Multilingual Automatic Speech Recognition

601.764

4/11/2023

DOI: 10.1121/1.1906946 • Corpus ID: 121505424

Automatic Recognition of Spoken Digits

K. Davis, R. Biddulph, S. Balashek • Published 1 November 1952 • Physics • Journal of the Acoustical Society of America

The recognizer discussed will automatically recognize telephone-quality digits spoken at normal speech rates by a single individual, with an accuracy varying between 97 and 99 percent. After some preliminary analysis of the speech of any individual, the circuit can be adjusted to deliver a similar accuracy on the speech of that individual. The circuit is not, however, in its present configuration, capable of performing equally well on the speech of a series of talkers without recourse to such adjustment.Circuitry involves division of the speech spectrum into two frequency bands, one below and the other above 900 cps. Axis-crossing counts are then individually made of both band energies to determine the frequency of the maximum syllabic rate energy with each band. Simultaneous two-dimensional frequency portrayal is found to possess recognition significance. Standards are then determined, one for each digit of the ten-digit series, and are built into the recognizer as a form of elemental memory. By means of... Collapse

In the early 1960s, IBM developed and demonstrated "Shoebox" -- a forerunner of today's voice recognition systems.

Click to enlarge

Dr. E. A. Quade, manager of the advanced technology group in IBM's Advanced Systems Development Laboratory in San Jose, Calif., demonstrates Shoebox, an experimental machine that performed arithmetic on voice command.

This innovative device recognized and responded to 16 spoken words, including the ten digits from "0" through "9." When a number and command words such as "plus," "minus" and "total" were spoken, Shoebox instructed an adding machine to calculate and print answers to simple arithmetic problems. Shoebox was operated by speaking into a microphone, which converted voice sounds into electrical impulses. A measuring circuit classified these impulses according to various types of sounds and activated the attached adding machine through a relay system.

Shoebox was developed by William C. Dersch at IBM's Advanced Systems Development Division Laboratory in San Jose, Calif. He later demonstrated it to the public on television and at the IBM Pavilion of the 1962 World's Fair in Seattle.

Click to enlarge

IBM engineer William C. Dersch, shown above in 1961, demonstrates Shoebox, an experimental machine that performed arithmetic on voice command.

Fallout from these blasts

The first idea: **Try Artificial Intelligence . . .**

DARPA Speech Understanding Research Project (1972-75)

Used classical AI to try to "understand what is being said with something of the facility of a native speaker"

DARPA SUR was viewed as a failure; funding was cut off after three years

The second idea: **Give Up.**

1975-1986: No U.S. research funding for MT or ASR

https://www.ling.upenn.edu/courses/ling052/LibermanNAS072012a.pdf

- 1,000 words
- Human 3 year old

The HARPY Speech Recognition System

Thesis Summary

Bruce T. Lowerre

April, 1976

Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Submitted to Carnegie-Mellon University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Siri

2011

2014

2012

alexa

2014

Deep Speech 2: End-to-End Speech Recognition in English and Mandarin

2015

Baidu Research – Silicon Valley AI Lab*

Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, Zhenyao Zhu

Deep Speech 2

- English & Mandarin
- 10,000+ hours
- “Our English and Mandarin corpora are substantially larger than those commonly reported in speech recognition literature.”

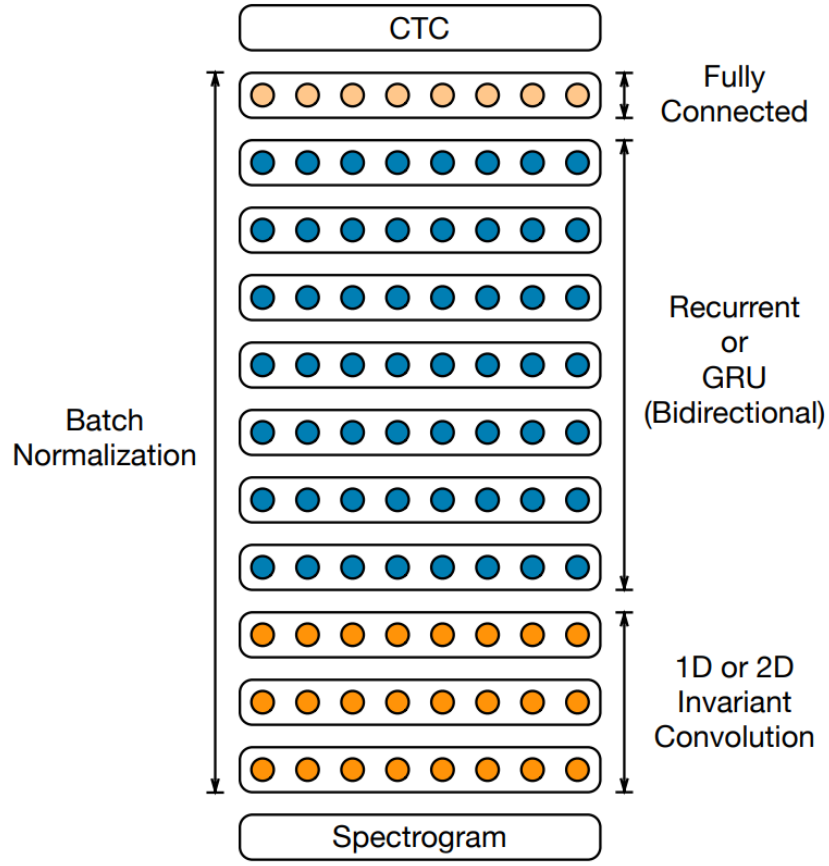


Figure 1: Architecture of the DS2 system used to train on both English and Mandarin speech. We explore variants of this architecture by varying the number of convolutional layers from 1 to 3 and the number of recurrent or GRU layers from 1 to 7.

Deep Speech 2

Dataset	Speech Type	Hours
WSJ	read	80
Switchboard	conversational	300
Fisher	conversational	2000
LibriSpeech	read	960
Baidu	read	5000
Baidu	mixed	3600
Total		11940

Table 9: Summary of the datasets used to train DS2 in English. The Wall Street Journal (WSJ), Switchboard and Fisher [13] corpora are all published by the Linguistic Data Consortium. The LibriSpeech dataset [46] is available free on-line. The other datasets are internal Baidu corpora.

