IndicSpeech: Text-to-Speech Corpus for Indian Languages

by

Nimisha Srivastava, Rudrabha Mukhopadhyay, K R Prajwal, C V Jawahar

in

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020),
: 1
-6

Report No: IIIT/TR/2020/-1

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2020

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6417–6422
Marseille, 11–16 May 2020

IndicSpeech: Text-to-Speech Corpus for Indian Languages

Nimisha Srivastava, Rudrabha Mukhopadhyay*, K R Prajwal*, C V Jawahar
IIIT Hyderabad
{radrabha.m, prajwal.k}@research.iiit.ac.in
{nimisha.srivastava, jawahar}@iiit.ac.in

Abstract

India is a country where several tens of languages are spoken by over a billion strong population. Text-to-speech systems for such languages will thus be extremely beneficial for wide-spread content creation and accessibility. Despite this, the current TTS systems for even the most popular Indian languages fall short of the contemporary state-of-the-art systems for English, Chinese, etc. We believe that one of the major reasons for this is the lack of large, publicly available text-to-speech corpora in these languages that are suitable for training neural text-to-speech systems. To mitigate this, we release a 24 hour text-to-speech corpus for 3 major Indian languages namely Hindi, Malayalam and Bengali. In this work, we also train a state-of-the-art TTS system for each of these languages and report their performances. The collected corpus, code, and trained models are made publicly available.

Keywords: Text-to-speech, Indian languages, TTS corpus

1. Introduction

India has always been known as a country with great diversity. It houses numerous races, castes, creeds, and religions. It is also a pluri-lingual country with over 19,500 languages or dialects being spoken as a mother tongue (Chandramouli and General, 2011). Even if we only consider the official languages as declared by the Indian constitution (Chandramouli and General, 2011), we have 22 languages, each of them being spoken by over a million people. Over the last decade, the enormous penetration of the Internet in the country means that a major chunk of this multi-lingual population is actively seeking to consume and interact with online content in their own languages (IMAI, 2017). However, the Internet today is severely deficit in terms of content for Indian local languages (IMAI, 2017). In fact, 53% of non-Internet owners in India state that they will start using the Internet if it has content available in their local languages (IMAI, 2017). The lack of local language content is particularly stark in the multimedia domain that consists of videos, podcasts and digital assistants. This can be attributed to the lack of localization of cutting-edge technologies like optical character recognition, neural machine translation and text-to-speech systems. While significant efforts are being made in the former two areas (Mathew et al., 2016; Philip et al., 2019a; Philip et al., 2019b), the progress in developing reliable text-to-speech systems for Indian languages has been relatively much slower.

One of the major hindrances to adopt existing state-of-the-art neural text-to-speech systems (Ping et al., 2017; Wang et al., 2017) for Indian languages, we believe, is the lack of large, reliable text-to-speech corpora in these languages. The largest publicly available resource available for training text-to-speech systems is the IndicTTS (Baby et al., 2016) corpus that contains about 8 hours of speech data for 13 Indian languages. While this corpus has been very useful for TTS systems in the past, it is insufficient to train modern neural TTS models that typically require over 24 hours of speech data (Ito, 2017) to be able to generate natural, accurate speech.

To this end, we release *IndicSpeech*, a large text-to-speech corpus for multiple Indian languages with about 24 hours of single-speaker speech data each. So far, the corpus has been curated for three languages: (i) Hindi, (ii) Malayalam, (iii) Bengali and a few more languages are also being collected. All the collected languages is released publicly along with this paper.

With the help of the *IndicSpeech* corpus, we explore the feasibility of state-of-the-art neural TTS systems (Ping et al., 2017) for multiple Indian languages. We demonstrate that we are able to generate natural speech for these languages. We hope that this work encourages subsequent efforts in this area to address some of the practical issues that arise when training TTS systems for Indian languages that are morphologically rich and agglutinative. In summary, our key contributions are:

- We release a text-to-speech corpus for multiple Indian languages. The IndicSpeech corpus is about $4\times$ larger than current corpora, allowing us to train neural text-to-speech systems.
- We adapt and train a state-of-the-art neural text-to-speech model to achieve high-quality speech in these languages.
- The code, trained models, and a live demo will be made available publicly.

