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

Nay San, Georgios Parakevopoulos, Aryaman Arora, Xiluo He,
Prabhjot Kaur, Oliver Adams, Dan Jurafsky

Stanford | Linguistics
SCHOOL OF HUMANITIES AND SCIENCES

AOHNA

WAYNE STATE UNIVERSITY

Atos zData

Low-resource ASR

- Transcribed speech is a key resource of ASR training
- ‘Low-resource’ (for ASR):
  - Limitation is *only* transcribed speech
  - Easy to source untranscribed speech and metadata about language
- Problems:
  1. Limited metadata (if under-described)
  2. Limited recordings (for self-supervised training)
- Proposal:
  - Use recordings from another language (helps with #2)
  - Use bottom-up approach (helps with #1)

Missing metadata

Family	Language	Information (Database)	
		Inventories (PHOIBLE)	Phonology (WALS)
Indo-Aryan	Punjabi	✓	✗
Sotho-Tswana	Setswana	✗	✗

Feature vectors in lang2vec are imputed (via k-NN)

Combating a curse
via continued pre-training

- Default go-to: fine-tune a pre-trained model
- Problem: ‘Curse of Multilinguality’
  - Under-representation in massively multilingual models
    - Worse downstream performance on under-represented languages
  - wav2vec 2.0 XLSR-128:
    - Pre-trained on 436k hours from 128 languages
    - 95% of data is Germanic/Romance
- Solution: Continued Pre-training (CPT) on target language
  - Ainu (200h: Nowakowski et al., 2023)
  - Greek (70h: Paraskevopoulos et al., 2024)
- Problem: what if we don’t even have 70-200h?
  - Can we add data from another language?

Experiment 1

Condition	Test set WER (WERR)		Data for continued pre-training
	Median	Range	
T. In-domain top-line	22.2 (11.2%)	-	70h Punjabi
B. Only target data baseline	25.0	-	10h Punjabi
U. Unadapted XLSR-128	30.8 (-23.2%)	-	-

All models fine-tuned for ASR using 1h of transcribed Punjabi

Experiment 1

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
B. Only target data baseline	25.0	-	10h Punjabi
U. Unadapted XLSR-128	30.8 (-23.2%)	-	-

All models fine-tuned for ASR using 1h of transcribed Punjabi

Experiment 1

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
U. Unadapted XLSR-128	30.8 (-23.2%)	-	-

All models fine-tuned for ASR using 1h of transcribed Punjabi

Experiment 1

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%)	-	-

All models fine-tuned for ASR using 1h of transcribed Punjabi

Experiment 1

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%)	-	-

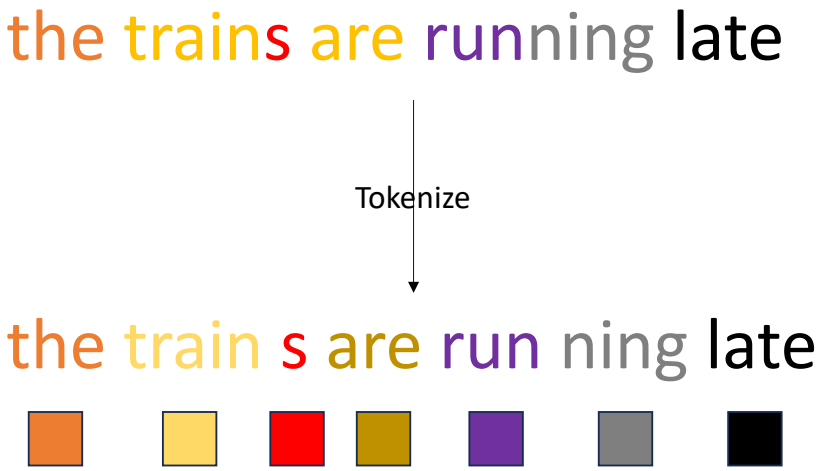
What similarity measure best predicts word error rate reduction (WERR)?