Deep Speech 2

"We find that our best Mandarin Chinese speech system transcribes short voice-query like utterances better than a typical Mandarin Chinese speaker. To benchmark against humans we ran a test with 100 randomly selected utterances and had a group of 5 humans label all of them together. The group of humans had an error rate of 4.0% as compared to the speech systems performance of 3.7%"

Architecture	Dev	Test
5-layer, 1 RNN	7.13	15.41
5-layer, 3 RNN	6.49	11.85
5-layer, 3 RNN + BatchNorm	6.22	9.39
9-layer, 7 RNN + BatchNorm + 2D conv	5.81	7.93

Table 16: Comparison of the improvements in DeepSpeech with architectural improvements. The development and test sets are Baidu internal corpora. All the models in the table have about 80 million parameters each

SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network

William Chan, Daniel S. Park, Chris A. Lee, Yu Zhang, Quoc V. Le, Mohammad Norouzi

2021

Google Research, Brain Team

{williamchan,danielspark,chrisalee,ngyuzh,qvl,mnorouzi}@google.com

- 7 pre-existing datasets
- 5,140 hours
- English

1. AMI [31]. AMI is approximately 100 hours of meeting recordings.
2. Common Voice [32]. Common Voice is a crowd-sourced open licensed speech dataset. We use the version 5.1 (June 22 2020) snapshot with approximately 1500 hours. The data was collected at 48 KHz, and we resampled it to 16 KHz.
3. English Broadcast News (LDC97S44, LDC97T22, LDC98S71, LDC98T28). English Broadcast News is approximately 50 hours of television news.
4. LibriSpeech [33]. LibriSpeech is approximately 960 hours of speech from audiobooks.
5. Switchboard/Fisher (LDC2004T19, LDC2005T19, LDC2004S13, LDC2005S13, LDC97S62). Switchboard/Fisher is approximately 2000 hours of telephone conversations. The data was collected at 8 KHz, and we upsampled it to 16 KHz.
6. TED-LIUM v3 [34, 35]. TED-LIUM is approximately 450 hours of TED talks.
7. Wall Street Journal (LDC93S6B, LDC94S13B). WSJ is approximately 80 hours of clean speech.

Speech Stew

BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition

2021

Yu Zhang*, Daniel S. Park*, Wei Han*,
James Qin, Anmol Gulati, Joel Shor, Aren Jansen,
Yuanzhong Xu, Yanping Huang, Shibo Wang, Zongwei Zhou,
Bo Li, Min Ma, William Chan, Jiahui Yu, Yongqiang Wang,
Liangliang Cao, Khe Chai Sim, Bhuvana Ramabhadran,
Tara N. Sainath, Françoise Beaufays, Zhifeng Chen, Quoc V. Le, Chung-Cheng Chiu,
Ruoming Pang† and Yonghui Wu

BigSSL

- YouTube Data
- YT-L: 350k hours of segmented, weakly-labeled audio, combined with 1000 hours of labeled audio
- YT-T: 500k hours of segmented, pseudo-labeled audio
- YT-U: 900k hours of segmented, unlabeled audio

BigSSL

TABLE VII: YouTube and Voice Search datasets.

Language	YouTube (hrs)	Voice Search (hrs)
Hungarian (HU)	400k	9k
Chinese (TW)	900k	20k
Hindi (IN)	800k	27k

BigSSL

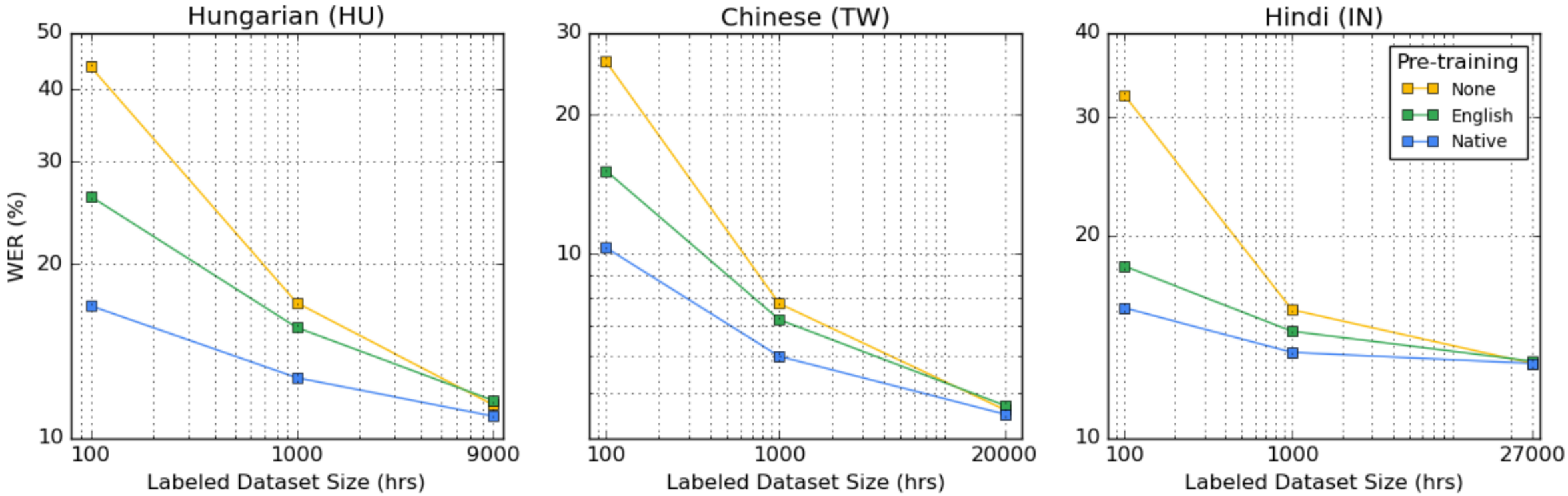


Fig. 6: Test WERs (%) from training pre-trained Conformer XL RNN-T networks on non-English Search datasets and subsets thereof. Both axes are plotted in log-scale.