The data and code are available at this link [1]

2. Related Works

In this section we discuss about current state-of-the-art text-to-speech systems and the popular corpora that are used to train them in various languages.

* Equal contribution

[1] http://cvit.iiit.ac.in/research/projects/cvit-projects/text-to-speech-dataset-for-indian-languages

Figure 1: We curate a text to speech corpus for three languages. Each language contains about 25 hours of high quality speech data spanning a rich vocabulary of over $11k+$ words. The most popular words in Hindi, Malayalam, and Bengali are depicted in the word-clouds.

In recent years, neural network based text-to-speech systems have garnered a lot of attention in the speech community. Works like Tacotron (Wang et al., 2017),Tacotron 2 (Shen et al., 2017), Deep Voice 3 (Ping et al., 2017) are capable of producing high quality natural speech. However, all of these methods are data hungry and require approximately 24 hrs of text-to-speech data for a single speaker. Tacotron (Wang et al., 2017) and Tacotron 2 (Shen et al., 2017) were originally trained on US English dataset (≈ 24.6 hours) which is not available publicly. Deep Voice 3 was also originally trained on a non-public dataset (≈ 20 hours). Public implementations of all of the three models have been trained on the LJSpeech dataset (Ito, 2017) and are able to produce high quality natural speech. Some of these networks were also trained on datasets of comparable size for other languages like Korean (Park, 2019), Japanese (Sonobe et al., 2017) etc.

One of the first efforts towards collecting a text-to-speech dataset for Indian languages was made in (Prahallad et al., 2012). Later a much larger corpus for Indian languages was developed in (Baby et al., 2016). This corpus was used to train text-to-speech systems for 13 languages were developed in (Pradhan et al., 2015). However, in this corpus, the amount of data provided per language is far too less ($\approx 25\%$ of recent TTS datasets) for training recent neural network based systems that can produce natural, accurate speech. In this work, we collect the first large scale text-to-speech corpus for multiple Indian languages aimed at training recent TTS systems like Tacotron 2 (Shen et al., 2017), Deep Voice 3 (Ping et al., 2017). In the next section, we describe the newly created IndicSpeech corpus in detail.

3. The IndicSpeech Corpus

In this section, we describe the efforts taken to curate the IndicSpeech corpus. We start with Hindi, as it is the most popular language in India, and second with Malayalam, as we decided to include a popular Dravidian language in our corpus as well. We also include a third language Bengali in our corpus. Note that the data collection steps are same for all the languages. We discuss how we collect the text sentences, the selection process for the speakers and details of the recording setup.

3.1. Text collection

We collect the sentences in multiple Indian languages from online newspapers. Newspapers contain texts from a wide variety of domains ranging from politics to sports and entertainment. We normalize the text before asking the speaker to read. In this normalization step, we replace all the numbers with words, abbreviations to their full forms etc. We then record speech corresponding to these normalized sentences. We do not choose very long sentences which need greater than 15 seconds to be uttered, as such sentences can make the training process inefficient. We also do not choose short sentences that are less than 3 words. These design choices are also made in the LJSpeech dataset (Ito, 2017). Additionally, we use the Indic NLP Library (Kunchukuttan, 2013) to transliterate the texts to English characters.

3.2. Speaker selection

Once we collect the text data and organize them into sentences, the next step is to record the audio corresponding to each of the textual samples. For each of the languages, we choose a single native speaker with an amicable voice. The native speaker is a female for Hindi and Malayalam. For Bengali we choose a male speaker. All the speakers are proficient in their native languages.

