The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali

Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, Linne Ha

Google Research

{oddur,mungkol,thammaknot,mjansche,linne}@google.com

Abstract

We present speech corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. Each corpus consists of an average of approximately 200k recorded utterances that were provided by native-speaker volunteers in the respective region. Recordings were made using portable consumer electronics in reasonably quiet environments. For each recorded utterance the textual prompt and an anonymized hexadecimal identifier of the speaker are available. Biographical information of the speakers is unavailable. In particular, the speakers come from an unspecified mix of genders. The recordings are suitable for research on acoustic modeling for speech recognition, for example. To validate the integrity of the corpora and their suitability for speech recognition research, we provide simple recipes that illustrate how they can be used with the open-source Kaldi speech recognition toolkit. The corpora are being made available under a Creative Commons license in the hope that they will stimulate further research on these languages.

Index Terms: Malayo-Polynesian languages, Indo-Aryan languages, speech corpora, speech recognition

1. Introduction

Demand for data is increasing due to the advent of more sophisticated machine learning systems, which makes it even harder for lesser resourced languages to keep up. Having access to good resources is paramount for the speech community to reach feature parity for those lesser resourced languages. The challenges of developing speech technologies for under-resourced languages are summarized in [1], which provides valuable insights into why collecting these resources can be problematic.

Speech recognition systems depend heavily on data. At Google we have collected speech corpora in collaboration with external partners in the past, where crowd sourcing efforts were employed to collect the data quickly and cost efficiently. Crowd sourcing the work is not new when it comes to collecting speech corpora; our existing approach and tools have been described in [2].

This paper gives an overview of several new speech recognition corpora which we have collected in collaboration with external partners. We give a brief overview of what is in the datasets and how the data were collected. At the end of the paper we give an example of simple Kaldi [3] recipes which were used to test the validity of the Javanese and Sundanese data.

This paper is organized as follows: Related work is discussed in Section 2. A general overview of resources needed for speech recognition systems is given in Section 3. Descriptions of the corpora and details of how they were collected are provided in Section 4. Section 5 describes the tools for adapting the corpora to be used with Kaldi. Section 6 describes preliminary experiments with training simple Kaldi speech recognition models on three of our corpora. In Section 7 we conclude.

2. Related Work

Different approaches have been tried for collecting corpora. [4] gives an overview of how an Icelandic dataset was created using crowd sourced methodology. [5] describes the work done on verifying the quality of the Icelandic corpus. [6] discusses a crowd sourced approach to collect text-to-speech corpora for Javanese and Sundanese. Work on verifying the quality of the data being collected can be found in [7], where a simple system was bootstrapped which attempts to identify if the incoming data are of good quality.

3. Resources for Building Speech Recognition Corpora

For speech recognition system the following resources are needed:

- Waveforms (audio files) to train an acoustic model.
- Transcriptions of the audio.
- Phonology of the target language.
- A lexicon for the target language.
- A language model, or text to generate a language model for the target language.

The amount of audio data required depends on the scope of the system. The audio can be collected in different ways, ranging from a professional studio setup to asking volunteers to download and use data collection software on their mobile devices.

The transcriptions can either be done by listening to the audio after it has been collected, which is a time consuming effort, or prompting the people being recorded with text and collecting read speech. The latter eliminates the need to transcribe the audio, but does not eliminate the chance of errors in reading the text, or other problems which might affect the quality of the audio such as loud background noises.

Before working on the lexicon, the phonology should be described, which is minimally an enumeration of the phonemes that are present in the language.

The lexicon provides phonemic transcriptions of words in the target phonology. The minimum requirement is a lexicon which covers all the words present in the script used to record the corpus. For a medium sized corpus of about 200k utterances, one can expect to have tens of thousands of words in the lexicon. Transcribing these words by hand would be a sizable task on its own. This could also done by writing grapheme-to-phoneme (G2P) rules, or learning rules from an existing smaller lexicon. Building the lexicon before the text prompts gives the advantage of being able to verify the frequency of phonemes in the text prompts, and balancing the prompts before recording.

Another important part of a speech recognition system is the textual Language Model (LM). In classical speech recognition, the textual language model provides an estimate of the

10.21437/SLTU.2018-11

Table 1: *Overview of Corpora.*

Language	Locale	Collected in	Year	Recordings	Hours	Speakers	URL
Javanese	jv-ID	Yogyakarta	2016	185,076	296	1019	openslr.org/35
Sundanese	su-ID	Bandung	2016	219,156	333	542	openslr.org/36
Sinhala	si-LK	Sri Lanka	2015	185,293	224	478	openslr.org/52
Bengali	bn-BD	Bangladesh	2015	232,537	229	508	openslr.org/53
Nepali	ne-NP	Nepal	2015	157,905	165	527	openslr.org/54

prior probability of seeing a particular sequence of words, regardless of their acoustic realization. The language model combines with the acoustic model, which in turn provides the likelihood of the observed waveform given a particular hypothesis of the words. The language model is built using text corpora for the language, minimally providing the probability of *n*-grams of words, or of whole sentences. A language model can also be a hand-crafted grammar, which can be the case for limited-domain applications.

4. Overview of the Corpora

We have collected and released speech corpora in five languages of South and Southeast Asia: Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. These corpora are available under a liberal Creative Commons license and can be downloaded from OpenSLR.org. Table 1 provides a summary overview of the corpora. The rest of this section describes them in further detail.

Each corpus consists of the following items:

- Audio files containing the recorded utterances.
- A TSV file called `utt_spk_text.tsv`, which consists of one row per utterance and three tab-separated columns. The columns contain the `UtteranceID`, anonymized `SpeakerID`, and `Text` transcriptions of the utterances.

A lexicon and phonology (`non_silent_phonemes`) are derived *naively* and automatically as part of our Kaldi recipes (see below) from the `utt_spk_text.tsv` file.

Each line in the file `utt_spk_text.tsv` corresponds to one line of read text, and the basename of the recording is the `UtteranceID`. The audio data were recorded in mono as 16-bit linear PCM with a 16kHz sampling rate. They are stored in Free Lossless Audio Codec (FLAC) format[1] with file extension `.flac`. As part of the Kaldi recipes, the `.flac` files are decompressed on-the-fly to RIFF `.wav` format.

The recordings were provided by native-speakers in the respective regions. In all cases we found local volunteers who were willing to help out with the data collection. The aim was to find between 5 and 20 people interested in becoming data specialists. The data specialists were trained in using the data collection tools, and the specialists then reached out to additional volunteers who supplied recordings. The text prompts came from open and available sources which can be found online. The text prompts used were usually not much longer than about 10 words, so that they could be displayed in a readable way on a smartphone.

For Bengali, Sinhala, Nepali and Sundanese a tool developed at Google called DataHound – described in [2] – was used for the data collection. All data were collected using standard consumer smartphones. No specialized or additional hardware was used for the data collection. The audio was first saved to local storage on the device and then uploaded to a server once a connection to the internet was established.

The data collection for Javanese was performed in collaboration with the Javanese Literature Department of Universitas Gadjah Mada (UGM) in Yogyakarta, Indonesia, a Special Administrative Region in central Java. Our colleagues at UGM connected us with volunteers who we helped train as data specialists for this effort.

Two tools were used in the data collection for Javanese: in addition to DataHound we used the open-source tool Eyra [7] developed at Reykjavik University and available on GitHub[2]. The ASR data collection was done at the same time as a text-to-speech data collection for Javanese [6].

The data collection for Sundanese was performed in a similar fashion in collaboration with colleagues from Universitas Pendidikan Indonesia (UPI) in Bandung, West Java, Indonesia.

The languages covered in the present set of corpora have the potential to reach hundreds of millions of speakers:

- Javanese is the second largest language of Indonesia (after the national language, Indonesian), spoken by about 90 million people. The language is a regional language on the island of Java in Indonesia. Javanese belongs to the Malayo-Polynesian subgroup of the Austronesian languages. Our corpus uses the contemporary Javanese writing system based on the Latin alphabet.
- Sundanese is the third largest language of Indonesia, spoken by about 40 million people. Similar to Javanese, it is a regional language on the island of Java in Indonesia, in the Malayo-Polynesian subgroup of Austronesian. Our corpus uses the contemporary Sundanese writing system based on the Latin alphabet.
- Nepali is an official language of Nepal and a regionally official language in neighboring regions of India. It is spoken by about 20 million people. Nepali belongs to the Eastern Pahari subgroup of the Indo-Aryan languages and is written in Devanagari script.
- Sinhala is one of the two official languages of Sri Lanka, alongside Tamil. Sinhala is spoken by about 14 million people. Sinhala belongs to the Insular Indic subgroup of the Indo-Aryan languages, and is written in Sinhala script, a southern Brahmic script.
- Bengali is the national language of Bangladesh, as well as an official language in India. Bengali is spoken as a first or second language by about 160 million people in Bangladesh, and about 100 million people in India, making it one of the most widely spoken languages in the

[1] `https://xiph.org/flac`

[2] `https://github.com/Eyra-is/Eyra`

world. Bengali belongs to the Bengali-Assamese subgroup of the Indo-Aryan languages. Our corpus uses the standard Bengali-Assamese (Eastern Nagari) script.

5. Kaldi Recipes

In addition to the speech corpora, we are further providing scripts for making the corpora usable with Kaldi [3]. The goal of these recipes is to validate the integrity of the corpora and to show how they can be straightforwardly used for research in speech recognition. Building competitive ASR systems was not a goal.

Our recipes are available on GitHub[3] under an Apache license. They consist of scripts for data preparation. The acoustic model training itself re-uses existing Kaldi recipes for training on Wall Street Journal [8] and Resource Management [9] corpora; those recipes are available as part of the example recipes in the Kaldi source distribution. Most of the tools used here simply set up our corpora for use with those existing recipes. Minor cosmetic changes and simplifications have been made to the script which handles the training and testing part of the recipe.

The bulk of our additional tools is concerned with consuming the file `utt_spk_text.tsv` distributed as part of each corpus, containing the utterance and anonymized speaker identities as well as the utterance text. Our recipe generates the files expected by the existing Kaldi recipes in the formats used by Kaldi.

For the languages we are concerned with, large enough pronunciation lexicons were not immediately available which would be expected to cover all the words occurring in the textual prompts. In order to train a simple classic speech recognition system, a grapheme-based approach was used, akin to the one described in [10]. Instead of splitting the words into characters, the words were split into pairs of of adjacent characters. For example the word `Matur` would be transcribed as `_m ma at tu ur r_`. The grapheme-based lexicons are trivially generated placeholders that can and should be replaced with proper phonemic lexicon if and when those become available. In a similar fashion, the phonological description simply assumes that all character pairs occurring as part of the transcriptions in the lexicon are non-silent "phonemes".

Our main goal here is to validate that basic Kaldi recipes can be run which will train classical monophone and triphone acoustic models. This in turn provides feedback about the quality of the data, e.g. in the form of alignment information that can be used to check whether waveforms and training transcriptions match. Our approach is insufficient for producing competitive ASR baseline results; if that were the goal, the placeholder files for lexicon and phonology should be replaced and more sophisticated models should be trained. However, that is beyond the aim and scope of this paper.

5.1. Data Conversion Tools

Two main scripts were developed to convert to the format the existing recipes use to consume the corpora. These are basic Python scripts, which read input files and write out to standard output the format needed.

- `kaldi_converter.py` is the main script which consumes the corpus index and outputs data file for consumption by Kaldi.
- `corpus_util.py` is a utility library with a container class for the corpora.
- `simpleg2g.py` is a helper script which generates the graphemic lexicon described above.

Example usage of the `kaldi_converter.py` script would be:

```
$ python kaldi_converter.py \
    -d [corp. dir]  \
    -f [corp. file] \
    --utt2spk
```

For this one of our corpora must have been downloaded and unpacked (helper scripts are provided for that) into a corpus directory. The converter script reads the `utt_spk_text.tsv` file (called corpus file here) and outputs the `utt2spk` mapping file required by Kaldi. The generated file is a two-column TSV file which provides a mapping from utterance ID to speaker ID – information that can be trivially derived from the corpus file `utt_spk_text.tsv`. Along the same lines, all other files required for the Kaldi recipes are generated by the converter script.

To reduce storage and transmission cost and time, the downloadable corpora provide the recorded utterances in losslessly compressed FLAC format. Our recipes set them up to be uncompressed on-the-fly to RIFF `.wav` format as part of Kaldi feature extraction.

6. Experiments and Analysis

We experimented with simple classical monophone and triphone recognizers for Sinhala, Javanese, and Sundanese in order to validate the integrity of our corpora and associated data preparation scripts. As noted above, this by itself is not sufficient for producing competitive ASR models, which was not one of our goals.

