Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6494–6503
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems

Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Işın Demirşahin, Cibu Johny, Martin Jansche†, Supheakmungkol Sarin, Knot Pipatsrisawat
Google Research
Singapore, United States and United Kingdom
{oddur,rivera,agutkin,isin,cibu,mungkol,thammaknot}@google.com

Abstract

We present free high quality multi-speaker speech corpora for Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu, which are six of the twenty two official languages of India spoken by 374 million native speakers. The datasets are primarily intended for use in text-to-speech (TTS) applications, such as constructing multilingual voices or being used for speaker or language adaptation. Most of the corpora (apart from Marathi, which is a female-only database) consist of at least 2,000 recorded lines from female and male native speakers of the language. We present the methodological details behind corpora acquisition, which can be scaled to acquiring data for other languages of interest. We describe the experiments in building a multilingual text-to-speech model that is constructed by combining our corpora. Our results indicate that using these corpora results in good quality voices, with Mean Opinion Scores (MOS) $>$ 3.6, for all the languages tested. We believe that these resources, released with an open-source license, and the described methodology will help in the progress of speech applications for the languages described and aid corpora development for other, smaller, languages of India and beyond.

Keywords: speech corpora, low-resource, text-to-speech, Gujarati, Kannada, Marathi, Malayalam, Tamil, Telugu, open-source

1. Introduction

Voice communication is one of the most natural and convenient modes of human interaction. As technologies in this field have advanced, computer applications that can use natural speech to communicate with users have become increasingly popular. In this work, we deal with six out of the twenty two official languages of India (Mohanty, 2006; Mohanty, 2010): Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu, which have combined speaker population of close to 400 million people. Although the situation with the speech corpora availability for these languages has been improving, these languages are still considered by many to be low-resource (Besacier et al., 2014; Srivastava et al., 2018). Furthermore, the resources available for building speech technology (and text-to-speech (TTS) applications, in particular) for these languages are still relatively scarce compared to those of Hindi, the most widely-spoken language of India. We had published the Bangla speech corpora previously (Gutkin et al., 2016; Kjartansson et al., 2018) and these are the next six largest languages of India.

There are four main resource components required to construct a classical TTS system: a speech corpus, a phonological inventory, a pronunciation lexicon and a text normalization front-end. Among these four components, speech corpora are usually the most expensive to develop. In the conventional approach, one would need to carefully design the recording script with the help of a linguist, recruit a voice talent, rent a professional studio and manage the recordings making sure the good quality is maintained throughout (Pitrelli et al., 2006; Ni et al., 2007; Sonobe et al., 2017). The whole operation would typically take months and is a major effort and investment, especially if state-of-the-art quality acceptable in the industry is required.

The process of assembling a high-quality TTS corpora for a low-resource language often becomes even more involved, both in terms of time required to collect the data (e.g., difficulty finding the professional voice talent or recording environment) and potentially higher cost of procuring or building from scratch the necessary linguistic components, e.g., a detailed tonal pronunciation dictionary for Burmese (Watkins, 2001) or Lao (Enfield and Comrie, 2015), either due to the scarcity of such resources or due to the difficulty of finding people with the necessary linguistic expertise to undertake such work (Dijkstra, 2004; Zanon et al., 2018).

Potential issues with constructing TTS corpora can be alleviated thanks to the recent advances in utilizing the found data (Cooper, 2019; Baljekar, 2018), adaptation of the existing corpora to TTS needs (Zen et al., 2019) and development of novel techniques exploiting multilingual sharing, such as transfer learning (Baljekar et al., 2018; Chen et al., 2019; Nachmani and Wolf, 2019; Prakash et al., 2019). Because the crawled data or general audio corpora often results in TTS models that have quality somewhat below current state-of-the-art, we are primarily interested in the corpora that is significantly smaller in size, but has higher recording quality, with the aim of combining several such corpora within a single model. Previous research on the subject (Li and Zen, 2016; Gutkin, 2017; Achanta, 2018; Wibawa et al., 2018; Nachmani and Wolf, 2019) established the feasibility of utilizing the audio data not just from one person but from multiple speakers, as well as leveraging the existing audio data from related languages.

This approach is comparatively cost-effective, since we can utilize multiple volunteer speakers recorded relatively cheaply using a simple setup consisting of a microphone, a laptop and a quiet room instead of relying on one professional voice talent recorded in a dedicated studio. Since

†The author contributed to this paper while at Google.

none of the volunteer speakers are professional voice talents, it is difficult for them to record big volumes of consistent (in terms of quality) audio in a single or even multiple sessions. Hence, by relaxing the requirement on the amount of data recorded by an individual speaker, we can scale the size of the dataset to any required size by simply recruiting more volunteers instead of increasing the recording burden on the existing ones. This work builds upon our previous initiatives in constructing speech corpora for low-resourced languages in South Asia and beyond: Bangladeshi Bangla, Nepali, Khmer and Sinhala (Wibawa et al., 2018; Kjartansson et al., 2018), Javanese and Sundanese (Sodimana et al., 2018) and Afrikaans, isiXhosa, Sesotho and Setswana (van Niekerk et al., 2017).

This paper is organized as follows: The next section provides a brief survey of the related corpora. Section 3 introduces the datasets. Then, in Sections 4 and 5, we provide the details of the data acquisition process, starting from recording script building to the audio recording and quality control processes. We provide the corpora details and present the results of quality evaluations in Section 6. Section 7 concludes this paper.