3.3. Recording Setup

The recording is done using professional-grade microphones and each recording is manually verified for any errors in the text or speech. The sentence is read in one go, and the speaker is instructed to articulate the sentence as naturally as possible. Some sentences are re-recorded if any errors or unnatural speech is observed such as sudden change in pitch, pronunciation errors or abnormal pauses. As the task is quite laborious, we give regular breaks to the speaker in order to minimize any deterioration in speech quality due to fatigue. Recording about 26 hours of data takes about three weeks to complete.

3.4. Corpus Statistics

We plot the text word cloud of our corpus in Figure 1 and also report a few statistics in Table 1. We compare our corpus with other publicly available resources in Indian languages for text-to-speech synthesis in Table 2.
From both Table 1 and Table 2, it is evident that our corpus is significantly larger than current existing corpora for

Language	#Hours	Vocab size	#Sentences	Mean word freq.	Mean audio length	Mean #words / sample
Hindi	25.6	11234	9916	12.6	7.2 sec	18.3
Malayalam	29.1	41216	19954	8.5	6.8 sec	10.3
Bengali	22.3	20161	12176	9.9	7.6 sec	12.6

Table 1: Descriptive statistics of our IndicSpeech corpus. We see that the corpus consists of a diverse vocabulary and is at a scale well-suited for state-of-the-art neural TTS models.

Language	Corpus	#Hours	#Total Words	#Sentences
Hindi	(Baby et al., 2016)	7.5	15153	5240
	Ours	**25.6**	**105115**	**9916**
Malayalam	(Baby et al., 2016)	8.77	13738	5132
	Ours	**29.1**	**109245**	**19954**
Bengali	(Baby et al., 2016)	10.03	12901	5316
	Ours	**22.3**	**104891**	**12176**
English	LJSpeech (Ito, 2017)	23.3	225715	13100

Table 2: Comparison of current TTS corpora for Indian languages with our proposed IndicSpeech corpus. We also compare with an English TTS dataset which is widely used to train state-of-the-art single speaker neural text-to-speech models. As we can see our corpus is over 4 times larger in terms of number of hours than previous corpora for the same languages.

Indian languages and can be used to train recent neural network based text-to-speech synthesis models. It is also worth noting that the Malayalam corpus contains significantly more vocabulary, with lesser mean word frequency. Thus, we argue that this language would pose a bigger challenge for the TTS models. In the next section, we show the performance of Deep Voice 3 (Ping et al., 2017) model when trained on our newly collected corpus.

4. Experiments

We train Deep Voice 3 (Ping et al., 2017) for three languages, Hindi, Malayalam and Bengali from our corpus. We discuss the model architecture, training methodology and results in this section. The models for all the collected languages will be trained and released publicly.

4.1. Network Architecture

We adopt the NYANKO-BUILD[2] implementation of Deep Voice 3 for training our TTS model. The problem is formulated as the standard sequence-to-sequence learning paradigm. The block diagram of the architecture is depicted in Figure 2.

The architecture of Deep Voice 3 (Ping et al., 2017) consists of a text encoder and a speech decoder. The text encoder takes a sequence of characters as input. The character embeddings of these characters are then concatenated together to form a $N \times D$ embedding matrix. Here, N is the number of characters and D is the dimension of the learnable character embedding. This matrix is then passed through multiple Highway 1D convolutions. The output from the text encoder is passed through an attention based speech decoder. The speech decoder generates a melspectrogram as output. The encoder and decoder are trained end-to-end. We train the network by minimizing the standard L1 loss between the generated melspectrogram and the ground truth melspectrogram. Finally, during inference we use Griffin-Lim (Griffin and Jae Lim, 1984) to convert the generated melspectrogram to a raw waveform. For more information about the architecture and the training methodology, we refer the reader to (Ping et al., 2017) and the open-source implementation[2].