MLS: A LARGE-SCALE MULTILINGUAL DATASET FOR SPEECH RESEARCH

A PREPRINT

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, Ronan Collobert
Facebook AI Research, Menlo Park & Paris, USA & France
{vineelkpratap,qiantong,anuroops,gab,locronan}@fb.com

December 22, 2020

MLS

- 8 Languages
- 44.5K Hours English
- 6K Hours other Langs
- “Based on the number of audiobook hours and the availability of the corresponding text sources of the audiobooks, we have selected English, German, Dutch, Spanish, French, Portuguese, Italian, Polish for the MLS dataset preparation.”

Table 2: Statistics of Train/Dev/Test partitions of each language. Below lists for each partition: the total duration in hours (left), number of speakers in each gender (middle) and duration per gender in dev and test sets (right).

Language	**Duration (hrs)**			**# Speakers**						**# Hours / Gender**			
	train	dev	test	train		dev		test		dev		test	
				M	F	M	F	M	F	M	F	M	F
English	44,659.74	15.75	15.55	2742	2748	21	21	21	21	7.76	7.99	7.62	7.93
German	1,966.51	14.28	14.29	81	95	15	15	15	15	7.06	7.22	7.00	7.29
Dutch	1,554.24	12.76	12.76	9	31	3	3	3	3	6.44	6.32	6.72	6.04
French	1,076.58	10.07	10.07	62	80	9	9	9	9	5.13	4.94	5.04	5.02
Spanish	917.68	9.99	10	36	50	10	10	10	10	4.91	5.08	4.78	5.23
Italian	247.38	5.18	5.27	22	43	5	5	5	5	2.5	2.68	2.38	2.90
Portuguese	160.96	3.64	3.74	26	16	5	5	5	5	1.84	1.81	1.83	1.90
Polish	103.65	2.08	2.14	6	5	2	2	2	2	1.12	0.95	1.09	1.05

Table 1: LibriVox audiobooks statistics for the top 15 languages (* - audiobooks with mix of multiple languages)

Language	Hours	Books	Speakers
English	71,506.79	12421	4214
German	3,287.48	593	244
Dutch	2,253.68	206	91
Spanish	1,438.41	285	120
French	1,333.35	224	114
Multilingual*	516.82	130	19
Portuguese	284.59	68	31
Italian	279.43	61	28
Russian	172.34	44	29
Latin	138.93	20	16
Polish	137.00	25	16
Church Slavonic	136.42	8	2
Hebrew	125.72	23	13
Japanese	97.67	38	24
Ancient Greek	69.77	43	8

- 24 Languages
- Mostly from conversational telephone speech.
- 25 - 65 hours per language
- Data not released
- (MLS summary of this)

2019

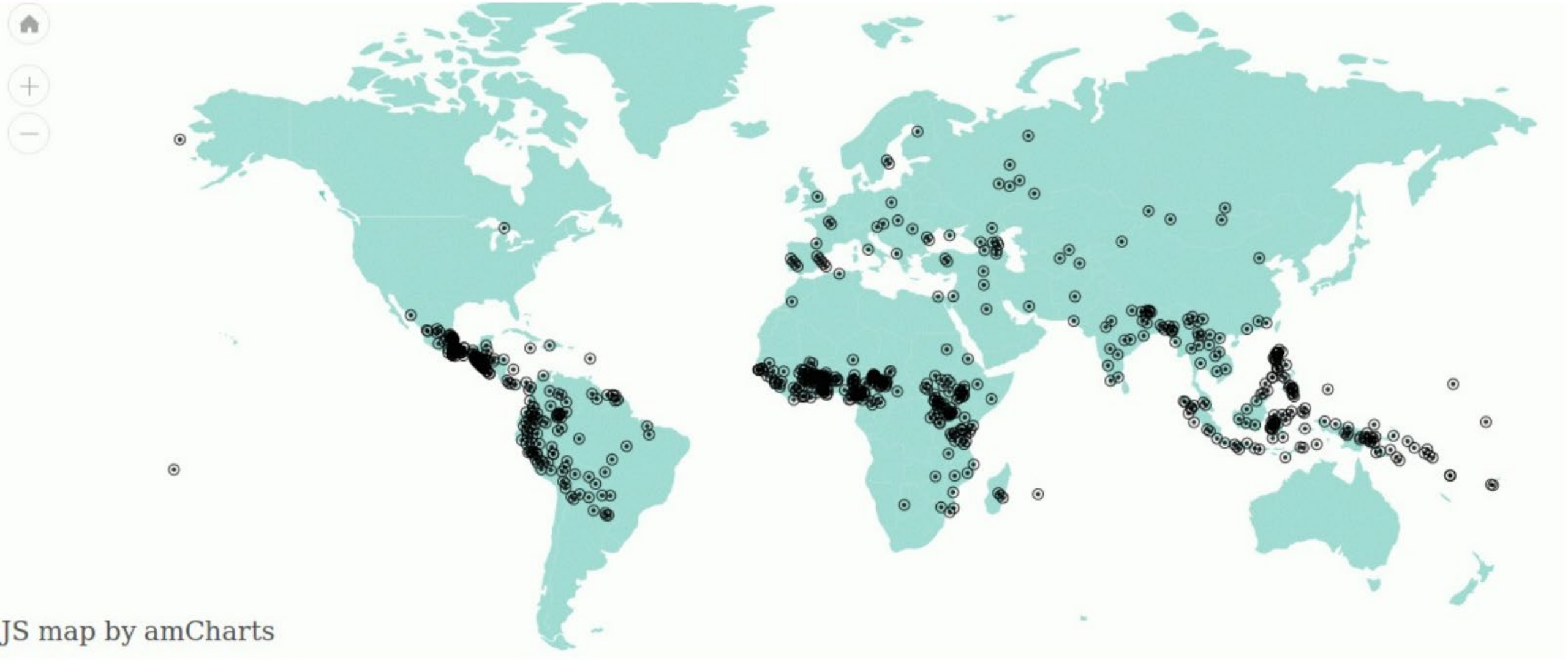
CMU WILDERNESS MULTILINGUAL SPEECH DA

Alan W Black

Language Technologies Institute, Carnegie Mellon Univ
Pittsburgh, PA, USA. awb@cs.cmu.edu

CMU Wilderness

- 700 Languages!
- Audio, Aligned Text, and Word Pronunciations
- “On average each language provides around 20 hours of sentence lengthed transcriptions”
- Read New Testaments available from the www.bible.is



699 Languages Successfully Aligned

The M-AILABS Speech Dataset

9 Langs,
1,000 Hours

Posted on 3.01.2019 by Imdat Solak

The following is the text that accompanied the M-AILABS Speech DataSet:

The M-AILABS Speech Dataset is the first large dataset that we are providing free-of-charge, freely usable as training data for **speech recognition** and **speech synthesis**.