We followed an identical procedure for all languages: The full dataset was divided into two subsets, one for training, and the other for testing. The test data was produced by grouping the data by `SpeakerID` and holding the first 2000 lines out; the remainder was used for training.

The graphemic lexicon and phonology were generated using the tools described in Section 5: we first auto-generated the graphemic lexicon with character pairs, then generated the "phoneme" inventory to comprise those grapheme pairs present in the lexicon.

The lexicon, phonology, and language model were derived from all the lines available in the full dataset. In other words, by design there cannot be any out-of-vocabulary words, and all test sentences are present in the language model. Needless to say, this would not be the right methodology in a competitive ASR evaluation, but that was not our goal. Instead, we wanted to use alignment information and analyze mistakes on previously seen data to scan for systematic data problems. The word-error rates reported below are highly optimistic and should not be used as baselines for future work.

Once all the resources have been assembled, the recipe converts the corpus files into the required format, runs MFCC feature extraction, performs flat-start training of a monophone model, uses that to align the data, and trains a triphone model. The Resource Management recipe provided with the Kaldi distribution would perform further training steps, but for our purposes (and for the reasons above) the simple monophone and

[3] `https://github.com/googlei18n/asr-recipes`

Table 2: *WER for the models built.*

Language	Monophone WER	Triphone WER
Sinhala	35.18	23.29
Javanese	19.31	9.45
Sundanese	17.92	3.16
Javanese + Sundanese	31.21	10.70

triphone models are sufficient for exploring the data. For reference, we show word-error rates in Table 2.

In future work we would like to provide similar recipes for the rest of the languages described here; due to the nature of the writing systems, our naive graphemic approach turns out to be insufficient, creating a lexical challenge that needs to be addressed first. Using the data to build more sophisticated state-of-the-art ASR systems would be of interest, as would jointly training models with some or all of the languages provided. Other work of interest would be using methodologies similar to those described in [5] and [7] to assess the quality of the data in more detail.

7. Conclusion

We have collected and released speech corpora for five languages of South and Southeast Asia. The corpora are publicly available under a liberal Creative Commons open-source license. We have provided simple tools for using the corpora with the popular open-source Kaldi speech-recognition toolkit. The combination of existing open-source toolkits with our open-source corpora and recipes greatly reduces the initial cost of embarking on research in these traditionally under-resourced languages. We hope that these efforts will facilitate future research by the broader scientific community and will encourage other researchers to make similar datasets available.

8. References

[1] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[2] T. Hughes, K. Nakajima, L. Ha, A. Vasu, P. J. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 1914–1917.

[3] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[4] J. Guðnason, O. Kjartansson, J. Jóhannsson, E. Carstensdóttir, H. H. Vilhjálmsson, H. Loftsson, S. Helgadóttir, K. Jóhannsdóttir, and E. Rögnvaldsson, "Almannarómur: An open Icelandic speech corpus," in *Third Workshop on Spoken Language Technologies for Under-resourced Languages, SLTU 2012, Cape Town, South Africa, May 7-9, 2012*, 2012, pp. 80–83.

[5] S. Steingrímsson, J. Guðnason, S. Helgadóttir, and E. Rögnvaldsson, "Málrómur: A manually verified corpus of recorded Icelandic speech," in *Proceedings of the 21st Nordic Conference on Computational Linguistics, NODALIDA 2017, Gothenburg, Sweden, May 22-24, 2017*, 2017, pp. 237–240.

[6] J. A. E. Wibawa, S. Sarin, C. Li, K. Pipatsrisawat, K. Sodimana, O. Kjartansson, A. Gutkin, M. Jansche, and L. Ha, "Building open Javanese and Sundanese corpora for multilingual text-to-speech," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*, 2018.

[7] M. Petursson, S. Klüpfel, and J. Guðnason, "Eyra – Speech data acquisition system for many languages," in *SLTU-2016, 5th Workshop on Spoken Language Technologies for Under-resourced languages, 9-12 May 2016, Yogyakarta, Indonesia*, 2016, pp. 53–60.

[8] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "Continous speech recognition (CSR-I) Wall Street Journal (WSJ0) news, complete," *Linguistic Data Consortium, philadelphia*, 1993.

[9] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1988, pp. 651–654.

[10] M. Killer, S. Stuker, and T. Schultz, "Grapheme based speech recognition," in *Eighth European Conference on Speech Communication and Technology*, 2003.