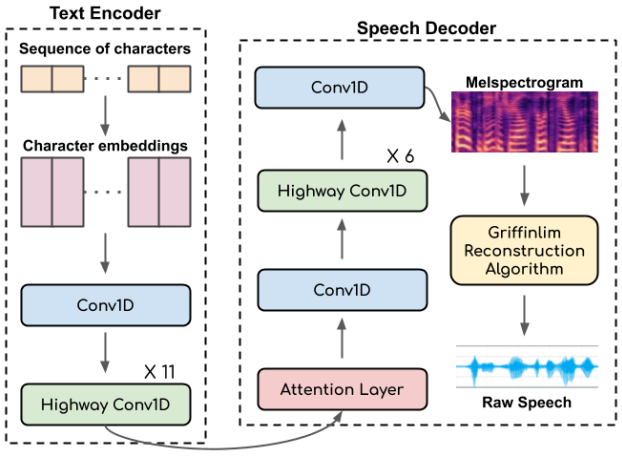

Figure 2: We use the NYANKO BUILD implementation of Deep Voice 3 as our base model. The model has a text encoder which takes sequence of characters as input. The speech decoder is used to generate melspectrograms. Griffin-Lim algorithm is then used to convert generated melspectrograms to raw audio.

4.2. Results

We evaluate the performance of the TTS trained for different languages with both subjective and objective human evaluation. We create a neutral test set for each language for this purpose. Our test set for each language contains 100 sentences which were not seen by the network during training.

4.2.1. Objective Human Evaluation

As stated in (Ping et al., 2017), attention based neural TTS systems often runs into three types of errors. These are -

[2] github.com/r9y9/deepvoice3_pytorch

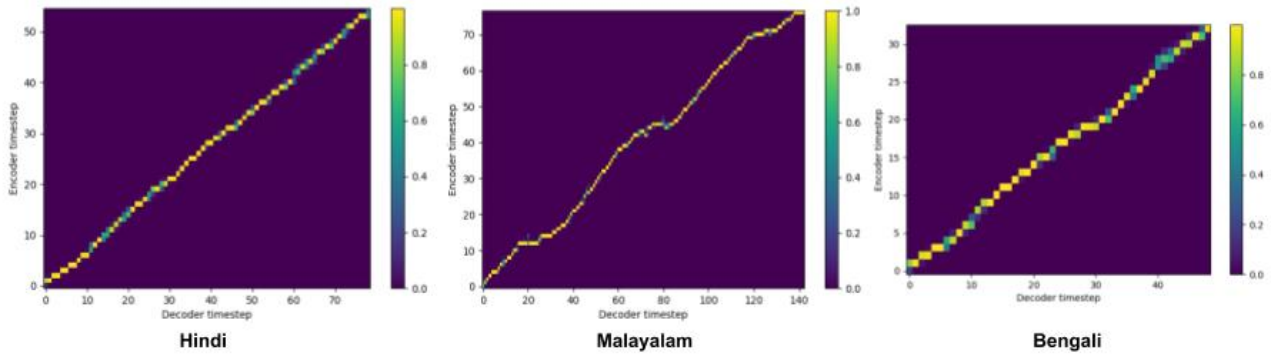

Figure 3: We provide the attention-alignment curve for Hindi, Malayalam and Bengali TTS for a sample test sentence. It is evident from all of the plots that the decoder roughly attends to the correct encoder time-step while generating output for a particular decoder time-step.

(i) repeated words, (ii) mispronunciations and (iii) skipped words. The participants listen to the test set predictions in each language and manually calculate the occurrence of each of the aforementioned errors. The statistics of all of these three errors are given in Table 3.

Error Type	Hindi	Malayalam	Bengali
Repeated words	4	10	6
Mispronunciations	11	18	14
Skipped words	1	6	3

Table 3: Scores for objective human evaluations. The participants were asked to identify the number of repeated words, mispronunciations and skipped words. One or more repeats, mispronunciations and skips count as a single mistake per utterance.

Model	Hindi	Malayalam	Bengali
Public API	2.98	3.32	3.63
Ours	**4.31**	**3.87**	**3.96**

Table 4: Average Mean-Opinion-Scores (MOS) for both Hindi, Malayalam and Bengali. The participants rate each of the speech outputs from both the models for each of the languages. The minimum score that can be given to an output is 1 and the maximum is 5. In our experiment, higher the score better the output from a model in terms on naturalness. (Murthy and others, 2016) has been used for the publicly available API.