Most of the data is based on LibriVox and Project Gutenberg. The training data consist of nearly thousand hours of audio and the text-files in prepared format.

A transcription is provided for each clip. Clips vary in length from 1 to 20 seconds and have a total length of approximately shown in the list (and in the respective `info.txt` -files) below.

The texts were published between 1884 and 1964, and are in the public domain. The audio was recorded by the LibriVox project and is also in the public domain – **except for Ukrainian**.

2020

Common Voice: A Massively-Multilingual Speech Corpus

Rosana Ardila,[†] Megan Branson,[†] Kelly Davis,[†] Michael Henretty, Michael Kohler, Josh Meyer,[°]
Reuben Morais,[†] Lindsay Saunders,[†] Francis M. Tyers,[‡] Gregor Weber[†]

[†] Mozilla [‡] Indiana University [°] Artie, Inc.
Various Cities Bloomington, IN, USA Los Angeles, CA, USA
{rosana, mbranson, kdavis, reuben, lsaunders, gweber}@mozilla.com,
ftyers@iu.edu, michael.henretty@gmail.com, me@michaelkohler.info, josh.meyer@artie.com

- 38 Languages
- 50,000 people
- 2,500 hours of audio
- "To our knowledge this is the largest audio corpus in the public domain for speech recognition, both in terms of number of hours and number of languages"

CommonVoice

“By applying transfer learning from a source English model, we find an average Character Error Rate improvement of 5.99 ± 5.48 for twelve target languages (German, French, Italian, Turkish, Catalan, Slovenian, Welsh, Irish, Breton, Tatar, Chuvash, and Kabyle). For most of these languages, these are the first ever published results on end-to-end Automatic Speech Recognition”

CommonVoice

- Website or iPhone App
- Read Sentence
- Mozilla

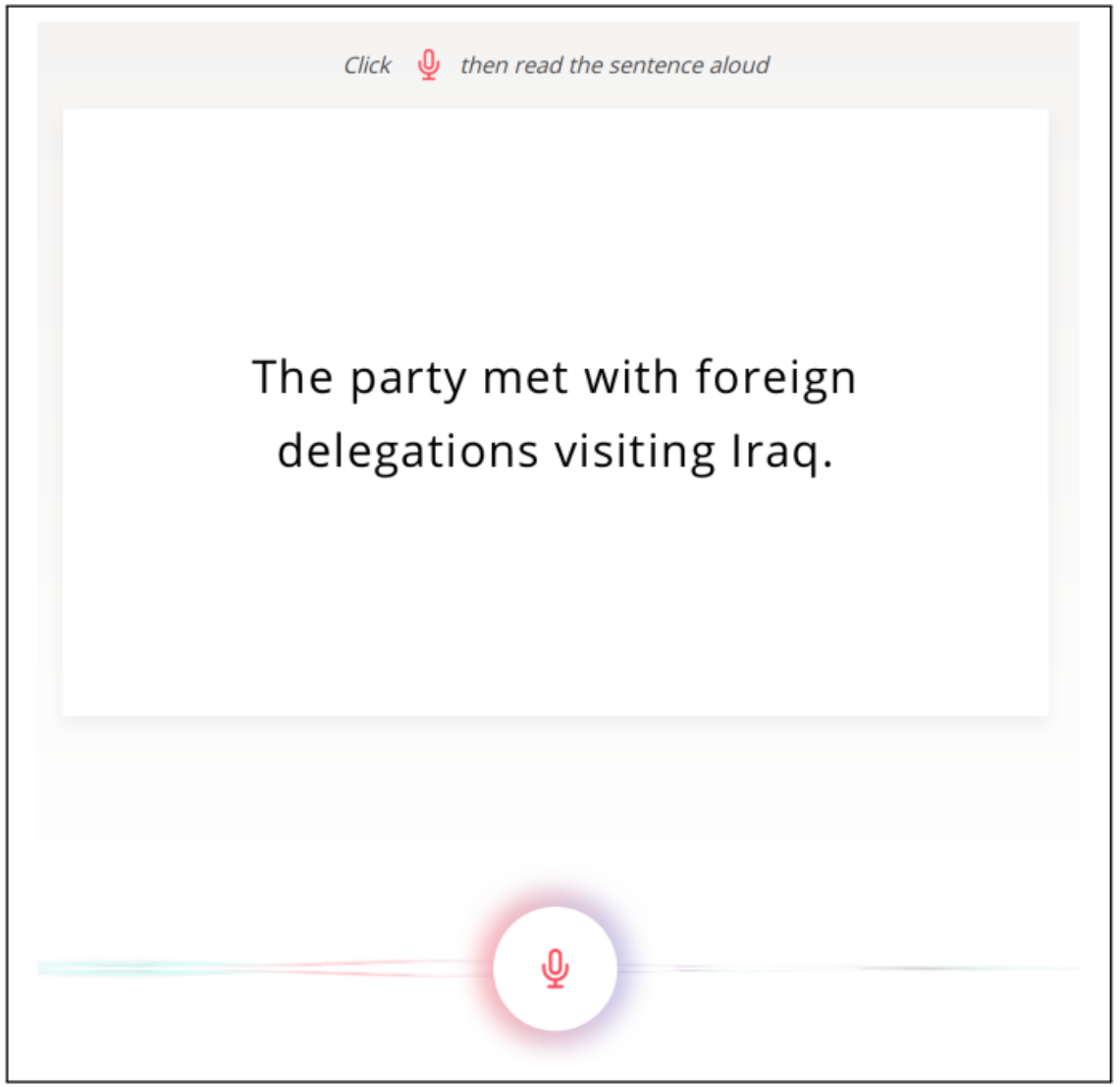


Figure 1: Recording interface for Common Voice. Additionally, it is possible to skip or report as problematic any audio or sentence.

