Speech Data from Radio Broadcasts for Low Resource Languages

Bismarck Bamfo Odoom
Johns Hopkins University
bodoom1@jhu.edu

Paola Leibny Garcia
Johns Hopkins University
lgarci27@jhu.edu

Prangthip Hansanti
Meta AI
prangthiphansanti@meta.com

Loïc Barrault
Meta AI
loicbarrault@meta.com

Christophe Ropers
Meta AI
chrisropers@meta.com

Matthew Wiesner
Johns Hopkins University
wiesner@jhu.edu

Kenton Murray
Johns Hopkins University
kenton@jhu.edu

Alex Mourachko
Meta AI
alexmourachko@meta.com

Philipp Koehn
Johns Hopkins University
phi@jhu.edu

Abstract

We created a collection of speech data for 48 low resource languages. The corpus is extracted from radio broadcasts and processed with novel speech detection and language identification models based on a manually vetted subset of the audio for 10 languages. The data is made publicly available. [1]

1 Introduction

While automatic speech recognition systems have seen great gains in recognition accuracy, even under challenging acoustic conditions, this success is highly uneven across the languages in the world. For many languages in the world, even reliable audio training data is not easily available.

Motivated by this, we set out to collect and make publicly available speech data for languages that fall below the top one hundred languages, broadly measured by number of speakers and commercial relevance. We present a novel audio data set for 48 low resource languages. We report on manual efforts to vet collected audio data as well as automatic methods to extract speech from mixed audio data (especially discarding music) and language identification.

We collected this data mostly from radio broadcasts by recording audio streams available at Radio Garden[2]. These audio broadcasts are identified by location which gives us some guidance to which broadcasts are likely to contain audio in a desired language. We record audio snippets of 10–60 seconds in length. Since much of the audio data contains music, we developed a speech detection model to automatically identify audio files that consist of speech data and not music or other non-speech data.

Since there are no reliable speech language identification models or even identified speech data for a subset of these languages, we manually vetted audio data for 10 languages to create a corpus of about 5 hours of audio per language that has been verified by native speakers to be speech in each of the targeted languages.

With these tools in place (speech crawling, speech detection, speech language identification), we scaled up the effort to 48 languages. The resulting corpus of speech data consists of about 3000 hours of clean raw speech suspected to be in these low-resourced languages. Upon further filtering with language identification (LID) systems, this results in about 450 hours of clean speech.

2 Related Work

Foley et al. (2024) use audio data from Radio Garden to learn a mapping from speech to a geographic location. Conneau et al. (2022) create a dataset of 101 languages by recording audio from native speakers. The audio recorded stems from the Flores-101 dataset which consists of English sentences from Wikipedia translated into 101 languages. Pratap et al. (2023) introduce a massively multilingual dataset for over 1000+ languages based on recordings of publicly available religious texts. They further train self-supervised, automatic speech recognition, text-to-speech synthesis, and language identification models on this dataset. Radford et al. (2022) introduce a large-scale multilingual weakly supervised dataset consisting of about 680k hours of audio for speech

[1] https://huggingface.co/datasets/jhu-clsp/radio-broadcast

[2] https://radio.garden/

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 182–187

recognition. They showed that scaling the amount of data greatly improves the performance and robustness of speech recognition systems.

Unlabeled speech data has many uses in building speech applications. Representation learning methods like HuBERT (Hsu et al., 2021) and w2v-BERT (Chung et al., 2021) use raw speech data to distill semantic speech tokens from audio. Large-scale models such as Whisper (Radford et al., 2022), MMS (Pratap et al., 2023), or Seamless (Communication et al., 2023) rely partly on raw speech data to scale to hundreds of languages.

3 Corpus Collection

The sources of our data are radio broadcasts that are transmitted freely over the Internet. We use Radio Garden to discover and identify stations that broadcast in languages we target. Radio Garden identifies radio stations with the location from which they broadcast — which provides a pool of candidate stations for each language, based on the region where the language is spoken. The broadcasts are accessible through an API call.

We filter this pool of candidate stations by checking manually if they likely broadcast in the targeted language (opposed to, say, English) or exclusively broadcast music. Since this effort often relies on researchers that are not familiar with the languages, the process is necessarily imperfect. Another obstacle is that some radio station broadcasts are not reliably delivered over the Radio Garden platform, leading to gaps in the data collection.

We break up the audio signal into segments of different lengths, ranging from 10 to 60 seconds. The raw audio is also converted to the FLAC files and re-sampled to 16kHz. We collected this data throughout 2023 and early 2024.

