Greek read speech to be ineffective. However, they found that successful CPT-based adaptation could be achieved with the use of multi-domain data, e.g. 12 hours of read speech mixed with 70 hours of newscasts. Given these findings, we were motivated to investigate whether comparable results could be achieved by supplementing target language data with data from another language.

3.2 Method

3.2.1 Model training

As our primary interest was in examining how downstream ASR performance is affected by the dataset(s) used in continued pre-training (CPT), we carried out nearly identical experiments as those in Nowakowski et al. (2023) by obtaining their configuration file for wav2vec 2.0 pre-training using the official fairseq library.[4] Similarly, we follow the official fine-tuning recipe suitable for 1 hour of transcriptions. For both procedures, we made modifications to suit our hardware configuration and compute budget, as detailed in Appendix A.

3.2.2 Data

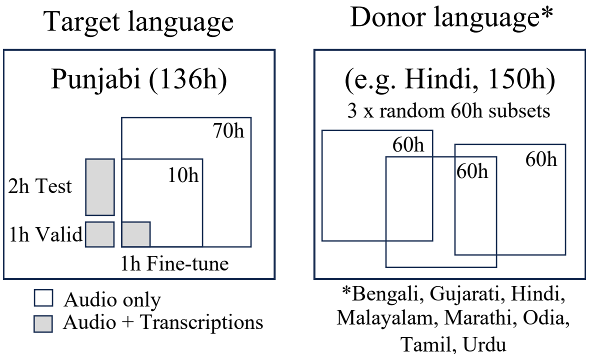
For a systematic investigation of how downstream ASR performance in a given target language varies with the choice of donor language added during CPT, we required a dataset with some specific characteristics: a) contains a variety of both similar and dissimilar languages with, b) at least 60 hours

[4] https://github.com/facebookresearch/fairseq

per language, c) and collected in relatively similar acoustic conditions, and d) not have already been used in the original pre-training process. Accordingly, for this set of experiments, we sourced data from IndicSUPERB (Javed et al., 2023), a dataset of 12 Indic languages containing 65–180 hours of read speech per language (Indo-Aryan: Bengali, Gujarati, Hindi, Marathi, Odia, Punjabi, Sanskrit, Urdu; Dravidian: Kannada, Malayalam, Tamil, Telugu).

Amongst these 12 IndicSUPERB languages, we selected Punjabi as our target language as it is relatively low-resourced (cf. Hindi, Tamil), its geographical and typological location (yielding a variety of closer and farther geographic/typological distances to the others), and also native speaker-linguist expertise on our research team for data validation and error analysis.

As illustrated below in Figure 2, we began by selecting for our top-line condition a random 70h subset of Punjabi from the total 136h available in IndicSUPERB, and then from this subset a random 10h selection for our baseline, and again a random 1h subset for fine-tuning. We also selected a random 1h validation and 2h test set both disjoint from each other and any data to be used for pre-training.



*Bengali, Gujarati, Hindi, Malayalam, Marathi, Odia, Tamil, Urdu

Figure 2: Data selection for transfer experiments

Additionally, for each donor language (Bengali, Gujarati, Hindi, Malayalam, Marathi, Odia, Tamil, Urdu), we created three random 60h subsets. Given the total amounts of data available in IndicSUPERB for each language (e.g. 87h for Urdu but 129h for Gujarati), there is however some unavoidable overlap between these subsets for each language (i.e. they are not disjoint 60h splits). Using each of the 60h subsets, we conducted 3 separate CPT runs per language to obtain estimates for both within- and between-donor language differences on downstream ASR performance.

3.3 Results and discussion

Compared to directly fine-tuning the XLSR-128 model, adapting the model first via continued pre-training (CPT) with 70 hours of Punjabi yields a large improvement in downstream ASR performance (unadapted 30.8% vs. 22.2% CPT-adapted, 70h). This constitutes an absolute WER difference of 8.6% and is consistent with improvements reported in previous CPT experiments (Nowakowski et al., 2023; Paraskevopoulos et al., 2024). We found that model adaptation with only 10 hours of Punjabi still yielded an appreciable improvement over the unadapted model: 5.8% absolute (unadapted 30.8% vs. 25.0% CPT-adapted, 10h).

We now turn to our experiment conditions in which 10h of Punjabi is supplemented with 60h of data from another language. As summarised below in Table 1, the effects of donor language on Punjabi ASR performance can be divided into roughly three strata. In the bottom stratum (E3), we find unrelated Dravidian languages (Tamil, Malayalam) or dissimilar Indo-Aryan languages (Bengali, Odia). When data from these languages are added during CPT, we find no meaningful difference compared to using only 10 hours of Punjabi (WERR: -0.8–0.0%).