Language	Code	Voices	Hours	
			Total	Validated
Abkhaz	ab	3	<1	<1
Arabic	ar	225	15	9
Basque	eu	508	83	46
Breton	br	118	10	3
Catalan	ca	1,834	120	107
Chinese (China)	zh-ZH	288	12	11
Chinese (Taiwan)	zh-TW	949	43	33
Chuvash	cv	38	2	1
Dhivehi	dv	92	8	5
Dutch	nl	502	23	18
English	en	39,577	1,087	780
Esperanto	eo	129	16	13
Estonian	et	225	12	11
French	fr	3,005	184	173
German	de	5,007	340	325
Hakha Chin	cnh	280	4	2
Indonesian	id	54	5	4
Interlingua	ia	11	2	1
Irish	ga	63	3	2
Italian	it	602	40	36
Japanese	ja	48	2	1
Kabyle	kab	584	192	181
Kinyarwanda	rw	32	1	<1
Kyrgyz	ky	97	20	8
Latvian	lv	82	8	6
Mongolian	mn	230	9	8
Persian	fa	1,240	70	67
Portuguese	pr	316	30	27
Russian	ru	64	31	27
Sakha	sah	35	6	3
Slovenian	sl	42	5	2
Spanish	es	611	31	27
Swedish	sv	44	3	3
Tamil	ta	89	5	3
Tatar	tt	132	26	22
Turkish	tr	344	10	9
Votic	vot	2	<1	<1
Welsh	cy	748	48	42
TOTAL		58,250	2,508	2,019

Table 1: Current data statistics for Common Voice. Data in *italics* is as of yet unreleased. Other numbers refer to the data published in the June 12, 2019 release.

		Dataset Size					
		Audio Clips			Unique Speakers		
Language	**Code**	Dev	Test	Train	Dev	Test	Train
Slovenian	sl	110	213	728	1	12	3
Irish	ga	181	138	1,001	4	12	6
Chuvash	cv	96	77	1,023	4	12	5
Breton	br	163	170	1,079	3	15	7
Turkish	tr	407	374	3,771	32	89	32
Italian	it	627	734	5,019	29	136	37
Welsh	cy	1,235	1,201	9,547	51	153	75
Tatar	tt	1,811	1,164	11,187	9	64	3
Catalan	ca	5,460	5,037	38,995	286	777	313
French	fr	5,083	4,835	40,907	237	837	249
Kabyle	kab	5,452	4,643	43,223	31	169	63
German	de	7,982	7,897	65,745	247	1,029	318

Table 2: Data used in the experiments, from an earlier multilingual version of Common Voice. Number of audio clips and unique speakers.

	Character Error Rate					
	Number of Layers Copied from English					
Lang.	None	1	2	3	4	5
sl	23.35	21.65	26.44	19.09	**15.35**	17.96
ga	31.83	31.01	32.2	27.5	25.42	**24.98**
cv	48.1	47.1	44.58	42.75	**27.21**	31.94
br	21.47	19.16	20.01	18.06	**15.99**	18.42
tr	34.66	34.12	34.83	31.79	**27.55**	29.74
it	40.91	42.65	42.82	36.89	**33.63**	35.10
cy	34.15	31.91	33.63	30.13	**28.75**	30.38
tt	32.61	31.43	30.80	27.79	**26.42**	28.63
ca	38.01	35.21	39.02	35.26	**33.83**	36.41
fr	43.33	43.26	43.51	43.24	43.20	**43.19**
kab	25.76	25.5	26.83	25.25	**24.92**	25.28
de	43.76	43.69	43.62	**43.60**	43.76	43.69

Table 3: Fine-Tuned Transfer Learning Character Error Rate for each language, in addition to a baseline trained from scratch on the target language data. Bolded values display best model per language. Shading indicates relative performance per language, with darker indicating better models.

XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale

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[□‡], Michael Auli[△‡]

[△] Meta AI [□] Google AI [◊] Outreach [♣] Hugging Face

2021

- 2B parameters
- ~500k hours of publicly available speech audio
- 128 languages, an order of magnitude more public data than the largest known prior work

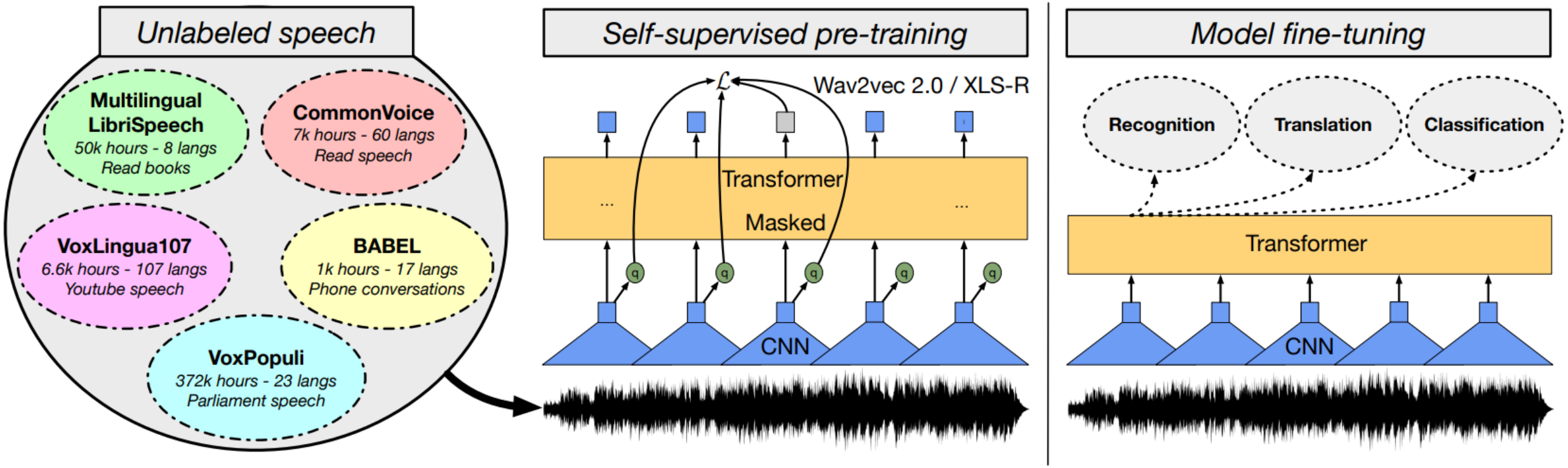


Figure 1: **Self-supervised cross-lingual representation learning.** We pre-train a large multilingual wav2vec 2.0 Transformer (XLS-R) on 436K hours of unannotated speech data in 128 languages. The training data is from different public speech corpora and we fine-tune the resulting model for several multilingual speech tasks.

mSLAM: Massively multilingual joint pre-training for speech and text

Ankur Bapna [*1] **Colin Cherry** [*1] **Yu Zhang** [*1] **Ye Jia** [1] **Melvin Johnson** [1] **Yong Cheng** [1] **Simran Khanuja** [1]
Jason Riesa [1] **Alexis Conneau** [1]

2022

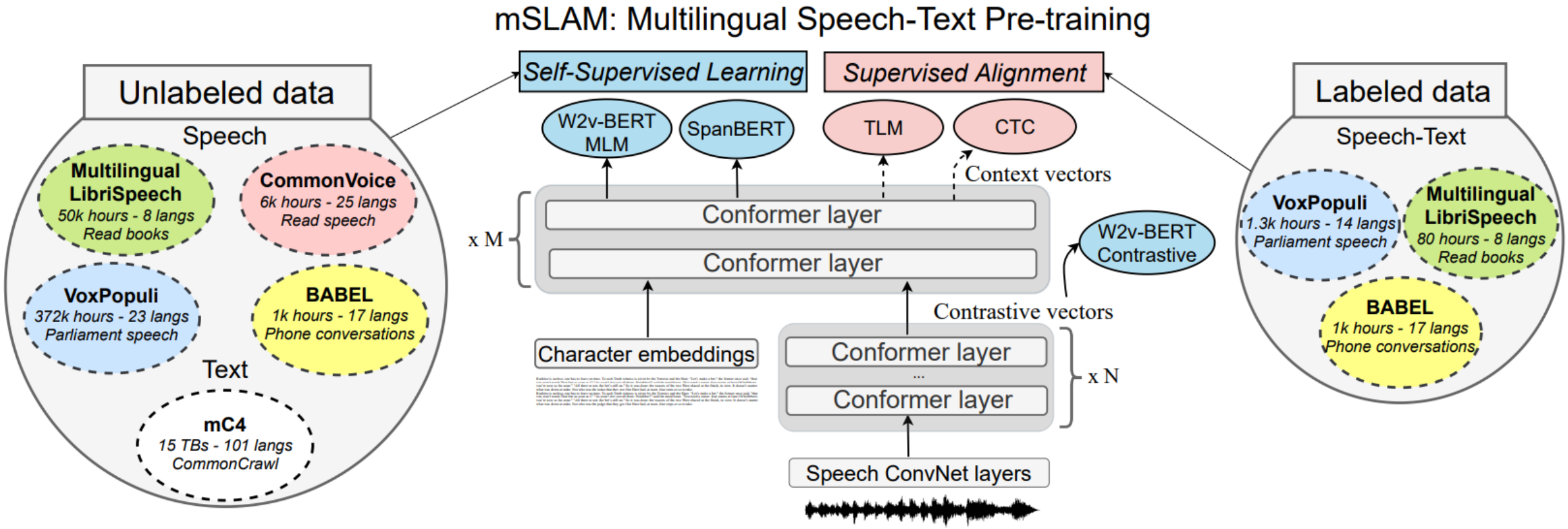


Figure 1: Multilingual Speech-Text Pretraining We pre-train a large multilingual speech-text Conformer on 429K hours of unannotated speech data in 51 languages, 15TBs of unannotated text data in 101 languages, as well as 2.3k hours of speech-text ASR data.

"Our unlabeled speech data closely follows the pre-training data used for XLS-R (Babu et al., 2021) with one major difference: we do not use VoxLingua. As a consequence our model is pre-trained on speech from 51 languages as compared to 128 for XLS-R, and our pretraining set is smaller by 6.6k hours. "

Robust Speech Recognition via Large-Scale Weak Supervision

Alec Radford [*1] **Jong Wook Kim** [*1] **Tao Xu** [1] **Greg Brockman** [1] **Christine McLeavey** [1] **Ilya Sutskever** [1]

- 680,000 hours
- … of what?
- 117,000 hours 96 Non-English Langs
- 125,000 X → English Translation

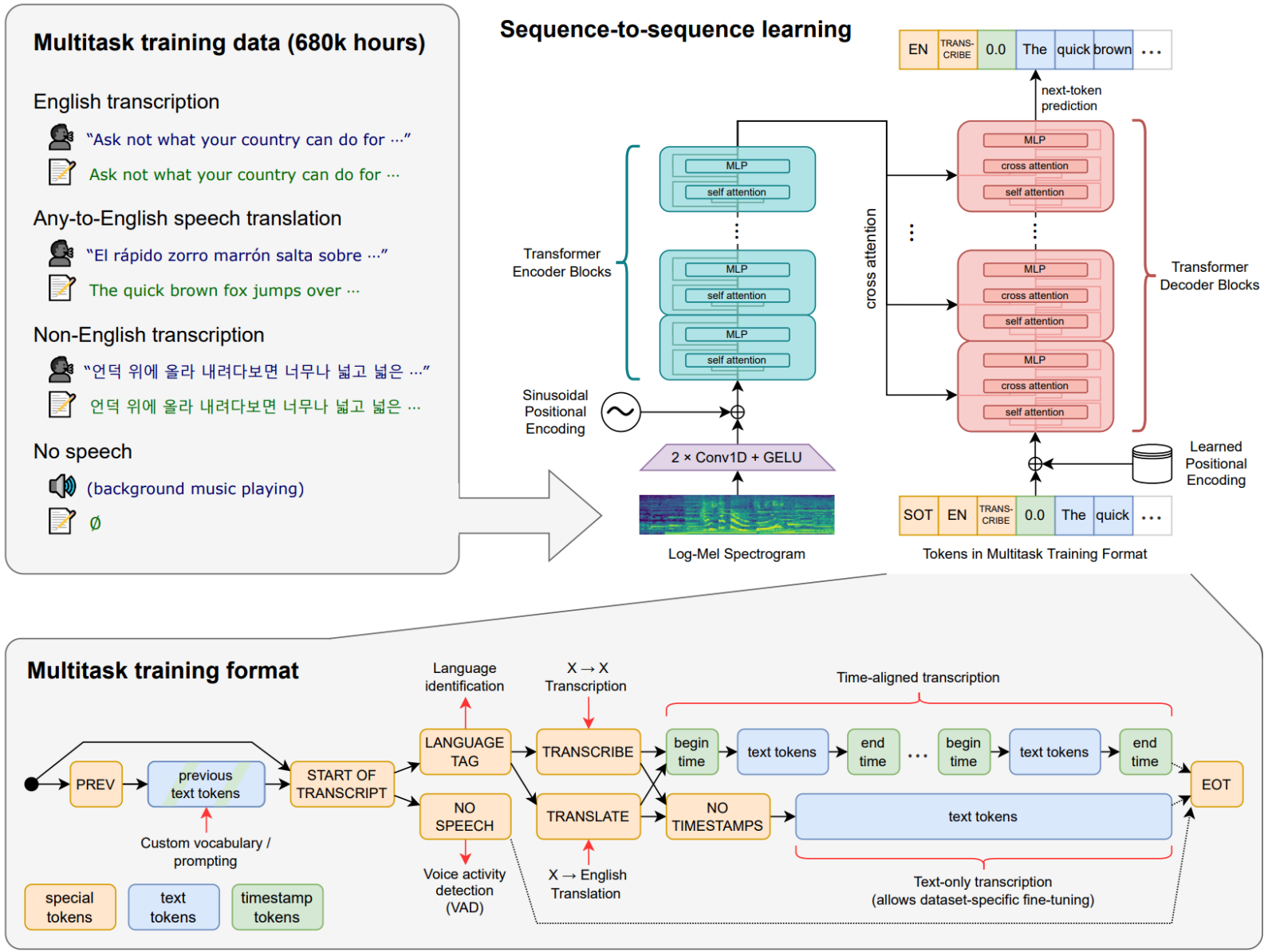


Figure 1. **Overview of our approach.** A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets, as further explained in Section 2.3.

Model	MLS	VoxPopuli
VP-10K + FT	-	15.3
XLS-R (1B)	10.9	10.6
mSLAM-CTC (2B)	9.7	9.1
Maestro	-	**8.1**
Zero-Shot Whisper	**7.3**	13.6

Table 3. **Multilingual speech recognition performance.** Zero-shot Whisper improves performance on Multilingual LibriSpeech (MLS) but is still significantly behind both Maestro, XLS-R, and mSLAM on VoxPopuli.