4 Speech Detection

To filter out audio files containing music, we use a convolutional recurrent neural network (CRNN) (Hung et al., 2022) which was trained on a high-quality dataset (Hung et al., 2022) of speech and music activity labels. The CRNN model predicts the probability of music and speech for each audio frame.

We also use a feature-based model that calculates the average energy in each chunk of the audio spectrogram. This energy level indicates the intensity of the audio within that chunk. Chunks with energy levels higher than 0.5 are classified as music.

We set the detection threshold of the CRNN model to 0.9 and that of the feature-based model to 0.5. Audio files classified as not having music in them by both models are kept and the rest are discarded.

5 Manual Vetting

We are addressing several languages for which we do not have reliable language identification methods, or even any speech data that is verified to be in the presumed language. Hence, we engaged speakers of these languages to verify that speech audio that we presumed to be in their language was indeed in their language.

We carried out this manual vetting for Igbo, Luo (a.k.a. Dholuo), Ganda (a.k.a. Luganda), Nyanja, Maithili, Marwari, Santali, Meitei (a.k.a. Manipuri), Yue Chinese, and Central Kurdish. We recruited native speakers of these languages through language service providers. We carried out this vetting process through three phases, with increasingly larger quantities and more detailed questions.

Phase 1 Since we collected audio from only a few radio stations, our first question was to know which of them are reliable sources of speech data in the targeted languages. We sampled about a hundred 30-second speech segments per language and asked the language experts to assess whether those were indeed in their language. We also encouraged them to identify other language(s) that may be present in utterances, as well as the presence of non-speech or incomprehensible audio. For several languages, the experts also reported code-mixing with other languages, especially for Maithili, Marwari, Meitei, and Santali. Table 1(a) shows the results of the study. We considered as *good* those samples that have at least 90% audio in the targeted language. For 3 languages, we repeated the exercise since the first phase did not yield sufficient positively identified audio segments.

Phase 2 In the second phase, we scaled up the experiment to more audio samples. Here, the audio samples were of different lengths (10s, 20s, 30s, and 60 seconds). We also asked detailed questions about music being present in the background, speech being spontaneous or scripted, and about the presence of multiple speakers. Table 1(b) shows the results of the study. For most of the languages,

(a) Phase 1: Language identification

Language	Good	Total	Other languages detected
Central Kurdish	1+45	67+119	Arabic, Kurdish Bahdini, Kurdish Kurmanji, English
Ganda	47	95	English, Swahili
Igbo	12	90	Nigerian Pidgin English, Latin American Spanish, English-Spanish (Spanglish), Yoruba, US English, Pidgin, Nigerian English, British English
Luo	73	94	Swahili, English
Maithili	80	104	Nepali, Hindi, English
Marwari	55+92	120+120	-
Meitei	94	99	Hindi
Nyanja	58	91	English
Santali	0+45	107+120	Bengali, Hindi, English
Yue	59	91	Mandarin

(c) Phase 2: Larger sample, more detailed questions

Language	Total	Good	Music (yes/no)		Scripted/Spontaneous		Speakers (1/more)	
Central Kurdish	640	407	44	363	71	336	190	217
Ganda	645	577	296	281	262	315	306	271
Igbo	636	235	185	50	157	78	96	139
Luo	645	473	463	10	441	32	396	77
Maithili	480	352	31	321	195	157	245	107
Marwari	640	208	176	32	173	35	139	69
Meitei	624	516	89	427	175	341	263	253
Nyanja	644	435	282	153	267	169	256	180
Santali	640	309	105	204	248	61	125	184
Yue	646	354	58	296	24	272	51	248

(c) Phase 3: Scaling up data sizes for some languages with cleaner sources

Language	Total	Good	Music (yes/no)		Scripted/Spontaneous		Speakers (1/more)	
Central Kurdish	240	237	4	213	41	196	131	106
Ganda	105	102	17	85	55	47	60	42
Igbo	216	195	11	184	0	195	145	50
Maithili	337	331	21	263	47	284	72	191
Meitei	222	222	8	213	164	57	172	49
Nyanja	222	216	15	201	138	78	115	101
Santali	640	640	57	573	42	598	380	260
Yue	285	284	4	280	17	267	41	243

Table 1: Manual vetting of speech data by language experts: The goal of this study was to identify 5 hours of vetted audio in the targeted language to be able to train language identification models.