2. Related Corpora

Similar to observations by Wilkinson et al. (2016), we note that although there exist various TTS corpora for languages of India intended for research and applications, such as (Shrishrimal et al., 2012), they are generally proprietary, or available for research purposes only. One of the examples of such corpora is the Enabling Minority Language Engineering (EMILLE) corpus that has been constructed as part of a collaborative venture between Lancaster University, UK, and the Central Institute of Indian Languages (CIIL), Mysore, India (Baker et al., 2003). Part of the corpus includes audio data collected from daily conversations and radio broadcasts in Gujarati, Tamil and other languages of South Asia.

To the best of our knowledge, when it comes to Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu TTS corpora, the open-source corpora options, which are not encumbered by restrictive licenses, are not that many.

IIIT-H Datasets Perhaps the best known and to date the most widely used corpus is the TTS corpus from IIIT Hyderabad (Prahallad et al., 2012), which, among other languages, provides single-speaker male recordings of the languages in question, with the exception of Gujarati. The dataset for each language consists of 16 kHz audio recordings of 1,000 Wikipedia sentences selected for phonetic balance. This corpus served as de-facto standard TTS corpus for Indian languages for a number of years (Prahallad et al., 2013).

DeitY Datasets Alternative resource was produced by consortium of universities led by the Indian Ministry of Information Technology (DeiTY) (Baby et al., 2016). The resource has single-speaker TTS corpora for 13 Indian languages (including our languages of interest) consisting of 1,992 to 5,650 utterances per language. The audio was recorded at 48 kHz by professional voice talents in an anechoic chamber. This resource is becoming increasingly popular with the speech researchers dealing with Indian languages (Rallabandi and Black, 2017; Baljekar et al., 2018; Mahesh et al., 2018).

Language	Code	ISLRN	SLR Id
Gujarati	gu	276-159-489-933-8	SLR78
Kannada	kn	494-932-368-282-1	SLR79
Malayalam	ml	246-208-077-317-5	SLR63
Marathi	mr	498-608-735-968-0	SLR64
Tamil	ta	766-495-250-710-3	SLR65
Telugu	te	598-683-912-457-2	SLR66

Table 1: Dataset languages and the corresponding codes.

CMU Wilderness Dataset This speech dataset consists of aligned pronunciations and audio for about 700 different languages based on readings of the New Testament by volunteers (Black, 2019). Each language provides around 20 hours of speech. The dataset can be used to build single or multilingual TTS and automatic speech recognition (ASR) systems. Unfortunately at present this very interesting dataset does not include Gujarati and Kannada languages, but includes other lower-resource South Asian languages, such as Oriya (Pattanayak, 1969) and Malvi (Varghese et al., 2009).

Our Contributions Compared to the IIIT Hyderabad dataset, our corpora are multi-speaker and multi-gender, with almost twice the number of higher quality 48 kHz recordings for each gender and language. From our experience, the corpus of 1,000 utterances may not be enough to train a neural acoustic model, such as LSTM-RNN (Zen and Sak, 2015), let alone the state-of-the-art models (Oord et al., 2016; Wang et al., 2017). In addition, the crowdsourcing process we describe in this paper is more scalable than the process employed during for the construction of DeitY dataset. This is because it is easy to record more volunteer speakers if more data for a particular language is desired. Also, our data provides more variability in terms of the recording script coverage compared to the CMU Wilderness dataset that is restricted to Bible text. Finally, because the audio quality of our recordings is high, our data can be used as part of a larger multi-speaker multilingual corpus, which can be used to train systems such as the one reported by Gibiansky et al. (2017).

The key contributions of this work are:

- Methodology for affordable construction of text-to-speech corpora.
- The release of speech corpora for six important Indian languages with an open-source unencumbered license with no restrictions on commercial or academic use.

We hope that the release of this data will provide a useful addition to the Indian language corpora for speech research.

3. Brief Overview of the Datasets

The released datasets consist of Gujarati (Google, 2019a), Kannada (Google, 2019b), Malayalam (Google, 2019c), Marathi (Google, 2019d), Telugu (Google, 2019f) and

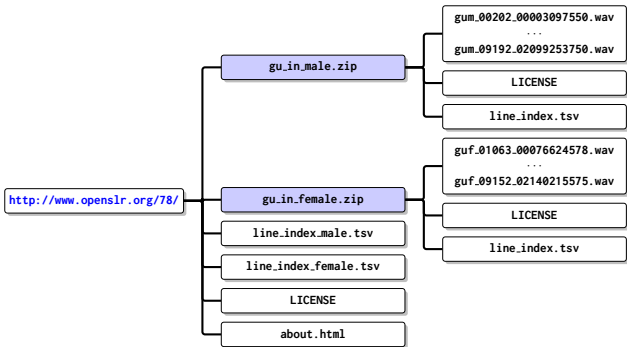


Figure 1: Layout of the Gujarati corpus.

Tamil (Google, 2019e). The brief synopsis of the released datasets is given in Table 1, where each of the six datasets is shown along the corresponding BCP-47 language code (Phillips and Davis, 2009), the International Standard Language Resource Number (ISLRN) (Mapelli et al., 2016) and the Speech and Language Resource (SLR) identifier from the Open Speech and Language Resources (OpenSLR) repository where these datasets are hosted (Povey, 2019). The ISLRN is a 13-digit number that uniquely identifies the corpus and serves as official identification schema endorsed by several organizations, such as ELRA (European Language Resources Association) and LDC (Linguistic Data Consortium).

The corpora are open-sourced under "Creative Commons Attribution-ShareAlike" (CC BY-SA 4.0) license (Creative Commons, 2019). The corpora structure follows the same lines for each language, similar to Figure 1, which shows the structure for Gujarati distribution. Collections of audio and the corresponding transcriptions are stored in a separate compressed archive for each gender (for Marathi only the female recordings are released). Transcriptions are stored in a *line index* file, which contains a tab-separated list of pairs consisting of the audio file names and the corresponding unnormalized transcriptions. The name of each utterance consists of three parts: the symbolic dataset name (e.g., Gujarati male is denoted `gum`), the five-digit speaker ID and the 11-digit hash.