In the middle stratum (E2), we find relatively similar Indo-Aryan languages (Marathi, Gujarati, Urdu). When data from these languages are added, we find modest improvements over using only 10 hours of Punjabi (WERR: 1.6–2.4%). In the top stratum (E1), we find Hindi — the most similar Indo-Aryan language amongst our candidate donors. When data from Hindi is added, we find a large improvement over using only 10 hours of Punjabi (WERR: 6.0%). In fact, adding the 60h of Hindi results in ASR performance that is in absolute terms close to the 70h Punjabi topline: 23.5% vs. 22.2%, respectively.

We have established that there are observable differences in target language ASR performance that vary with the donor language added during continued pre-training and that these differences appear to align with qualitative notions of language similarity. In the next section, we investigate quantitative measures of similarity and evaluate their correlations to these observed differences.

4 A bottom-up approach to similarity

4.1 Motivation

A common method for calculating similarities between languages is to use language vectors from the

Condition	Test set WER (WERR)		Data for continued pre-training
	Median	Range	
T. In-domain top-line	22.2 (11.2%)	-	70h Punjabi
E1. Most similar	**23.5 (6.0%)**	23.4–23.8	10h Punjabi + 60h Hindi
E2. Similar	24.4 (2.4%)	24.3–24.5	10h Punjabi + 60h Urdu
	24.4 (2.4%)	24.2–24.4	10h Punjabi + 60h Gujarati
	24.6 (1.6%)	24.5–24.7	10h Punjabi + 60h Marathi
B. Only target data baseline	25.0	-	10h Punjabi
E3. Unrelated/dissimilar	25.0 (0.0%)	25.0–25.2	10h Punjabi + 60h Odia
	25.1 (-0.4%)	25.0–25.4	10h Punjabi + 60h Tamil
	25.1 (-0.4%)	25.0–25.3	10h Punjabi + 60h Malayalam
	25.2 (-0.8%)	25.1–25.2	10h Punjabi + 60h Bengali
U. Unadapted XLSR-128	30.8 (-23.2%)	-	-

Table 1: Automatic speech recognition (ASR) results from fine-tuning wav2vec 2.0 XLSR-128 (Babu et al., 2022) with and without adaptation via continued pre-training (CPT). CPT-adapted models were trained for 10k updates using 70 hours of Punjabi for the topline (T), 10 hours of Punjabi for the baseline (B), and 10 hours of Punjabi combined with 60 hours of data from another language for the experiment conditions (E1, E2, E3). All models were fine-tuned with the same 1 hour of Punjabi data. ASR performance reported in word error rate (WER) and relative word error rates (WERR), relative to the 10 hour CPT baseline (B). For each experiment condition, median and range were obtained from 3 CPT runs per language with different donor data in each run.

lang2vec database (Littell et al., 2017), which itself draws on other databases (e.g. phonological information from WALS: Haspelmath, 2009). Wu et al. (2021) investigated how well measures based on lang2vec and other data sources correlated with successful transfer learning for ASR. Of the lang2vec similarity metrics, they found that genetic and geographic measures correlated highly with better ASR performance but, surprisingly, inventory and phonological measures did not. Acoustic similarities as derived from embeddings of a pre-trained spoken language identification model were also found to correlate strongly with better ASR performance. In the context of continued pre-training, we questioned whether the to-be-adapted model could be used for this purpose.

Investigating an analogous question in the text domain, Gogoulou et al. (2023) evaluated various measures for predicting transfer characteristics for transformer language models initially pre-trained on one language (e.g. English) and subsequently adapted to another (e.g. Icelandic), and how these characteristics varied according to the distributions of data in the respective language corpora. They propose a novel metric: the Token Distribution Similarity (TDS), which correlated with positive transfer. As illustrated below in Figure 3 (a), the TDS is derived by 1) using the pre-trained model's tokeniser to process a sample of data from each language, 2) then generating a token frequency vector for each language, and 3) taking the cosine similarity between these two vectors. Given these promising results for predicting positive cross-lingual transfer for continued pre-training on text, we investigated whether they extended to the speech domain.

4.2 Induction and analysis of acoustic tokens

In order to compute token distribution similarity for two untranscribed speech corpora, we first need to 'tokenise' the corpora. For this purpose, we can leverage speech representations extracted using a middle transformer layer (e.g. Layer 12 of 24) of a pre-trained wav2vec 2.0 model such as XLSR-128. As previously highlighted above (in §2), these representations are useful for fine-grained comparisons of acoustic-phonetic content and, to make use for these representations for inducing phone-like tokens, they must first be grouped in the high-dimensional latent space and then again in time based on their co-occurrences.