	FLEURS									
	C.Kurdish	Ganda	Igbo	Luo	Marwari	Maithili	Meitei	Nyanja	Santali	Yue
MMS	98.3	99.8	98.3	99.6	-	-	-	95.1	-	99.9
Ours	87.7	88.9	10.6	0.4	-	-	-	46.3	-	88.5
	RADIO BROADCAST									
	C. Kurdish	Ganda	Igbo	Luo	Marwari	Maithili	Meitei	Nyanja	Santali	Yue
MMS	99.2	62.8	85.1	61.6	-	54.5	43.4	98.1	86.1	99.9
Ours	**99.9**	**92.4**	64.7	**93.2**	-	**97.1**	**99.9**	88.7	**95.2**	99.3

Table 2: Comparing the accuracy of our LID model to the MMS LID model (Pratap et al., 2023) on the FLEURS and radio broadcasts test sets

Language	Hours
Central Kurdish	3.30
Ganda	4.96
Igbo	1.14
Luo	4.00
Maithili	1.95
Manipuri	4.38
Marwari	1.65
Nyanja	3.56
Santali	2.63
Sorani	3.39
Yue	2.96

Table 3: Amount of data per language used to train our LID models.

there is often some music in the background. The amount of scripted vs. spontaneous speech as well the number of speakers in the audio varies by language.

Phase 3 Since our goal was to collect at least 5 hours of vetted audio, we repeated the Phase 2 study on additional audio samples using the same vetting protocol. Table 1(c) shows the results. Given the feedback from the second phase, we were able to identify generally cleaner audio sources to be vetted, resulting in a much larger ratio of them assessed to be good and without background music. For logistical reasons, we were not able to do this for Luo and Marwari.

We will release the audio with meta data from the annotation effort publicly.

6 Language Identification

The LID system follows Villalba et al. (2023). Essentially, our LID uses log-Mel-filter banks with 64 filters as feature extractor. The features were short-time mean normalized with a 3-second window. Silence portions (frames) were removed using an energy voice activity detector (VAD) based on Kaldi. This VAD classifies each frame as speech or non-speech based on the average log-energy in a window.

The language embedding architecture follows the x-vector process (Snyder et al., 2017, 2018) as described by Villalba et al. (2023). It consists of an encoder that extracts frame-level discriminant embeddings, a pooling mechanism, and a classification head. We used the Res2Net architecture as the encoder. The system uses the datasets in the *Training Open* condition for training the language embedding. For the backend, the system employs a linear Gaussian classifier with a single Gaussian per target language, and a shared-covariance across languages. The system is trained on about 30 hours of audio in 10 languages. Table 3 shows the distribution of data per language.

As shown in Table 2, we compare the performance of our LID model to the MMS LID model (Pratap et al., 2023) on the FLEURS (Conneau et al., 2022) benchmark and a carefully selected test set comprising radio broadcast recordings. FLEURS is in a similar domain to the data used to train the MMS model, and the test set of radio broadcasts is in the same domain as the data used to train our model. The MMS LID model was trained on 1000 times more data as compared to ours.

Luo's severe performance drop on FLEURS is due to the difference in the dialects in FLEURS and radio broadcast test sets. The poor performance of Igbo on both test sets is due to the small amount of data in Igbo used in training the LID system. For most languages, our LID model outperforms the

MMS model on the radio broadcast data.

7 Corpus

With all the tools in place, we scaled up the effort to collect audio speech data for all the targeted 48 languages. Table 4 gives details about the number of hours of audio data we handled at various processing stages: (1) the number of hours of crawled audio expected to be in the targeted language, (2) the number of hours after speech detected, and (3) what remained after a language ID filter.

For the 10 targeted languages (bold in the table), we collected substantial amounts of data, ranging from 12.43 hours (Marwari) to 178.55 hours (Maithili) after music detection and language ID filtering.

Scaling up to 48 language was challenging as we could not repeat the expensive first stage of annotations to identify radio stations which broadcast in the languages of interest. We randomly pick radio stations within locations we believe speak the languages of interest and collect data from them. Since we did not run annotations for the new languages we did not have ground-truth data to train LID models for those languages. We rely on the MMS LID model for these languages. Specifically, we use the variant trained on 4017 languages.

The amount of data collected per language varies due to the number of radio stations we collected data from at each time. For some languages, we identified many radio stations that broadcast in the language of interest, enabling us to collect hundreds of hours of data. Also, we aggressively filtered the corpus for music, which greatly affected the amount of data we collected for some languages. We could not report on the amount of data after LID for Egyptian, Morrocan, and Pashto as the MMS model does not support them. Other languages with no data after LID had none of the top predictions of the audio files to be in the language. This data was collected from early 2023 to early 2024.

8 Conclusion

We collected a large corpus of speech audio for 48 languages from audio sources. We focused special attention to 10 languages for which we built language identification models based on manually vetted audio data. We will release all audio data (manually vetted and automatically filtered) open source with a liberal license for research and commercial use. We hope that this data fosters research

Languages	Crawled	Clean	LID
Amharic	83.74	20.44	7.94
Armenian	82.35	9.03	2.13
Assamese	85.03	16.77	0.13
Azerbaijani	96.71	4.45	1.79
Belarusian	101.53	0.84	0.10
Bosnian	63.48	3.67	1.29
Cebuano	64.53	1.00	0.02
C. Kurdish	**75.53**	**46.74**	**23.51**
Egyptian	108.19	10.32	-
Galician	75.35	31.60	0.69
Ganda	**293.65**	**125.97**	**24.25**
Georgian	65.25	1.42	0.05
Gujarati	95.99	0.13	0.02
Icelandic	134.99	11.22	5.47
Igbo	**137.95**	**12.12**	**4.21**
Irish	200.41	15.62	0.06
Javanese	25.37	5.97	0.14
Kannada	40.53	1.94	0.96
Kazakh	83.67	4.07	1.58
Khmer	21.99	2.59	2.07
Konkani	72.93	4.01	-
Kyrgyz	51.05	6.75	1.26
Lao	108.27	10.19	1.91
Luo	**409.3**	**243.38**	**48.46**
Macedonian	62.66	0.51	0.24
Maithili	**2860.84**	**1722.91**	**178.55**
Maltese	89.75	14.68	4.51
Meitei	**299.50**	**129.97**	**18.13**
Marathi	139.25	25.06	9.24
Marwari	**155.46**	**118.05**	**12.43**
Mongolian	33.25	2.91	0.66
Moroccan	184.80	11.73	-
Nepali	53.15	3.61	0.81
Nyanja	**251.11**	**79.20**	**22.41**
Odia	106.61	1.20	-
Oromo	117.52	14.77	0.18
Panjabi	45.63	0.57	-
Pashto	40.81	6.58	-
Santali	**272.65**	**120.06**	**20.45**
Shona	70.19	15.71	3.17
Sindhi	33.22	10.38	0.19
Swiss German	584.60	86.86	-
Tajik	26.34	1.21	0.49
Telugu	28.98	0.51	0.10
Uzbek	49.71	5.88	2.44
Welsh	67.29	2.14	0.12
Yue	**117.28**	**101.21**	**64.70**
Zulu	49.51	24.03	0.04

Table 4: Statistics of the collected audio data (in hours). The focus languages for which we performed manual vetting and more thorough radio station selection are in bold.

in low resource speech technology.

Limitations

The legal status of web crawled data is currently in a gray area. We argue that the released data set falls under fair use since we are releasing disconnected snippets and do not interfere with the commercial use of the original broadcasts.

References

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *CoRR*, abs/2108.06209.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussá, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation. Technical report, Meta FAIR.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech.

Patrick Foley, Matthew Wiesner, Bismarck Bamfo Odoom, Leibny Paola Garcia, Kenton Murray, and Philipp Koehn. 2024. Where are you from? geolocating speech and applications to language identification. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistcs (NAACL)*.

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. *CoRR*, abs/2106.07447.

Yun-Ning Hung, Chih-Wei Wu, Iroro Orife, Aaron Hipple, William Wolcott, and Alexander Lerch. 2022. A large TV dataset for speech and music activity detection. In *J AUDIO SPEECH MUSIC PROC*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. 2017. Deep Neural Network Embeddings for Text-Independent Speaker Verification. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association, INTERSPEECH 2017*, pages 999–1003, Stockholm, Sweden. ISCA.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors : Robust DNN Embeddings for Speaker Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018*, pages 5329–5333, Alberta, Canada. IEEE.

Jesús Villalba, Jonas Borgstrom, Maliha Jahan, Saurabh Kataria, Leibny Paola Garcia, Pedro Torres-Carrasquillo, and Najim Dehak. 2023. Advances in Language Recognition in Low Resource African Languages: The JHU-MIT Submission for NIST LRE22. In *Proc. INTERSPEECH 2023*, pages 521–525.