4. Recording Script Development

4.1. Linguistic Aspects

Indian languages belong to several language families. In our set of languages, Gujarati and Marathi belong to the Indo-Aryan language family (Cardona and Jain, 2007; Dhongde and Wali, 2009), while Kannada, Malayalam, Tamil and Telugu are under the Dravidian tree (Steever, 1997). Apart from Gujarati, spoken in the central western part of the country, these languages are spoken mainly in the southern part of India. The numbers of native (L1) and second-language (L2) speakers are estimated to be around 374 millions and 47 millions, respectively (SIL International, 2019).

One important goal during the recording script preparation was to cover all phonemes in each language. We used a unified phoneme inventory for South Asian languages introduced by Demirsahin et al. (2018), where the unification capitalizes on the original observation by Emeneau (1956) that, on the one hand, the languages in question exhibit considerable phonological variation within each language group, and on the other, share several cross-group similarities. For example, the retroflex consonants of the six languages in question overlap significantly. In addition, our phoneme inventory has a large overlap between phonologically close languages, namely Telugu and Kannada, and Gujarati and Marathi. Table 2 shows the total size of the phonemic inventory for each language and the corresponding numbers of consonants and vowels. Difference in the counts between Marathi and Gujarati is due the presence of several consonantal phonemes which are specific to Marathi.

Language	Phonemes	Consonants	Vowels
Gujarati	40	32	8
Kannada	45	34	11
Malayalam	42	30	12
Marathi	49	41	8
Tamil	37	27	10
Telugu	45	33	11

Table 2: Number of phonemes (divided into consonants and vowels) in the language phonologies.

4.2. Recording Script Sources

This project was carried out with the intention to open-source the data from the start. Therefore, we avoided using copyrighted material to develop our corpora. Besides the absence of copyright, our objectives were (a) to have a variety of sentences (b) to include the most common words of the language and (c) to minimize the amount of manual review required. There are four sources of our script: (1) Wikipedia, (2) organic sentences that were handcrafted, (3) sentences created from templates (this process is explained in more detail in the next section) and (4) real-world sentences from various potential TTS application scenarios such as weather forecasts, navigation and so on. For Gujarati, Kannada, Malayalam, Telugu and Tamil, we only used source (1) (Wikipedia). The Marathi corpus was developed later on and included sentences from all of the aforementioned sources. To reduce the amount of human effort needed to create the corpus, we used source (3) (template-based sentences) as the main approach for Marathi script creation.

4.3. Template-based Recording Script Creation

To create sentences from templates, we first asked native speakers to list common named entities and numbers in each language, such as celebrity names, organization/place names, telephone numbers, time expressions, and so on. We then asked them to create 20–50 sentence templates that used these entities. The following are a few examples of such templates (given in English, for illustration purposes):

- *person name* was with *person name* on *time expression* for a meal at *place name*,
- *person name* is an officer of *organization name* in *country name* from *time expression* to *time expression*,

Figure 2: Recording equipment and environment.

Lang.	Female			Male		
	Duration		Spkrs	Duration		Spkrs
	total	avg		total	avg	
gu	4.30	6.97	18	3.59	6.30	18
kn	4.31	7.11	23	4.17	7.89	36
ml	3.02	5.17	24	2.49	4.43	18
mr	3.02	6.92	9		–	
ta	4.01	6.18	25	3.07	5.66	25
te	2.73	4.28	24	2.98	4.98	23

Table 3: Properties of the recorded speech corpora. Total durations are measured in hours, whereas average durations are measured in seconds.

- *person name* ordered *food name* and *drink name* at *location name*.

Italic words indicate placeholders that would be substituted with actual entities and expressions. Each template was carefully reviewed to make sure every entity/expression from the specified groups could be used as a fill in without causing any grammatical errors. Since Marathi is a highly inflectional language and requires grammatical agreement between phrases (Dhongde and Wali, 2009), extra attention had to be paid to devise the templates in such a way as to preserve the grammatical agreement in the resulting sentences. Once the templates were ready, sentences were then generated from these templates. For example, the first template above may yield the following sentence: "Theresa May was with Bill Gates on Monday for a meal at the Four Seasons Hotel."

4.4. Quality Control

We ensured that all sentences contained between five and twenty words. For sentences that were either manually created or needed to be reviewed (e.g., Wikipedia sentences), we asked native speakers to filter out typos, nonsensical or sensitive content and hard-to-pronounce sentences. We ensured that each script contained all the phonemes represented in the phoneme inventory for the language (briefly introduced in Section 4.1). We did not ensure an even coverage of phonemes within each script, as demonstrated by Figure 4 in Section 6, where the details of our experiments are provided.