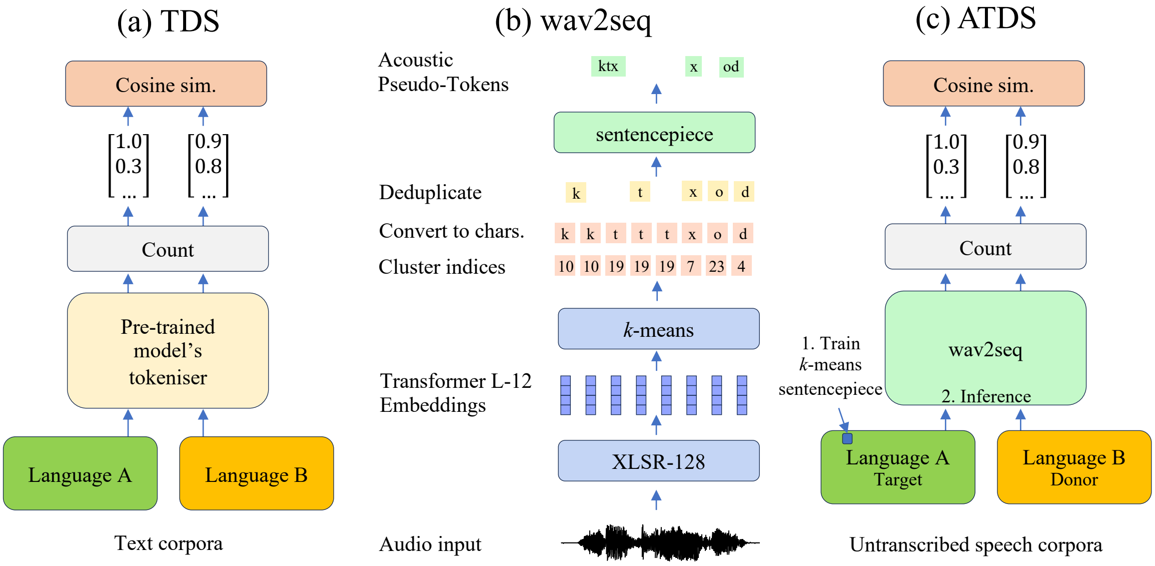


Figure 3: Derivation of the Acoustic Token Distribution Similarity (ATDS) measure for predicting positive transfer between two languages resulting from continued pre-training (CPT) of a pre-trained speech model (e.g. XLSR-128). ATDS extends to the speech domain the concept of Token Distribution Similarity (TDS: Gogoulou et al., 2023), shown to predict positive transfer for CPT in the text domain. To account for the token-less nature of untranscribed speech, pseudo-tokens are derived using the wav2seq process proposed by Wu et al. (2023). As is the case for text tokenisation, the goal is to derive units of practical granularity based on the raw input. While typical text tokenisation sub-divides words into sub-words (e.g. *eating* → *eat,ing*), the analogous process for raw speech data involves grouping sub-frames into more phone-like units: first based on featural similarity (e.g. using a k-means model on embeddings) then again based on distributional similarity (e.g. using a sub-word modelling).

Accordingly, we use the wav2seq procedure proposed by Wu et al. (2023).[5] As illustrated below in Figure 3 (b), the first step is using a pre-trained model (e.g. XLSR-128) to extract speech embeddings from a transformer layer. The second step involves training a k-means model to cluster these embeddings (i.e. group similar sounds together). The cluster indices are converted to characters (by a simple Unicode table lookup, e.g.10→*k*, 19→*t*, etc.) and then deduplicated. To discover frequently occurring sound sequences, the third step involves using these character strings to train a subword model (e.g. via sentencepiece: Kudo and Richardson, 2018). Once these models are trained on a subset of the corpus, they are used to derive pseudo-tokens for the rest of the corpus and, in our application, also for candidate donor corpora.

To discover whether these pseudo-tokens exhibited both within- and cross-language consistency, we conducted an analysis using the Punjabi and Hindi datasets of the CommonVoice corpus (Ardila et al., 2019), for which forced-aligned phoneme labels are available in the VoxCommunis corpus (Ahn and Chodroff, 2022). For this analysis, we trained the k-means and subword models on Punjabi (the target language) and then induced pseudo-tokens for both Punjabi and Hindi.[6]

Results of our analyses revealed that the most frequent pseudo-token in Punjabi, t_1, consistently corresponded to the /ɑ/ label, i.e. $P(/\text{ɑ}/|t_1) = 0.69$. Similarly, we found that t_2 consistently corresponded to high-back vowels: i.e. $P(/\text{o}/|t_2) = 0.56$, $P(/\text{u}/|t_2) = 0.18$, $P(/\text{ʊ}/|t_2) = 0.09$. For Hindi, we found that t_2 also consistently corresponded to the same vowel labels, while t_1 consistently corresponded to /aː/ — also a low vowel. Such minor differences are likely attributable to VoxCommunis labels being automatically derived via grapheme-to-phoneme conversion and thus do not represent narrow phonetic transcriptions. Given their broad categorical consistency, these tokens may be useful for cross-language comparisons.

4.3 Acoustic Token Distribution Similarity

We propose the *Acoustic* Token Distribution Similarity (ATDS) measure, which, as illustrated above in Figure 3 (c), is a straightforward composition of

[5] Originally developed for inducing pseudo-tokens for use in a self-supervised pseudo speech recognition task for jointly pre-training a speech encoder and text decoder.

[6] Appendix A details the analysis procedure.

	Punjabi (PAN)									
	Donor Lang.	Median WERR (of 3 runs)	*Similarity Measure*							
			ATDS	SB	lang2vec					
					Syn.	Geo.	Feat.	Inv.	Gen.	Phon.
E1.	Hindi	6.0	0.96	0.96	0.67	1.0*	0.6	0.67	0.38	0.41
E2.	Gujarati	2.4	0.93	0.82	0.46			0.72	0.43	1.0*
	Urdu	2.4	0.93	0.88	0.51	0.9		0.67		
	Marathi	1.6	0.92	0.89	0.47			0.65		
E3.	Bengali	-0.8	0.90	0.81			0.5	0.66	0.38	0.38
	Malayalam	-0.4	0.89	0.83	0.32			0.64	0.00	1.0*
	Odia	0.0	0.87	0.71				0.65	0.43	
	Tamil	-0.4	0.86	0.76	0.47			0.59	0.00	
Correlation of measure to WERR:			**0.89**	0.78	0.79	0.77	0.83	0.55	0.48	-0.31

Table 2: Acoustic Token Distribution Similarity (ATDS) measure between Punjabi and donor language predicts downstream speech recognition performance as measured by relative word error rate (WERR) when fine-tuning the wav2vec 2.0 XLSR-128 model adapted using continued pre-training (CPT) on 10 hours of target and 60 hours of donor language speech. Other similarity measures for comparison are derived embeddings of the SpeechBrain language identification model and from the lang2vec database (syntactic, geographic, featural, inventory, genetic, and phonological). * indicate erroneous similarity scores resulting from identical, imputed vectors within the database. Correlations (Pearson's r) calculated using 24 data points (8 donor languages x 3 CPT runs per language with different donor data in each run).

wav2seq (Wu et al., 2023) and Token Distribution Similarity (TDS: Gogoulou et al., 2023). We provide two analyses showing that the ATDS between a target language and its candidate donors precisely predicts downstream ASR performance in the target language resulting from continued pre-training of a speech model on target and donor data.

In our first analysis, we examined how well ATDS can account for the results of the Indic language experiments and how ATDS compares to other measures. Accordingly, for ATDS, we trained the relevant wav2seq models on Punjabi (the target language), induced tokens on Punjabi and all donor languages, then calculated the token frequency vectors and computed the pairwise cosine similarities. Similar to Wu et al. (2021), we also computed similarities using lang2vec and corpus-level means of utterance-level embeddings extracted using a pre-trained spoken language identification model (SpeechBrain: Ravanelli et al., 2021).

Results of this analysis revealed that the two bottom-up acoustic measures provide overall a finer-grained ranking than top-down lang2vec measures. For example, as shown above in Table 2 Feat. (a combination of all lang2vec data sources), the featural similarity, splits the four languages into two groups: similar (0.6: Hindi–Marathi) and dissimilar (0.5: Bengali–Tamil). Additionally, we also found erroneous, perfect similarity values (1.0), resulting from identical language vectors for the category (e.g. Phon.: Marathi-Punjabi), likely an artefact of data imputation. These results demonstrate that while top-down data may be suitable for more noise-tolerant applications (e.g. large-scale typological comparisons), they may not be well suited for helping select donors for a specific under-described target language if the relevant metadata is unavailable or inaccurate.

While both acoustic measures accurately select Hindi as the most suitable donor, ATDS provides a better ranking of the donor languages. As shown in Table 2 (SB), the measure based on SpeechBrain embeddings ranks Gujarati as being as dissimilar to Punjabi as Bengali/Malayalam. We make a similar observation as above that using embeddings from a different pre-trained model than the one to be adapted via CPT risks adding unwanted noise to an inherently hard task. Leveraging the representations of the pre-trained model to be adapted reduces this risk, reflected in ASR improvement being most correlated with ATDS (r = 0.89).

In our second analysis, we examined whether the ATDS measure generalised beyond the Indic languages through identical CPT experiments on a typologically varied set of target languages. We selected triplets of languages consisting of 1) a target language, 2) a language more similar to the target as measured by ATDS, and 3) another language

	Galician (GLG)		Iban (IBA)		Setswana (TSN)	
		WER (WERR)		WER (WERR)		WER (WERR)
E1.	SPA (0.96) 10h GLG + 60h SPA	**13.7 (8.7%)**	ZSM (0.91) 7h IBA + 60h ZSM	**15.9 (4.2%)**	SOT (0.96) 10h TSN + 56h SOT	**11.6 (7.9%)**
E2.	POR (0.89) 10h GLG + 60h POR	13.9 (7.3%)	IND (0.88) 7h IBA + 60h IND	16.4 (1.2%)	NSO (0.88) 10h TSN + 56h NSO	12.0 (4.8%)
B.	10h GLG	15.0	7h IBA	16.6	10h TSN	12.6
U.	-	15.4 (-2.7%)	-	21.4 (-28.9%)	-	20.8 (-65.1%)

Table 3: Validation of the Acoustic Token Distribution Similarity (ATDS) measure for predicting target language automatic speech recognition (ASR) performance as a result of continued pre-training (CPT) of the wav2vec 2.0 XLSR-128 model using mix target and donor language data. For each target language (Galician, Iban, Setswana), U. indicates ASR performance from using the unadapted XLSR-128 model, B. indicates performance from CPT adaptation with only target language data (7–10 hours), and E1–E2 using target language data supplemented with 56–60 hours of donor language data. Parentheses next to donor language names indicate ATDS to the target language. Percentages within parentheses indicate relative word error rate (WERR), relative the baseline word error rate (B) within the same column. Donor language codes are: Spanish (SPA), Portuguese (POR), Malay (ZSM), Indonesian (IND), Sesotho (SOT), Sepedi (NSO).

relatively farther. As summarised below in Table 3, the target languages were Galician (West-Iberian), Iban (Malayic), and Setswana (Sotho–Tswana). For Galician, ATDS predicted that Spanish (SPA: 0.96) was more similar than Portuguese (POR: 0.89); for Iban, Malay (ZSM: 0.91) more than Indonesian (IND: 0.88); and for Setswana, Sesotho (SOT: 0.96) more than Sepedi (NSO: 0.88).

Results of these CPT experiments are summarised below in Table 3. We first note the large difference between Galician and the other languages in the improvement yielded by CPT baselines (B) compared to fine-tuning the unadapted XLSR-128 (U). As we sourced Galician data from CommonVoice (on which XLSR-128 was already pre-trained), CPT yields little further gain (U. 15.4% vs. B. 15.0%). By contrast, ASR performance was much improved via CPT adaptation for both Iban (U. 21.4% vs. B. 16.6%) and Setswana (U. 20.8% vs. B. 12.6%). These results constitute further evidence that directly fine-tuning massively multilingual models can yield sub-optimal performance for under-represented languages and that continued pre-training can help close this performance gap.

We found that the ATDS predictions are borne out for all three target languages (even for Galician in spite of relatively reduced benefits). As shown above in Table 3 for each of the target language columns (Galician, Iban, Setswana), larger improvements in target language ASR performance are observed as a result of continued pre-training on target language data supplemented with data from a more similar language as measured by ATDS than a less similar one (respectively, rows E1. vs. E2). Combined with results for Punjabi above, our findings altogether provide strong evidence for the effectiveness of ATDS for predicting positive transfer between target and donor languages for CPT-based model adaptation.

5 Limitations and future directions

We limited the scope of this paper to exploring the transfer between languages and as such used the standard wav2vec 2.0 pre-training recipe for model adaptation. We acknowledge that this requires a large compute budget beyond what is affordable in many low resource scenarios. In future, we hope to investigate whether or to what extent transfer learning can be combined with more compute-efficient adaptation methods.

To conduct a systematic study of pairwise transfer, we used domain-matched, high-quality ASR datasets containing mostly read speech and only examined sourcing data from a single donor language. Questions relating to multi-donor and multi-domain transfer and how such interactions affect downstream performance in the target language will need to be addressed in future research.

6 Conclusion

For developing automatic speech recognition (ASR) systems for languages with very few resources, we demonstrated that massively multilingual pre-trained models for speech can be successfully adapted via continued pre-training using a mix of data from the target language and supplemental data from a similar, higher-resource 'donor' language. Additionally, we motivate and propose the Acoustic Token Distribution Similarity (ATDS) — a novel measure of similarity to predict positive transfer between a donor and target language. Across a set of typologically different target languages (Punjabi, Galician, Iban, Setswana), we show that the ATDS between the target language and its candidate donors precisely predicts target language ASR performance.

We attribute this predictive capability of ATDS to leveraging the knowledge of the pre-trained model to be adapted, its inductive biases and training objectives, and the distributions within the candidate datasets that will be used to adapt it. It is indeed expected that this measure then is able to predict downstream task improvements better than measures based on other models or external information — the latter of which is not always available or reliable for under-described languages. Given the high cost associated with continued pre-training, however, we argue that using a well-calibrated, task-specific measure minimises the chance of costly, unexpected outcomes.

We make a final observation here that various target-donor language pairs where we observed successful transfer exist in linguistic situations with significant, sustained contact. For example, Galician is a minoritised language in Spain and virtually all Galician speakers are bilingual in Spanish (de la Fuente Iglesias and Pérez Castillejo, 2020). Similarly, Macaire et al. (2022) report that finetuning a model pre-trained on French was particularly successful for Gwadloupéyen and Morisien, two French-based creole languages. These findings suggest that to develop truly inclusive speech technologies in a resource efficient manner, what will be required is a nuanced understanding of what factors linguistic and non-linguistic yield sufficiently high levels of cross-lingual similarity which in turn permit positive transfer.

Acknowledgements

We thank Karol Nowakowski for helpful correspondence and also making available experiment configuration files. We thank Tolúlọpẹ́ Ògúnrẹ̀mí for assistance with using the Stanford NLP cluster as well as Stanford UIT for providing Google Cloud research credits, both without which extensive continued pre-training experiments would not have been possible. We are very grateful for support from both Alexis Michaud at the Centre National de la Recherche Scientifique (CNRS, France) and the Institut des langues rares (ILARA) at École Pratique des Hautes Études. We are thankful for the continued support from Dr. Weisong Shi from University of Delaware.

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.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

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.

Etienne Barnard, Marelie H Davel, Charl van Heerden, Febe De Wet, and Jaco Badenhorst. 2014. The NCHLT speech corpus of the South African languages. Workshop Spoken Language Technologies for Under-resourced Languages (SLTU).

Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2022. Neural representations for modeling variation in speech. *Journal of Phonetics*, 92:101137.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

Martijn Bartelds and Martijn Wieling. 2022. Quantifying language variation acoustically with few resources. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3735–3741, Seattle, United States. Association for Computational Linguistics.

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.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. 2022. Development of automatic speech recognition for the documentation of Cook Islands Māori. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3872–3882, Marseille, France. European Language Resources Association.

Monica de la Fuente Iglesias and Susana Pérez Castillejo. 2020. Phonetic interactions in the bilingual production of Galician and Spanish /e/ and /o/. *International Journal of Bilingualism*, 24(2):305–318.

Mitchell DeHaven and Jayadev Billa. 2022. Improving low-resource speech recognition with pretrained speech models: Continued pretraining vs. semi-supervised training. *arXiv preprint arXiv:2207.00659*.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. 1993. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Web Download.

Evangelia Gogoulou, Timothée Lesort, Magnus Boman, and Joakim Nivre. 2023. A study of continual learning under language shift. *arXiv preprint arXiv:2311.01200*.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for automatic speech recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178.

Séverine Guillaume, Guillaume Wisniewski, and Alexis Michaud. 2023. From 'Snippet-lects' to Doculects and Dialects: Leveraging Neural Representations of Speech for Placing Audio Signals in a Language Landscape . In *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 29–33.

Martin Haspelmath. 2009. The typological database of the world atlas of language structures. *The Use of Databases in Cross-Linguistic Studies*, 41:283.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M. Khapra. 2023. IndicSUPERB: A speech processing universal performance benchmark for Indian languages. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Towards building ASR systems for the next billion users. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Sarah Samson Juan, Laurent Besacier, Benjamin Lecouteux, and Mohamed Dyab. 2015. Using resources from a closely-related language to develop ASR for a very under-resourced language: A case study for Iban. In *Interspeech 2015*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Cécile Macaire, Didier Schwab, Benjamin Lecouteux, and Emmanuel Schang. 2022. Automatic speech recognition and query by example for creole languages documentation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2512–2520, Dublin, Ireland. Association for Computational Linguistics.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. 2023. Adapting multilingual speech representation model for a new, underresourced language through multilingual finetuning and continued pretraining. *Information Processing & Management*, 60(2):103148.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5206–5210.

Georgios Paraskevopoulos, Theodoros Kouzelis, Georgios Rouvalis, Athanasios Katsamanis, Vassilis Katsouros, and Alexandros Potamianos. 2024. Sample-efficient unsupervised domain adaptation of speech recognition systems: A case study for modern Greek. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:286–299.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

Caitlin Richter and Jón Guðnason. 2023. Relative dynamic time warping comparison for pronunciation errors. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Nay San, Martijn Bartelds, Mitchell Browne, Lily Clifford, Fiona Gibson, John Mansfield, David Nash, Jane Simpson, Myfany Turpin, Maria Vollmer, et al. 2021. Leveraging neural representations for facilitating access to untranscribed speech from endangered languages. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Tien-Ping Tan, Xiong Xiao, Enya Kong Tang, Eng Siong Chng, and Haizhou Li. 2009. MASS: A Malay language LVCSR corpus resource. In *2009 Oriental COCOSDA International Conference on Speech Database and Assessments*, pages 25–30. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Felix Wu, Kwangyoun Kim, Shinji Watanabe, Kyu J. Han, Ryan McDonald, Kilian Q. Weinberger, and Yoav Artzi. 2023. Wav2seq: Pre-training speech-to-text encoder-decoder models using pseudo languages. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Peter Wu, Jiatong Shi, Yifan Zhong, Shinji Watanabe, and Alan W Black. 2021. Cross-lingual transfer for speech processing using acoustic language similarity. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1050–1057.

A Materials and methods

A.1 Data

Galician, Spanish, Portuguese, and Indonesian data were sourced from CommonVoice (Ardila et al., 2019); Setswana, Sesotho, and Sepedi from NCHLT (Barnard et al., 2014); Malay from MASS (Tan et al., 2009); and Iban from Juan et al. (2015). For Iban only 7 hours of target data was available and for Sesotho/Sepedi only 56 hours per language of donor data was available. Experiments were otherwise identical to the Indic experiments.

A.2 Continued pre-training

We carried out nearly identical experiments as the single-language experiments in Nowakowski et al. (2023) for Ainu, as we obtained their configuration file for wav2vec 2.0 pre-training using the fairseq library.[7] We made approriate modifications to suit our hardware configuration (4 x A100 40GB GPUs), setting the batch size to 1.5M samples per GPU and gradient accumulation to 16 steps, yielding an effective batch size of 100 minutes.

As in other wav2vec 2.0 multilingual pre-training configurations (Conneau et al., 2021; Babu et al., 2022; Javed et al., 2023), we form multilingual batches (specifically, *bi*-lingual in our case).[8] We set our sampling alpha to 0.0, which results in data being drawn uniformly from the two languages (i.e. target is over-sampled). We make this modification based on the CPT method in Paraskevopoulos et al. (2024), where in- and out-of-domain Greek data were evenly sampled in each batch. In this way, we consider this method akin to "similar-language regularisation" (Neubig and Hu, 2018), as we are more concerned with preventing over-fitting rather than learning about the donor data.

For each CPT run, we start from the official XLSR-128 model checkpoint and update the model for 10k steps. We determined this value based on our pilot runs. We found that 10k updates were sufficient to observe improved downstream ASR performance comparable to previous results (e.g. Nowakowski et al., 2023). This choice permitted us to maximise the number of languages compared in this paper, as each run required on average 15 hours (for 10k steps). For select runs, we also verified

[7] https://github.com/facebookresearch/fairseq

[8] Specifically, we use the implementation adapted from https://github.com/AI4Bharat/IndicWav2Vec/

that no further improvements could be obtained with additional updates (up to 20k).

A.3 ASR fine-tuning and evaluation

For ASR fine-tuning, we follow the official wav2vec 2.0 fine-tuning recipe suitable for 1 hour of transcriptions, modified for our hardware setup. We use a single A6000 48 GB GPU with a batch size of 5.6M samples per GPU and accumulate gradients for 2 steps, yielding an effective batch size of 11.6 minutes. As standard, the feature extractor is kept frozen across all updates and the transformer is frozen for the first 10k of 13k total updates, optimised using a CTC loss. Each fine-tuning run required on average 3.5 hours. For our findings to be applicable for languages with little external text data, we use Viterbi decoding without a language model to obtain transcriptions from the fine-tuned model for evaluation.

A.4 ATDS analyses

For all analyses, we trained the necessary *k*-means and sentencepiece models on a random 5-hour subset of target language data (except for CommonVoice Punjabi, which had in total 4 hours available in the latest 15.0 release). We adopted hyperparameters based on previous findings: extracting embeddings from the mid-point layer (12 of 24) of the XLSR-128 model (e.g. San et al., 2021; Bartelds et al., 2022), and *k*=500 for *k*-means and V=10k for the subword model which were reported as optimal values in Wu et al. (2023). Using a 12GB 3060 GPU, embedding extraction required about 10 minutes per data subset. The *k*-means models trained in about 8 minutes and subword models in less than a minute.

For CommonVoice (CV) Punjabi and Hindi, we conducted similar analyses as those reported by Baevski et al. (2020, Appendix D) for analysing correspondences between the wav2vec 2.0 code vectors and hand-aligned phone labels from TIMIT (Garofolo et al., 1993). In our case we used the labels for CV Punjabi and Hindi via grapheme-to-phoneme conversion, forced-aligned to the audio, and released as Praat TextGrids in the VoxCommunis corpus (Ahn and Chodroff, 2022). We then added tiers containing the wav2seq-induced labels. We hand-inspected several TextGrids for data validation then compiled the correspondences between the wav2seq induced tokens and phoneme labels. We make available all TextGrids as well as the aggregated data.

B Languages

B.1 Indo-Aryan and Dravidian

Punjabi (PAN) is a Northwestern Indo-Aryan language spoken by over 100 million people along the five major tributaries of the Indus river, spanning the state of Punjab in India and the province of Punjab in Pakistan. Hindi (HIN) and Urdu (URD) are mutually-intelligible yet sociolinguistically distinct registers of one Central Indo-Aryan language (usually termed Hindi–Urdu), spoken across the Indian subcontinent and by a majority in the northern part. Gujarati (GUJ) is a Central Indo-Aryan language and Marathi (MAR) is a Southern Indo-Aryan language, spoken in the western Indian states of Gujarat and Maharashtra, respectively. Bengali (BEN), spoken in Bangladesh and the Indian state of West Bengal, and Odia (ORI), spoken in the Indian state of Odisha, are both Eastern Indo-Aryan languages. Finally, Tamil (TAM) and Malayalam (MAL) are both Dravidian languages spoken in the Indian states of Tamil Nadu and Kerala, respectively. The Indo-Aryan and Dravidian language families are phylogenetically unrelated but have a long history of contact and cross-family bilingualism.