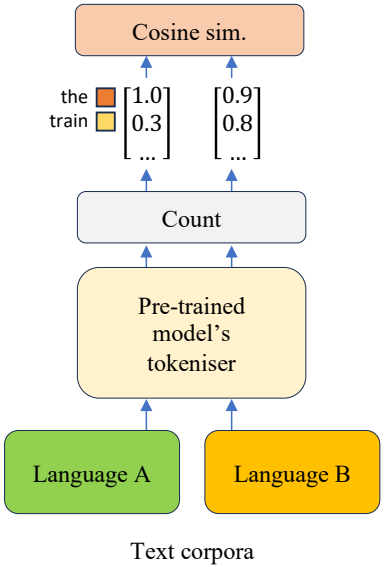
All models fine-tuned for ASR using 1h of transcribed Punjabi

Predicting positive transfer

Text domain

- Token Distribution Similarity can help predict positive transfer in the text domain (Gogoulou et al., 2023)





Text: top-down coarse-to-finer grained

Raw data

the trains are running late

the train s are run ning late

Text: top-down coarse-to-finer grained

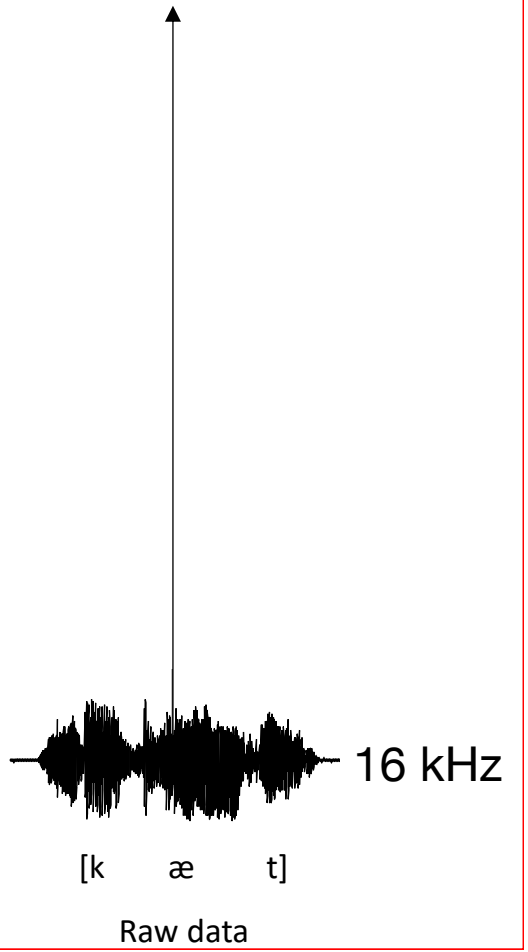
Raw data

the trains are running late

Tokenize

the train s are run ning late

Speech: bottom up fine-to-coarser grained



Text: top-down coarse-to-finer grained

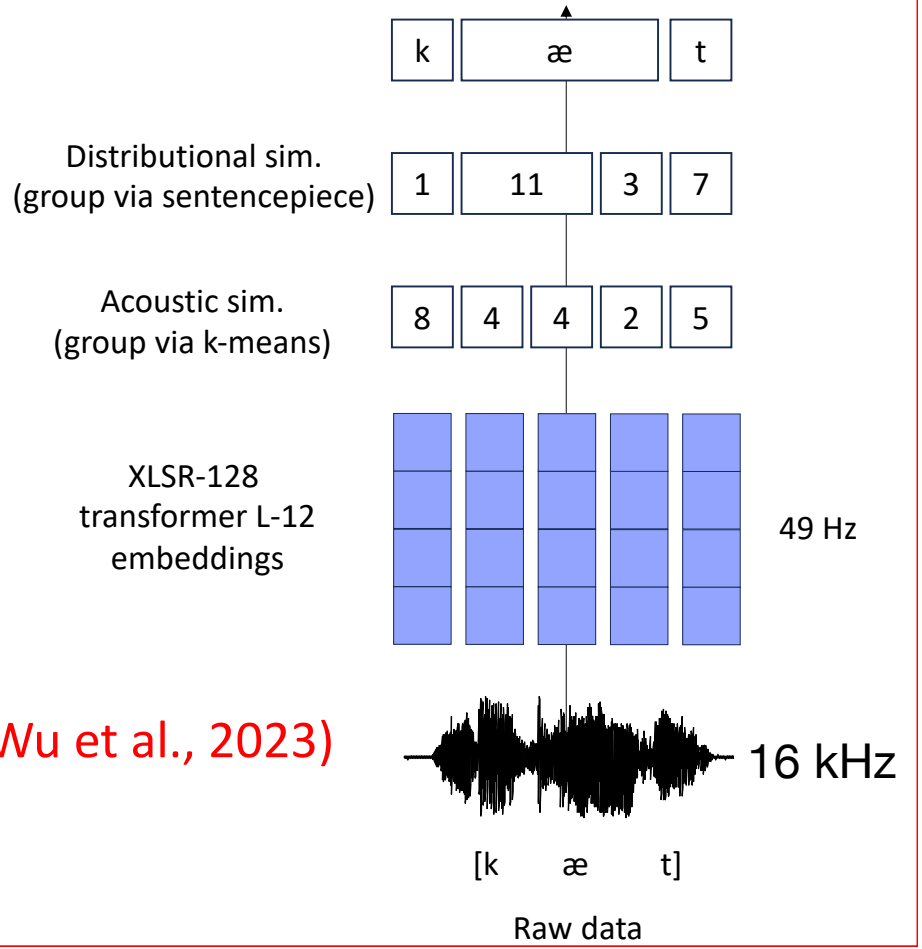
Raw data

the trains are running late

Tokenize

the train s are run ning late

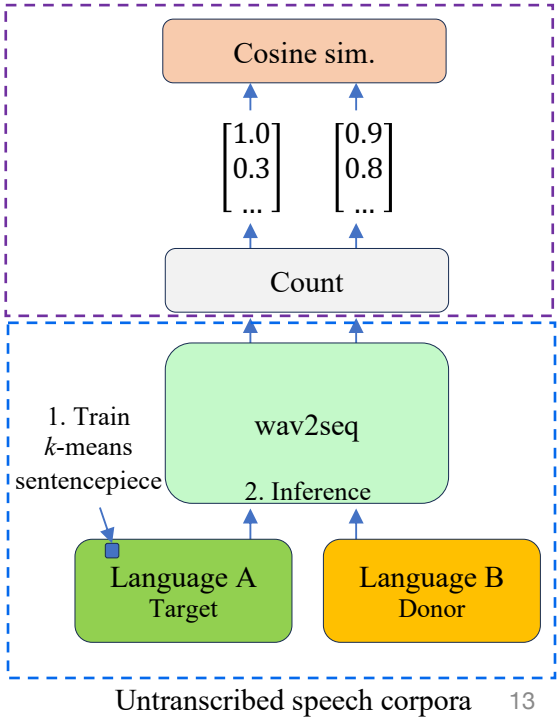
Speech: bottom up fine-to-coarser grained



wav2seq (Wu et al., 2023)

Acoustic Token Distribution Similarity (ATDS)

- **wav2seq**: Derive (pseudo-)tokens from untranscribed speech
- **TDS**: Predict positive transfer based on (text) token distributions
- **ATDS** (wav2seq + TDS):
  - Predict positive transfer using untranscribed speech corpora based on (pseudo-)token distributions



	Punjabi (PAN)		
			Similarity Measure
	Donor Lang.	Median WERR (of 3 runs)	ATDS
E1.	Hindi	6.0	0.96
E2.	Gujarati	2.4	0.93
	Urdu	2.4	0.93
	Marathi	1.6	0.92
E3.	Bengali	-0.8	0.90
	Malayalam	-0.4	0.89
	Odia	0.0	0.87
	Tamil	-0.4	0.86
Correlation of measure to WERR:			**0.89**

	Punjabi (PAN)				
			Similarity Measure		
	Donor Lang.	Median WERR (of 3 runs)	ATDS	lang2vec	
				Feat.	
E1.	Hindi	6.0	0.96	0.6	
E2.	Gujarati	2.4	0.93		
	Urdu	2.4	0.93		
	Marathi	1.6	0.92		
E3.	Bengali	-0.8	0.90	0.5	
	Malayalam	-0.4	0.89		
	Odia	0.0	0.87		
	Tamil	-0.4	0.86		
Correlation of measure to WERR:			**0.89**	0.83	

Bottom-up measure (ATDS) is more fine-grained than top-down (lang2vec)

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

Erroneous similarities from missing/imputed features in databases

	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

Acoustic measures based on model embeddings

	Punjabi (PAN)									
					Similarity Measure					
	Donor Lang.	Median WERR (of 3 runs)	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

XLSR-128 (Model used for CPT)

SpeechBrain LangID (External model)

Further validation of ATDS

For each target language (e.g. GLG), ATDS predicts best donor from two candidates (e.g. SPA, POR)

	Galician (GLG)		Iban (IBA)		Setswana (TSN)	
E1.	SPA (0.96) 10h GLG + 60h SPA	WER (WERR) **13.7 (8.7%)**	ZSM (0.91) 7h IBA + 60h ZSM	WER (WERR) **15.9 (4.2%)**	SOT (0.96) 10h TSN + 56h SOT	WER (WERR) **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%)

Summary

- Continued pre-training (CPT) alleviates under-representation
  - Evaluated on 4 target languages:
    - Punjabi (Indo-Aryan)
    - Galician (West Iberian)
    - Iban (Malayo-Polynesian)
    - Setswana (Sotho-Tswana)
  - More data is better for CPT
- Can source data from ‘donor’ language
- ATDS can help pick best donor
  - Better than 7 other measures