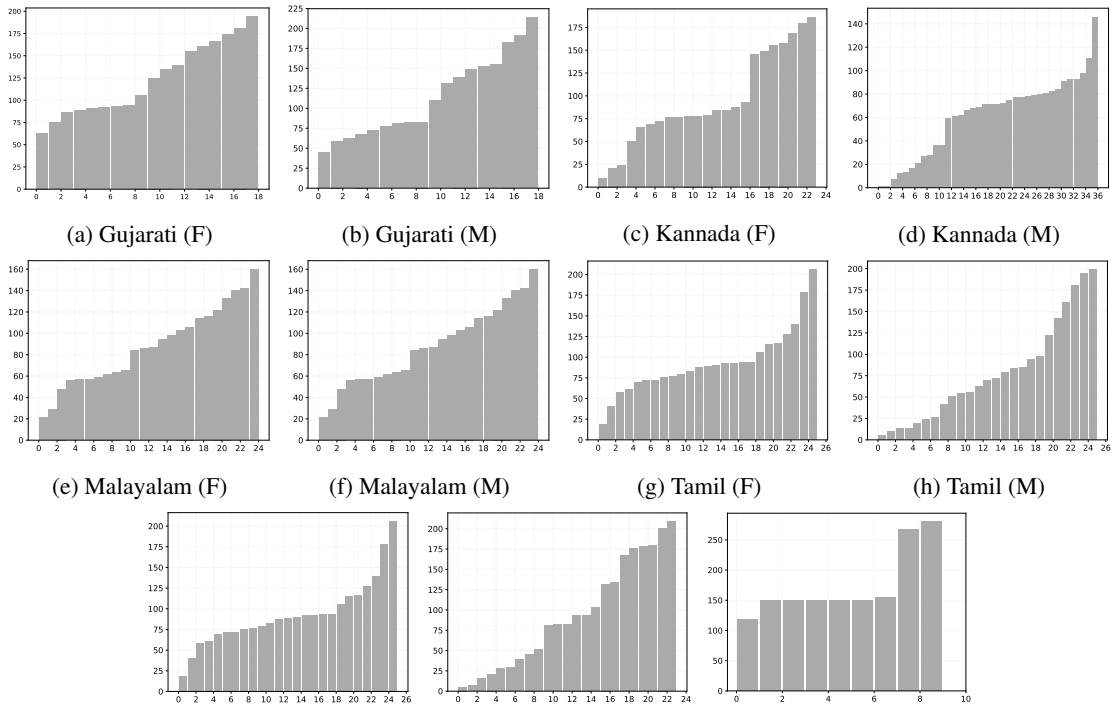
5. Recording Process

The speakers that we recorded were all volunteer participants. All the speakers were recorded at the Google offices. Using many speakers for the recording allowed us to obtain more data without putting too much burden on each volunteer, who was not a professional voice talent. Our speaker selection criteria were: (1) be a native speaker of the language with a standard accent and (2) be between 21 and 35 years of age. These criteria were adopted to be simple and make finding volunteers easy. We recorded the audio with an ASUS Zenbook UX305CA fanless laptop, a Neumann KM 184 microphone and a Blue Icicle XLR-USB A/D converter. Instead of renting an expensive studio, we simply used a portable 3x3 acoustic vocal booth. Figure 2 shows an example of our recording setup. The audio was recorded using our web-based recording software. Each speaker was assigned a number of sentences. The tool recorded each sentence at 48 kHz (16 bits per sample). We also used the in-house software for quality control where reviewers could check the recording against the recording script and provide additional comments when necessary.

A data release consent form was signed by every volunteer before each recording session. The equipment setup was designed to capture consistent volume and clear input, including keeping 30 cm mouth-to-mic distance between the volunteer and the microphone. The requirements for the position of the microphone were as follows: The microphone should point below the speaker's forehead and above their chin. The diaphragm of mic should be pointing directly at the mouth. The same distance between microphone and mouth should be kept for each recording session. We did so by marking these positions using a plastic tape.

The setup is kept identical throughout the entire recording session. Each volunteer read around 100 sentences in an hour. The volunteers were asked to speak with neutral tone and pace. They stood up during the recording and were asked to take a break every 20–30 minutes. We provided drinking water and apples for the speakers to help moisturize their mouths and to keep their voices clear. After each sentence was recorded, the volunteer played the recording to ensure that it was noise-free before continuing to the next sentence.

Since none of our speakers were professional voice talents, their recordings could contain problematic artifacts such as unexpected pauses, spurious sounds (like coughing or clearing the throat) and breathy speech. As a result, it was very important to conduct quality control (QC) of the recorded audio data. All recordings went through a quality control process performed by trained native speakers to ensure that each recording (1) matched the corresponding script (2) had consistent volume (3) was noise-free (free of background noise, mouth clicks, and breathing sounds) and (4) consisted of fluent speech without unnatural pauses or mispronunciations. The reviewers could use a QC tool to edit the transcriptions to match the recording (e.g., in the cases where the speaker skipped a word). Entries that could not be edited to meet the criteria were either re-recorded or dropped.



(a) Gujarati (F) (b) Gujarati (M) (c) Kannada (F) (d) Kannada (M)

(e) Malayalam (F) (f) Malayalam (M) (g) Tamil (F) (h) Tamil (M)

(i) Telugu (F) (j) Telugu (M) (k) Marathi (F)

Figure 3: Numbers of utterances recorded by each individual speaker arranged by language and gender (x-axis shows speaker indices, y-axis the frequency).

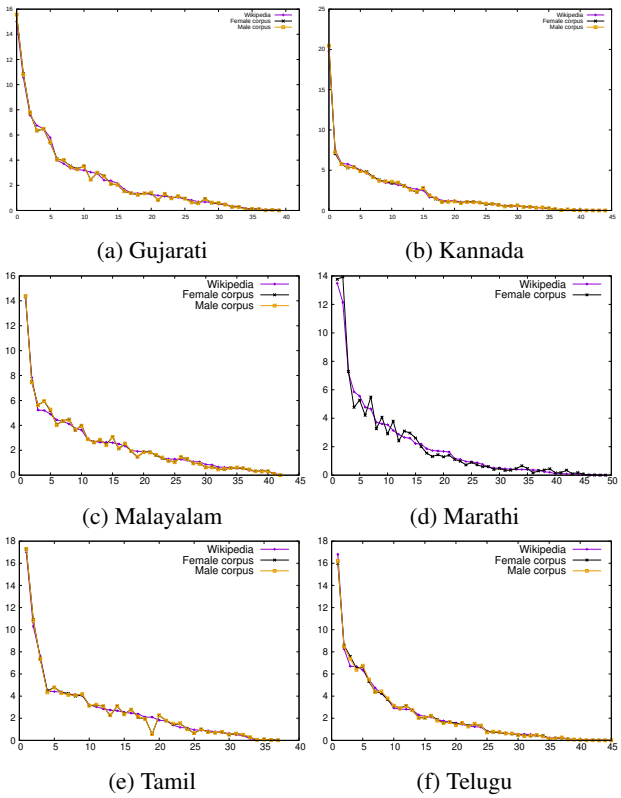
Language	Gender	Sentences	Words		Syllables		Phonemes	
			Total	Unique	Total	Unique	Total	Unique
gu	F	2,219	23,199	8,203	58,735	1,883	128,199	40
	M	2,053	21,518	7,818	54,473	1,893	119,111	40
kn	F	2,186	16,062	8,622	63,702	1,504	138,324	44
	M	2,214	14,413	7,381	57,004	1,372	123,709	44
ml	F	2,103	12,581	5,713	46,723	1,894	107,486	41
	M	2,023	12,749	6,407	47,620	1,999	109,626	41
mr	F	1,569	17,989	3,072	44,459	1,388	98,361	47
	M				–			
ta	F	2,335	15,880	6,620	56,607	1,696	126,659	37
	M	1,956	13,545	6,159	48,049	1,642	107,570	37
te	F	2,294	11,286	4,218	35,546	1,281	76,622	44
	M	2,154	11,172	4,336	34,828	1,310	75,250	44

Table 4: Various properties of the recording scripts for different languages.

6. Audio Corpora Details

The recordings of Gujarati, Kannada, Tamil, Telugu and Malayalam were done in Singapore, India (Bangalore) and the US (Mountain View), while the Marathi recordings were done in Singapore, the UK (London) and the US (Mountain View). We were able to get at least 9 volunteers for each locale. After the final round of quality control, we obtained at least 2.49 hours of verified recordings for each language. Table 3 shows the properties of the recorded audio files of each language and gender. On average an utterance is about four to seven seconds long, depending on the language and speaker's gender. Figure 3 shows the distribution of the number of sentences read by different speakers in different languages. The variation observed in the amount of read material contributed by each speaker was due to the availability of different speakers and the fact that some speakers produced a lot of bad recordings (e.g., containing a lot of breathing or mouth click artifacts).

Table 4 shows the properties of the corresponding transcriptions. Except for Marathi, all other corpora have around 2,000 sentences. They each contain a comparable number of total and unique words. Notice that, for Marathi, the number of unique words is clearly lower. This is because of the template-based sentence generation. The phonological information is displayed in terms of phonemes and the syllables. This information is based on the pronunciations derived from the phoneme inventories discussed in Section 4.1. We used in-house generated pronunciation dictionaries for each of the languages. The algorithm that we used for syllabifying the pronunciations is a standard syl-



(a) Gujarati (b) Kannada

(c) Malayalam (d) Marathi

(e) Tamil (f) Telugu

Figure 4: Phoneme distribution of each corpus with Wikipedia data as a reference. In each sub-figure, the x-axis indicates the phoneme indices and the y-axis indicates the percentages of the phonemes as they appear in the corpus.

labification approach based on the Maximum Onset Principle (Bartlett et al., 2009).

The numbers of unique phonemes reported in Table 4 differ from those in Table 2 for Kannada, Malayalam, Marathi, and Telugu. This is because some of the phonemes defined in the phoneme inventories for these languages are rarely attested in our data. For example, for Marathi, the breathy alveolar flap/trill /r$^{\text{h}}$/ and breathy alveolo-palatal approximant /j$^{\text{h}}$/ are missing from our corpus. Only 0.002% of phonemes derived from Wikipedia text are /r$^{\text{h}}$/, and /j$^{\text{h}}$/ does not appear at all. While there is an ongoing debate in the Marathi linguistic literature on how to precisely represent these sounds (Berkson, 2013), they do indeed happen to be very rare (Berkson and Nelson, 2017).

As mentioned earlier, we computed the phoneme distributions of the corpora to make sure that they were consistent with the natural distribution in the languages. We used Wikipedia texts of the respective languages as a reference. Figure 4 compares the phoneme distribution of each corpus against that of the text extracted from Wikipedia. In each plot, the phonemes are ordered in descending order by their frequencies in the text of Wikipedia. The plots show that the phoneme distributions in our corpora for both genders closely follow those of Wikipedia. The phoneme distribution for the Marathi corpus shows the most difference against Wikipedia's. This was because the majority of the sentences were generated from templates. However, overall, the difference was still relatively minor and was most pronounced only on a few phonemes.

Status	Source	Languages
Open-Source	This paper	Gujarati, Kannada, Malayalam, **Marathi**, Tamil, Telugu
Open-Source	Sodimana et al. (2018)	Nepali, Sinhala
Proprietary	Internal	Bengali (Bangladesh), Bengali (India), Hindi, Urdu (Pakistan)

Table 5: Structure of the training set used to train Marathi system. Eight open-source datasets where used.