B.2 Malayo-Polynesian

Iban (IBA) is a Malayo-Polynesian language spoken by over 2 million people in Brunei as well as the Indonesian and Malaysian parts of the island of Borneo. Iban has some use as a medium of education in the Malaysian state of Sarawak but does not possess official status. It is related to Indonesian (IND) and Malay (ZSM), which are the official languages of Indonesia and Malaysia, respectively.

B.3 Sotho-Tswana

Setswana (TSN) is a Bantu language spoken by over 8 million people Botswana, South Africa, and Zimbabwe, where it is an official language. Setswana also possesses minority language status in Namibia. Two other languages of the Sotho-Tswana subgroup of Bantu are Sesotho (SOT, also known as “Southern Sotho”) and Sepedi (NSO, “Northern Sotho”). Sesotho is an official language of South Africa, Lesotho, and Zimbabwe, and Sepedi is an official language of South Africa.

B.4 West Iberian

Galician (GLG) is a Romance language spoken in Galicia, an administrative division of northwestern

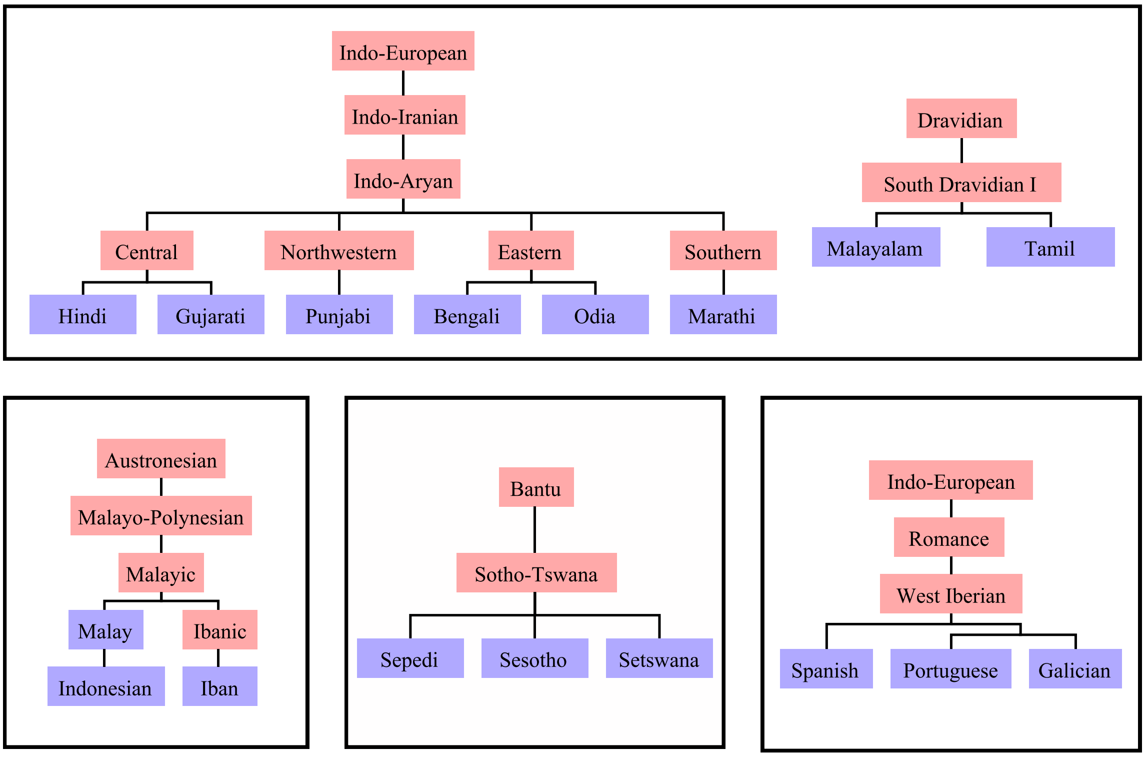


Figure 4: Family trees of the languages studied in this paper. Language families are in red and languages are in blue.

Spain bordering Portugal where it is the official language and spoken by over 2 million people. It is closely related to Spanish (SPA) and Portuguese (POR), and all three are classified under the West Iberian subgroup of Romance languages. Galician and Portuguese split in the late Middle Ages (c. 15th century) and thus are the most closely related pair of the three. Sociolinguistically, Galician speakers use Spanish in literary contexts and thus the two languages have a diglossic relationship.