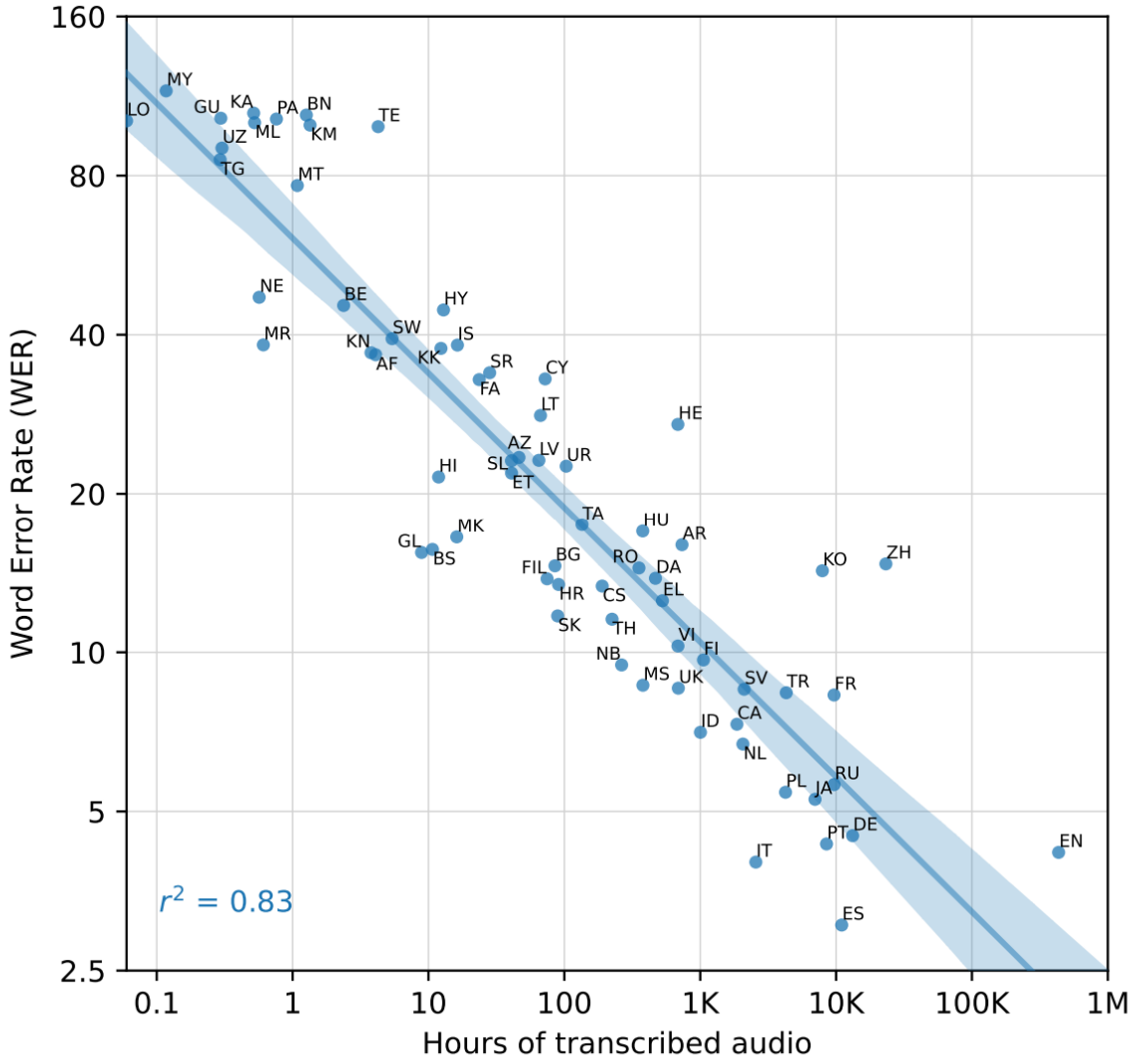


Figure 3. **Correlation of pre-training supervision amount with downstream speech recognition performance.** The amount of pre-training speech recognition data for a given language is very predictive of zero-shot performance on that language in Fleurs.

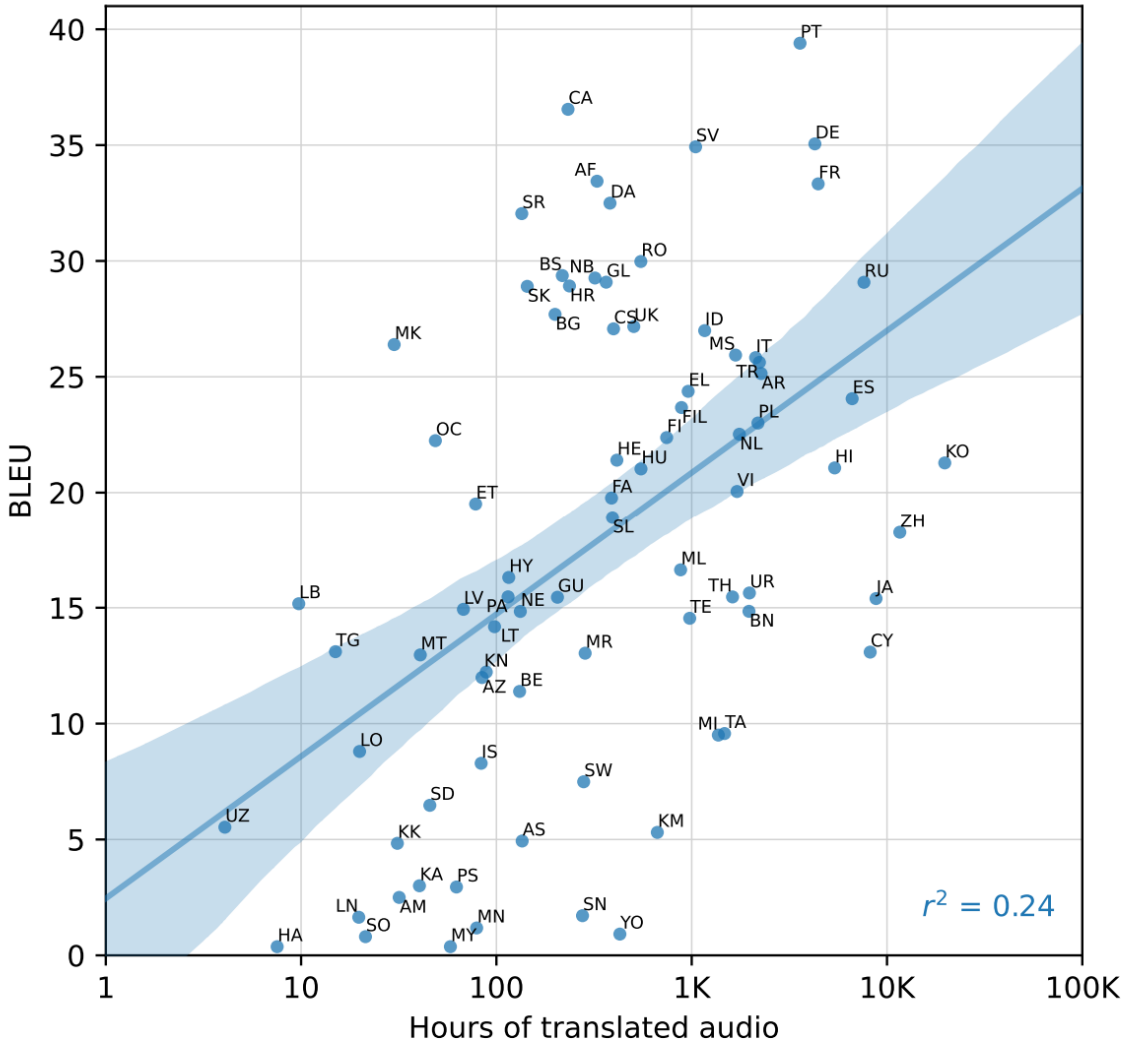


Figure 4. **Correlation of pre-training supervision amount with downstream translation performance.** The amount of pre-training translation data for a given language is only moderately predictive of Whisper's zero-shot performance on that language in Fleurs.

X → English	High	Mid	Low	All
XMEF-X	34.2	20.2	5.9	14.7
XLS-R (2B)	36.1	27.7	15.1	22.1
mSLAM-CTC (2B)	37.8	29.6	18.5	24.8
Maestro	**38.2**	31.3	18.4	25.2
Zero-Shot Whisper	36.2	**32.6**	**25.2**	**29.1**

Table 4. **X→en Speech translation performance.** Zero-shot Whisper outperforms existing models on CoVoST2 in the overall, medium, and low resource settings but still moderately underperforms on high-resource languages compared to prior directly supervised work.

Language ID	Fleurs
w2v-bert-51 (0.6B)	71.4
mSLAM-CTC (2B)	**77.7**
Zero-shot Whisper	64.5

Table 5. **Language identification performance.** Zero-shot Whisper's accuracy at language identification is not competitive with prior supervised results on Fleurs. This is partially due to Whisper being heavily penalized for having no training data for 20 of Fleurs languages.