6.1. Quality Evaluation of the Corpora

Combined Training Set Demirsahin et al. (2018) used the same corpora presented here (except for Marathi) to build multilingual TTS voices. In order to construct a multilingual acoustic model they combined the data for Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu with other South Asian languages, namely, Bangladeshi and Indian dialects of Bengali, Hindi, Nepali, Sinhala and Pakistani Urdu. The quality of the voices was reported in their work. It is important to note that the Marathi corpus investigated by Demirsahin et al. (2018) is proprietary and single-speaker, which is very different from our free Marathi multi-speaker corpus. Therefore, we built a Marathi voice that is based on the multilingual acoustic model trained using all the South Asian languages described by Demirsahin et al. (2018), but with the original proprietary Marathi single-speaker data replaced by our multi-speaker Marathi corpus. In other words, we trained our acoustic model on the training set constructed by pooling the data from all the available twelve datasets shown in Table 5. The second column in table, called "Source", refers to the source of documentation for the corpora. Eight out of twelve datasets that we used are in public domain. The open-source portion of the training set that we describe in this paper (six languages) is shown in blue in the first row of the table. Additional open-source multi-speaker datasets include Nepali[1] and Sinhala[2], which we released previously (Sodimana et al., 2018). The third row shows the proprietary portion of the training set consisting of two dialects of Bengali, Hindi and Urdu. It is worth noting that we previously released multi-speaker Bangladeshi and Indian Bengali datasets into public domain[3]. These open-sourced datasets are different from the proprietary Bengali corpora used in this work.

Model Details Neural network approach identical to the one described in (Gutkin, 2017; Demirsahin et al., 2018; Wibawa et al., 2018) was used. Briefly, we used long short term memory recurrent neural network (LSTM-RNN)

[1]Available from `http://www.openslr.org/43/`. ISLRN: `768-733-837-923-2`

[2]Available from `http://www.openslr.org/30/`. ISLRN: `182-897-524-187-4`

[3]Available from `http://www.openslr.org/37/`. ISLRN: `527-627-691-135-2`

Language	Female	Male
Gujarati	**3.950**±0.056	**4.269**±0.047
Kannada	**4.104**±0.063	**4.236**±0.059
Malayalam	**4.032**±0.118	**4.143**±0.108
Marathi	**4.065**±0.071	–
Tamil	**3.623**±0.103	**3.606**±0.095
Telugu	**4.152**±0.075	**3.793**±0.081

Table 6: Mean opinion scores (MOS) of the voices shown along with the 95% confidence intervals.

acoustic model configuration originally proposed by Zen and Sak (2015): Two unidirectional LSTM-RNNs for duration and acoustic parameter prediction are used in tandem in a streaming fashion. Given the input linguistic features, the goal of the duration LSTM-RNN is to predict the duration (in frames) of the phoneme in question. This prediction, together with the input features, is then provided to the acoustic model which predicts smooth acoustic vocoder parameter trajectories. The smoothing of transitions between consecutive acoustic frames is achieved in the acoustic model by using recurrent units in the output layer.

The input features used by both the duration and the acoustic models consist of one-hot linguistic features that describe the utterance including the phonemes, syllable counts, distinctive phonological features (such as manner of articulation) and so on. The one-hot features also include language and region encoded as two separate features. These features allow us to guide the acoustic model when the output for a particular language and region is desired at synthesis time. An additional important feature that we use is a one-hot speaker identity feature. When using a model trained on multiple speakers, this feature is instrumental in forcing the consistent speaker characteristics on the output of the model. In other words, it forces the voice to sound like the requested speaker.

Evaluation Results and Discussion We asked native speakers to identify the best speaker. During training, we used speaker IDs as an input feature. Then, during synthesis, we conditioned the speaker ID feature to be the best speaker. The voice was evaluated by the raters using a Mean Opinion Score (MOS) listening test (Streijl et al., 2016). For the test, we synthesized the audio for the 100 sentences unseen in the training data. Twelve native speakers were asked to rate each utterance on a 5-point scale (1: worst, 5: best). Table 6 shows the MOS results for our voices. Each mean opinion score (shown in bold) is shown along with the corresponding confidence interval statistics at 95% confidence level (Wonnacott and Wonnacott, 1990) computed using the recommendations in (Recommendation ITU-T P.1401, 2012). We included the scores for Gujarati, Kannada, Malayalam, Tamil and Telugu previously reported by Demirsahin et al. (2018) for completeness. The results for Marathi are shown in blue. These results show that we can build good quality voices using our corpora. For all the languages except Tamil some (or all) gender configurations achieved a high MOS of over 4.0 (in comparison, we consider MOS scores over 3.5 as reasonable).

7. Conclusions and Future Work

In this paper, we presented high quality open-source multi-speaker speech corpora for six official languages of India: Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu. The corpora has been designed with multilingual TTS applications in mind. We presented the details of the process used to construct the corpora. This process was designed to be practical for collecting data in low-resource scenarios, where limited linguistic or financial resources are often available. We show that a good-quality multilingual acoustic model can be trained by pooling the data from the six languages in question. The corpora are released with an open-source license with no limitations on academic or commercial use. We hope this data will contribute to research and development of speech applications for these six important languages of India.

In the future, tools such as the one described by Podsiadlo and Ungureanu (2018) could be used to streamline the recording script design process. Additional venues to explore include collecting high-quality data for other low-resource languages of India, such as Sindhi (Cole, 2006), which can be used in transfer learning scenarios, where a reasonably small amount of data from a target language is used as adaptation data. This will facilitate our efforts to expand the data for existing languages and release text-to-speech corpora for additional languages of India and beyond.

8. Acknowledgments

We would like to thank Richa Singh, Trisha Banerjee, Alena Butryna, Jaka Aris Eko Wibawa, Chenfang Li and Archana Amberkar for their help with the data collection and voice evaluation. The authors thank Rob Clark and Richard Sproat for their suggestions to the earlier drafts of this paper. Finally, the authors thank the anonymous reviewers for many useful suggestions.

9. Bibliographical References

Achanta, S. (2018). *Multilingual Text-to-Speech Synthesis using Sequence-to-Sequence Neural Networks*. Ph.D. thesis, International Institute of Information Technology, Hyderabad, India.

Baby, A., Thomas, A., N L, N., and Consortium, T. (2016). Resources for Indian languages. In *Proceeding of the 19th International Conference on Text, Speech and Dialogue*, pages 37–43, 09.

Baker, P., Hardie, A., McEnery, T., and Jayaram, B. (2003). Corpus data for South Asian language processing. In *Proceeding of the 10th Annual Workshop for South Asian Languages Processing, EACL*, pages 1–8.

Baljekar, P., Rallabandi, S., and Black, A. W. (2018). An Investigation of Convolution Attention Based Models for Multilingual Speech Synthesis of Indian Languages. In *Proc. Interspeech 2018*, pages 2474–2478, Hyderabad, India.

Baljekar, P. (2018). *Speech Synthesis from Found Data*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Bartlett, S., Kondrak, G., and Cherry, C. (2009). On the Syllabification of Phonemes. In *NAACL'09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 308–316, Boulder, Colorado. ACL.

Berkson, K. H. and Nelson, M. (2017). Phonotactic frequencies in Marathi. *IULC Working Papers*, 17.

Berkson, K. H. (2013). *Phonation types in Marathi: An acoustic investigation*. Ph.D. thesis, University of Kansas.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Black, A. W. (2019). CMU Wilderness Multilingual Speech Dataset. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975. IEEE.

Cardona, G. and Jain, D. (2007). *The Indo-Aryan Languages*. Routledge Language Family Series. Routledge.

Chen, Y.-J., Tu, T., Yeh, C., and Lee, H.-Y. (2019). End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning. In *Proc. Interspeech 2019*, pages 2075–2079, Graz, Austria.

Cole, J. S. (2006). Sindhi. In *Encyclopedia of Language & Linguistics*, pages 384–387. Elsevier.

Cooper, E. L. (2019). *Text-to-Speech Synthesis Using Found Data for Low-Resource Languages*. Ph.D. thesis, Columbia University, New York.

Creative Commons. (2019). Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). `http://creativecommons.org/licenses/by-sa/4.0/deed.en`.

Demirsahin, I., Jansche, M., and Gutkin, A. (2018). A Unified Phonological Representation of South Asian Languages for Multilingual Text-to-Speech. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 80–84.

Dhongde, R. V. and Wali, K. (2009). *Marathi*, volume 13. John Benjamins Publishing.

Dijkstra, J. (2004). FRYSS: A First Step Towards Frisian TTS. Technical Report 142, Institute of Phonetic Sciences, University of Amsterdam.

Emeneau, M. (1956). India as a Linguistic Area. *Language*, 32(1):3–16.

Enfield, N. J. and Comrie, B. (2015). *Languages of Mainland Southeast Asia. The state of the Art*. Pacific Linguistics. De Gruyter Mouton.

Gibiansky, A., Arik, S., Diamos, G., Miller, J., Peng, K., Ping, W., Raiman, J., and Zhou, Y. (2017). Deep Voice 2: Multi-Speaker Neural Text-to-Speech. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2962–2970.

Gutkin, A., Ha, L., Jansche, M., Pipatsrisawat, K., and Sproat, R. (2016). TTS for Low Resource Languages: A Bangla Synthesizer. In *10th edition of the Language Resources and Evaluation Conference (LREC)*, pages 2005–2010, Portorož, Slovenia, May.

Gutkin, A. (2017). Uniform Multilingual Multi-Speaker Acoustic Model for Statistical Parametric Speech Synthesis of Low-Resourced Languages. In *Proc. of Interspeech 2017*, pages 2183–2187, August 20–24, Stockholm, Sweden.

Kjartansson, O., Sarin, S., Pipatsrisawat, K., Jansche, M., and Ha, L. (2018). Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 52–55.

Li, B. and Zen, H. (2016). Multi-Language Multi-Speaker Acoustic Modeling for LSTM-RNN Based Statistical Parametric Speech Synthesis. In *INTERSPEECH*, pages 2468–2472.

Mahesh, M., Prakash, J. J., and Murthy, H. A. (2018). Resyllabification in Indian languages and its Implications in Text-to-speech Systems. In *Proc. of Interspeech 2018*, pages 212–216, Hyderabad, India, September.

Mapelli, V., Popescu, V., Liu, L., and Choukri, K. (2016). Language Resource Citation: the ISLRN Dissemination and Further Developments. In *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1610–1613, Portorož, Slovenia, May. ELRA.

Mohanty, A. K. (2006). Multilingualism of the Unequals and Predicaments of Education in India: Mother tongue or other tongue. *Imagining multilingual schools*, pages 262–283.

Mohanty, A. K. (2010). Languages, inequality and marginalization: Implications of the double divide in Indian multilingualism. *International Journal of the Sociology of Language*, 2010(205):131–154.

Nachmani, E. and Wolf, L. (2019). Unsupervised Polyglot Text-to-Speech. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7055–7059. IEEE.

Ni, J., Hirai, T., Kawai, H., Toda, T., Tokuda, K., Tsuzaki, M., Sakai, S., Maia, R., and Nakamura, S. (2007). ATRECSS: ATR English Speech Corpus for Speech Synthesis. In *Proc. of Blizzard Challenge 2007 Workshop*, pages 1–4, Bonn, Germany.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Pattanayak, D. P. (1969). Oriya and Assamese. In Murray B. Emeneau et al., editors, *Linguistics in South Asia*, pages 122–152. De Gruyter Mouton.

Phillips, A. and Davis, M. (2009). BCP 47 – Tags for Identifying Languages. *IETF Trust*.

Pitrelli, J. F., Bakis, R., Eide, E. M., Fernandez, R., Hamza, W., and Picheny, M. A. (2006). The IBM expressive text-to-speech synthesis system for American English. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1099–1108.

Podsiadlo, M. and Ungureanu, V. (2018). Experiments with Training Corpora for Statistical Text-to-Speech Systems. In *Interspeech*, pages 2002–2006.

Povey, D. (2019). Open SLR. http://www.openslr.org/resources.php. Accessed: 2019-03-30.

Prahallad, K., Elluru, N. K., Keri, V., Rajendran, S., and Black, A. W. (2012). The IIIT-H Indic Speech Databases. In *INTERSPEECH*, Portland, Oregon, USA.

Prahallad, K., Vadapalli, A., Elluru, N., Mantena, G., Pulugundla, B., Bhaskararao, P., Murthy, H., King, S., Karaiskos, V., and Black, A. (2013). The Blizzard Challenge 2013 – Indian Language Tasks. In *Proc. Blizzard Challenge Workshop*.

Prakash, A., Thomas, A. L., Umesh, S., and Murthy, H. A. (2019). Building Multilingual End-to-End Speech Synthesisers for Indian Languages. In *Proc. of 10th ISCA Speech Synthesis Workshop (SSW'10)*, pages 194–199, Vienna, Austria.

Rallabandi, S. K. and Black, A. W. (2017). On Building Mixed Lingual Speech Synthesis Systems. In *Proc. of Interspeech 2017*, pages 52–56, Stockholm, Sweden.

Recommendation ITU-T P.1401. (2012). Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. *International Telecommunication Union*, July.

Shrishrimal, P. P., Deshmukh, R. R., and Waghmare, V. B. (2012). Indian Language Speech Database: A Review. *International Journal of Computer Applications*, 47(5):17–21, June.

SIL International. (2019). Ethnologue. https://www.ethnologue.com. Accessed: 2019-03-25.

Sodimana, K., Silva, P. D., Sarin, S., Pipatsrisawat, K., Kjartansson, O., Jansche, M., and Ha, L. (2018). A Step-by-Step Process for Building TTS Voices Using Open Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 66–70.

Sonobe, R., Takamichi, S., and Saruwatari, H. (2017). JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis. *arXiv preprint arXiv:1711.00354*.

Srivastava, B. M. L., Sitaram, S., Mehta, R. K., Mohan, K. D., Matani, P., Satpal, S., Bali, K., Srikanth, R., and Nayak, N. (2018). Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages. In *6th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2018)*, pages 11–14, Gurugram, India, August.

Steever, S. B. (1997). *The Dravidian Languages*. Routledge Language Family Series. Routledge.

Streijl, R. C., Winkler, S., and Hands, D. S. (2016). Mean Opinion Score (MOS) revisited: Methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227.

van Niekerk, D., van Heerden, C., Davel, M., Kleynhans, N., Kjartansson, O., Jansche, M., and Ha, L. (2017). Rapid development of TTS corpora for four South African languages. In *Proc. Interspeech 2017*, pages 2178–2182.

Varghese, B., John, M., and Samuel, N. (2009). The Malvi-speaking people of Madhya Pradesh and Rajasthan: a sociolinguistic profile. *SIL Electronic Survey Report*.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. (2017). Tacotron: A fully end-to-end text-to-speech synthesis model. *arXiv preprint arXiv:1703.10135*.

Watkins, J. W. (2001). Burmese. *Journal of the International Phonetic Association*, 31(2):291–295.

Wibawa, J. A. E., Sarin, S., Li, C. F., Pipatsrisawat, K., Sodimana, K., Kjartansson, O., Gutkin, A., Jansche, M., and Ha, L. (2018). 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)*, pages 1610–1614, 7-12 May 2018, Miyazaki, Japan.

Wilkinson, A., Parlikar, A., Sitaram, S., White, T., Black, A. W., and Bazaj, S. (2016). Open-Source Consumer-Grade Indic Text To Speech. In *9th ISCA Speech Synthesis Workshop*, pages 190–195, Sunnyvale, USA, September.

Wonnacott, T. H. and Wonnacott, R. J. (1990). *Introductory Statistics*, volume 5. Wiley.

Zanon, B. M., Anastasopoulos, A., Villavicencio, A., Besacier, L., and Lekakou, M. (2018). A Small Griko-Italian Speech Translation Corpus. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 36–41, Gurugram, India, August.

Zen, H. and Sak, H. (2015). Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pages 4470–4474, Brisbane, Australia, April. IEEE.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. (2019). LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proc. of Interspeech 2019*, pages 1526–1530, Graz, Austria.

10. Language Resource References

Google. (2019a). *Crowd-sourced high-quality Gujarati multi-speaker speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), http://www.openslr.org/78, Google crowd-sourced resources, 1.0, ISLRN 276-159-489-933-8.

Google. (2019b). *Crowd-sourced high-quality Kannada multi-speaker speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), http://www.openslr.org/79, Google crowd-sourced resources, 1.0, ISLRN 494-932-368-282-1.

Google. (2019c). *Crowd-sourced high-quality Malayalam multi-speaker speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), http://www.openslr.org/63, Google crowd-sourced resources, 1.0, ISLRN 246-208-077-317-5.

Google. (2019d). *Crowd-sourced high-quality Marathi multi-speaker speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), http://www.openslr.org/64, Google crowd-sourced resources, 1.0, ISLRN 498-608-735-968-0.

Google. (2019e). *Crowd-sourced high-quality Tamil multi-speaker speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), http://www.openslr.org/65, Google crowd-sourced resources, 1.0, ISLRN 766-495-250-710-3.

Google. (2019f). *Crowd-sourced high-quality Telugu multi-speaker speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), http://www.openslr.org/66, Google crowd-sourced resources, 1.0, ISLRN 598-683-912-457-2.