4.2.2. Subjective Human Evaluation

We demonstrate the naturalness of our TTS system compared to current publicly available TTS systems for Indian languages (Murthy and others, 2016). We take inference for all the sentences present in our test set for all three languages using the API provided in (Murthy and others, 2016). For the human evaluation, 20 users assign scores ranging from 1 to 5 for each of the outputs from our model as well as from the publicly available Indian TTS system. In this experiment, higher score means more natural human-like speech. The average Mean-Opinion-Score (MOS) in terms of naturalness for both the languages are reported in Table 4. We choose a different set of 20 users for each language. All of the chosen users were proficient in their respective languages.

As we see from Table 4, Deep Voice 3 trained on our proposed corpus shows improvement to the current public Indian TTS systems. We also observe that the achieved scores for Malayalam is lower compared to the others. We attribute this to multiple issues as also noted in (Sankaran and Jawahar, 2013). Malayalam, being an inflectional language has a much more larger vocabulary and more characters per word. Further, it also has numerous morpho-phonemic changes during word formation. One of the solutions would be to increase the size of the Malayalam corpus to cover more vocabulary and also increase the word frequency. We believe overcoming these challenges for various Indian languages will be an interesting avenue for research.

4.2.3. Attention-Alignment Curves

Finally, in Figure 3 we also provide a sample attention-alignment curve for a sentence from the test set in both languages. The attention-alignment curve helps us visualize the encoder time-step the decoder attends while itself generating an output for a certain time-step. Ideally, the decoder while decoding a certain time-step should attend to a very similar time-step of the encoder. Hence, in such a situation the curve should be close to a $y = x$ line. It can also be seen that the alignment curve shape is inferior in the case of Malayalam, and this is also reflected in the lower MOS scores for Malayalam. On the other hand, Hindi and Bengali has near perfect alignment curves which corresponds to the higher MOS scores that we get for these languages.

5. Application of Text-to-Speech Systems in Indian Scenario

The general set of applications for TTS systems are widely applicable to Indian languages as well. This includes localized digital assistants, improved accessibility to the visually impaired to listen to text in the languages they are comfortable with and so on. Specifically for the visually impaired, the local language books can now be converted to a large audio book library for their consumption. This is very important given that large sections of the popula-

tion in India prefer to read in local languages (Youth, 2010) rather than in English. Similarly, the visually impaired can now also browse the web as well by listening to the text displayed in their local languages. Moreover, TTS models are an essential extension to produce speech from the text recognized by the image captioning systems (Anderson et al., 2018; Prajwal et al., 2019). In a country where several tens of languages are spoken by millions of people, content can quickly become inaccessible due to language barriers. TTS models form an essential element for speech-to-speech translation (Jia et al., 2019), speech-to-text translation (Weiss et al., 2017) and face-to-face translation (KR et al., 2019) pipelines. We hope that our corpus facilitates rapid development of TTS systems in Indian languages, thus enabling the aforementioned wide range of applications.

6. Ongoing Efforts

We are in the process of collecting data for more Indian languages. In the future we also aim to create large-scale multi-speaker text-to-speech corpus for some of the most popular Indian languages, similar to the English counterparts (Zen et al., 2019). This would enable training of multi-speaker TTS systems that can generate speech in the voice of any Indian speaker. In addition to data collection, we are also exploring various techniques that can allow us to scale TTS systems to the several tens of major Indian languages. One of the major directions is to reduce the need of speaker-specific data to train a neural TTS system. If we can train a neural model, with say, about two hours of data, then it becomes much easier to expand to numerous languages, accents, voices and dialects. Further, as many languages share an overlapping phoneme space, it would be interesting to study the transfer of learning across languages and speakers. We request the readers to check our website for updates along these directions.

7. Conclusion

In this work, we present a novel corpus in Indian languages for the task of neural text-to-speech synthesis. Our corpus contains three major Indian languages which are widely popular across the country. Our corpus is around $4\times$ times larger than the current corpora available publicly for the same languages. This facilitates the training of neural network based text-to-speech models like (Ping et al., 2017), (Wang et al., 2017), (Shen et al., 2017) etc. in Indian languages. Additionally, we train Deep Voice 3 (Ping et al., 2017) for three languages, namely Hindi, Malayalam and Bengali. We provide baseline results for all of these languages which can be used for comparison purposes in future. We will release the whole corpus and the trained models for future research. In future, we plan to expand the corpus by collecting more data for the existing languages. We also plan to include more Indian languages like English with Indian accent, Tamil, Kannada, Marathi, Oriya, Punjabi, etc.

8. Bibliographical References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Chandramouli, C. and General, R. (2011). Census of india 2011. *Provisional Population Totals. New Delhi: Government of India*.

Griffin, D. and Jae Lim. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, April.

IMAI. (2017). Local language internet users in india. `https://cms.iamai.in/Content/ResearchPapers/7cbb85e8-fa66-4a41-9705-20dfa8bfe2b9.pdf`.

Jia, Y., Weiss, R. J., Biadsy, F., Macherey, W., Johnson, M., Chen, Z., and Wu, Y. (2019). Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*.

KR, P., Mukhopadhyay, R., Philip, J., Jha, A., Namboodiri, V., and Jawahar, C. (2019). Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1428–1436. ACM.

Kunchukuttan, A. (2013). Indic nlp library. `https://github.com/anoopkunchukuttan/indic_nlp_library`.

Mathew, M., Singh, A. K., and Jawahar, C. (2016). Multilingual ocr for indic scripts. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 186–191. IEEE.

Murthy, H. et al. (2016). Online indian text-to-speech systems. `https://www.iitm.ac.in/donlab/hts/`.

Philip, J., Namboodiri, V. P., and Jawahar, C. (2019a). Cvit-mt systems for wat-2018. *arXiv preprint arXiv:1903.07917*.

Philip, J., Siripragada, S., Kumar, U., Namboodiri, V., and Jawahar, C. V. (2019b). Cvit's submissions to wat-2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 131–136, Hong Kong, China, November. Association for Computational Linguistics.

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J., and Miller, J. (2017). Deep voice 3: 2000-speaker neural text-to-speech. *arXiv preprint arXiv:1710.07654*.

Pradhan, A., Prakash, A., Aswin Shanmugam, S., Kasthuri, G. R., Krishnan, R., and Murthy, H. A. (2015). Building speech synthesis systems for indian languages. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6, Feb.

Prajwal, K., Jawahar, C., and Kumaraguru, P. (2019). Towards increased accessibility of meme images with the help of rich face emotion captions. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 202–210. ACM.

Sankaran, N. and Jawahar, C. (2013). Error detection in highly inflectional languages. In *2013 12th Interna-*

tional Conference on Document Analysis and Recognition, pages 1135–1139. IEEE.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. (2017). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.

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: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., and Chen, Z. (2017). Sequence-to-sequence models can directly translate foreign speech. *arXiv preprint arXiv:1703.08581*.

Youth, I. (2010). Demographics and readership–results from the national youth readership survey 2009, national book trust, india in association with national council of applied economic research, new 3. *See the survey carried out by the telecom equipment vendor Ericsson, polling*, 47.

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. *arXiv preprint arXiv:1904.02882*.

9. Language Resource References

Baby, A., Thomas, A. L., and Myrthy, H. (2016). Resources for indian languages. In *Proceedings of Text, Speech and Dialogue*.

Ito, K. (2017). The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Park, K. (2019). Korean single speaker speech dataset. `https://www.kaggle.com/bryanpark/korean-single-speaker-speech-dataset`.

Prahallad, K., Elluru, N. K., Keri, V., Rajendran, S., and Black, A. W. (2012). The iiit-h indic speech databases. In *INTERSPEECH*.

Sonobe, R., Takamichi, S., and Saruwatari, H. (2017). Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis.