IMaSC: A Malayalam Speech Corpus for High-Quality Text-to-Speech Synthesis

Deepa P Gopinath[1], **Thennal D K**[2], **Vrinda V Nair**[1,*], **Swaraj K S**[3,*], **Sachin G**[3,*]
[1]Kerala Development and Innovation Strategic Council
[2]Language Technology Group, University of Hamburg
[3]GPS Renewables
{deepapgopinath,thennal10,vrinda66nair,swarajks10,sachingracious}@gmail.com
*These authors contributed equally to this work.

Abstract

Modern text-to-speech (TTS) systems use deep learning to synthesize speech increasingly approaching human quality, but they require a database of high-quality audio-text sentence pairs for training. Malayalam, the official language of the Indian state of Kerala and spoken by 35+ million people, is a low-resource language in terms of available corpora for TTS systems. In this paper, we present *IMaSC*, a Malayalam text and speech corpora containing 49 hours and 37 minutes of recorded speech. With 8 speakers and a total of 34,473 text-audio pairs, *IMaSC* is larger than every other publicly available alternative. We evaluated the database by using it to train TTS models for each speaker based on a modern deep learning architecture. With an average mean opinion score of 4.50, we find that the synthesized speech of our model is close to human quality.

Keywords: text-to-speech, malayalam, text-speech corpus, indic languages

1. Introduction

Malayalam, one among the 22 scheduled languages[1] of India, is the official language of the state of Kerala and union territories Lakshadweep and Puducherry. According to the Census of India (2011), Malayalam is the native language of around 35 million people in India. A South Dravidian subgroup of the Dravidian language family, it is also spoken by bilingual communities in contiguous parts of Karnataka and Tamil Nadu and by Malayali communities in various parts of the world[2]. Malayalam orthography is phonemic, with a one-to-one mapping of graphemes and phonemes with very few exceptions (Manghat et al., 2020). In the articulation of all phonetic segments, a pulmonic egressive airstream mechanism is used (Asher, 2013). Malayalam language consists of 15 vowels and 36 consonants (Manghat et al., 2020). The plosive symbols point to a six-way contrast in terms of place of articulation (Asher, 2013). Malayalam has voiced and voiceless aspirated stops in five places of articulation, a distinctive feature which is not an element of Dravidian languages. This can be attributed to the presence of large number of loan words from Sanskrit (Asher, 2013). Over the years, Malayalam has incorporated various aspects from other languages, with Sanskrit and later English being the most noteworthy examples (Bright, 1999). Other major languages whose vocabulary was integrated over the millennia include Arabic, Dutch, Hindustani, Pali, Persian, Portuguese, Prakrit, and Syriac (Pillai, 1965).

Like many other Indian languages, Malayalam is also a low-resourced language in terms of the availability of text-to-speech (TTS) corpora. Choudhary (2021) has noted that the information technology support in Indian languages has been lagging by decades compared to other languages like English, Japanese, or Russian, because of several factors including a lack of sufficient language resources required for the development of such technology.

As an agglutinative language, complex words can be formed in Malayalam by combining smaller words and morphemes (Premjith et al., 2018). In Malayalam, there is no absolute limit on the length and extent of agglutination. This results in a wide variation in the number of graphemes per word, and inflections at word or morpheme boundaries. These features engender the formation of a large vocabulary that necessitates the compilation of large corpora for speech-related applications (Srivastava et al., 2020).

In this paper, we present the Malayalam text and speech corpus created for the development of TTS in Malayalam. Curated with the help of linguists, the corpus consists of 34,473 sentences and 49.63 hours of speech read in studio conditions by 8 speakers (4 male and 4 female).

Our corpus, named *IMaSC—ICFOSS Malayalam Speech Corpus*, was evaluated by building a TTS system for Malayalam based on a deep learning architecture. Deep learning is opted in this work since techniques based on it have been state-of-the-art in

[1]Languages included in the VIII schedule of the Constitution of India

[2]https://www.britannica.com/topic/Malayalam-language

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5864–5872
11–16 May 2026.

Figure 1: Word cloud representation of the text in *IMaSC*

speech synthesis for many years (Srivastava et al., 2020). Deep learning algorithms learn from the given training data and build models for achieving the required functionality, and as such the quality and size of the database for training are critical in the performance of the model (Marcus, 2018; Tan et al., 2021). As such, the ubiquity, effectiveness, and data-critical nature of TTS systems based on deep learning makes them a suitable option for evaluating the sufficiency and quality of a speech corpus.

Speaker ID	Gender	Age
M1	Male	28
F1	Female	43
M2	Male	26
F2	Female	22
M3	Male	48
F3	Female	23
M4	Male	25
F4	Female	24

Table 1: Details of speakers

2. Related Work

2.1. TTS corpus in Indian Languages

Owing to the lack of large text-to-speech corpora, the progress of developing reliable text-to-speech systems for Indian languages has been relatively slow (Srivastava et al., 2020). One of the first efforts in this area is reported by Prahallad et al. (2012). Baby et al. (2016) later developed a much larger resource for Indian languages, IndicTTS, which contains about 8 hours of speech data for 13 Indian languages. Pradhan et al. (2015) used this corpus to train text-to-speech systems for these 13 languages. However, Srivastava et al. (2020) assert that the data provided for each language in IndicTTS is insufficient for training recent neural-network-based systems that can produce natural accurate speech. To address this gap, they present IndicSpeech, a large-scale text-to-speech corpus for 3 Indian languages—Hindi, Malayalam, and Bengali—aimed at training neural TTS systems. They trained TTS models on their corpora, and the mean opinion score (MOS, Streijl et al., 2016) obtained for the Malayalam TTS model is lower than those obtained for Hindi and Bengali. They

Speaker	Time (HH:MM:SS)	Sentences	Words		Phonemes
			Total	Unique	Total
M1	06:08:55	4,332	28,508	15,912	239,066
F1	05:22:39	4,294	28,196	16,221	237,405
M2	05:34:05	4,093	26,742	15,223	226,715
F2	06:32:39	4,416	29,015	16,358	243,611
M3	05:58:35	4,239	27,777	15,937	235,163
F3	04:21:56	3,242	21,489	13,087	177,120
M4	06:04:43	4,219	27,390	15,599	233,467
F4	09:34:21	5,638	36,664	20,649	318,841
Total	**49:37:55**	**34,473**	**225,781**	**23,604**	**1,911,388**

Table 2: Details of Speech corpus and corresponding text

attribute the discrepancy to fundamental characteristics of Malayalam, such as the morpho-phonemic changes during word formation. They suggest that one of the solutions would be to increase the size of the Malayalam corpus to cover a larger vocabulary. Concurrent work by He et al. (2020) presents a multi-speaker corpus for 6 Indian languages, with approximately 6 hours of data for Malayalam, significantly less than that of IndicSpeech with 29 hours. Most recently, Sankar et al. (2024) presents IndicVoices-R, a TTS corpus for 22 Indian languages derived from a preexisting automatic speech recognition (ASR) corpus, with roughly 83 hours of speech for Malayalam across 462 speakers. However, only 5 hours of that is read speech, the rest being extempore, and each speaker only has 10 minutes of speech on average. They note that as it is derived from an ASR corpus, the quality is markedly lower in comparison to studio-quality TTS datasets.

2.2. TTS corpus evaluation using deep learning models

A TTS speech corpus can be evaluated by testing the quality of synthetic speech generated with the corpus (Dybkjær et al., 2007). Ahmad et al. (2021) prepared a phonetically balanced Bangla corpus and evaluated it using a Bangla neural synthesizer based on Merlin (Wu et al., 2016), an open-source speech synthesis toolkit using deep neural networks. Deep learning architectures based on Tacotron (Wang et al., 2017) and Tacotron 2 (Shen et al., 2018) were used for evaluation of CSS10 (Park and Mulc, 2019), LibriTTS (Zen et al., 2019), Latvian corpus created by Darǵis et al. (2020), DiDiSpeech (Guo et al., 2021), KazakhTTS (Mussakhojayeva et al., 2021), and TTS-Portuguese Corpus (Casanova et al., 2022). FastSpeech (Ren et al., 2019) was used for the evaluation of AISHELL-3 (Shi et al., 2020) and Didispeech (Guo et al., 2021). IndicSpeech, a corpus curated for 3 Indian languages—Hindi, Malayalam, and Bengali—trains Deep Voice 3 models (Ping et al., 2017) to evaluate the corpus. BU-TTS, a bilingual Welsh-English speech corpus developed by Russell et al. (2022), was evaluated by training VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) models (Kim et al., 2021).

3. Method

3.1. Corpus design

The task of compiling a phonetically rich corpus involves linguistic analysis of a large raw text corpus, which we did in collaboration with linguists from the Department of Linguistics, University of Kerala[3]. The linguistic analysis comprised text collection and preparation of corpus explained in the sections that follow.

3.1.1. Text collection

The text corpus was derived from the Malayalam Wikipedia. Launched in December 2002, it has grown to be an online encyclopedia containing 79,510 articles as of October 2022[4]. This choice was made primarily because the articles of Wikipedia are in the public domain and the scale of the wiki is sufficient to compile a phonetically balanced database. This enabled us to select a set of phonetically balanced sentences, record the corresponding speech, and release it in the public domain without any copyright infringements.

[3] https://www.keralauniversity.ac.in/home

[4] https://en.wikipedia.org/wiki/Malayalam_Wikipedia

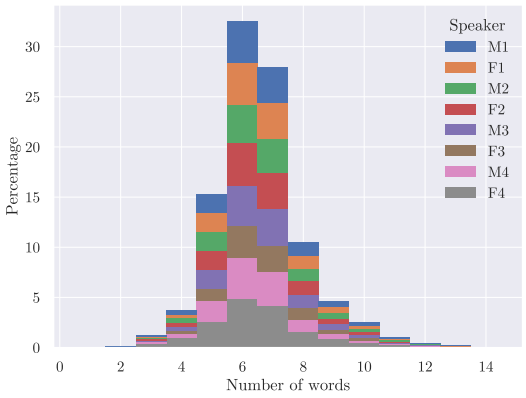

Figure 2: Histogram of the number of words in each sentence

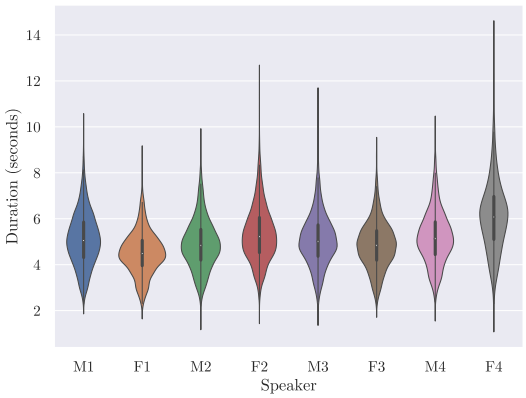

Figure 3: Violin plot indicating the distribution of duration of audio samples for each speaker

3.1.2. Preparation of text corpus

Preliminary data cleaning was done on the text obtained from Wikipedia. The cleaned data was separated into a set of sentences, and only sentences composed entirely of Malayalam characters and punctuation were kept. Sentences were sampled from this set and vetted by linguists for quality in terms of naturalness, semantics, and syntax. The phonetic representation of the database was enhanced by ensuring that all phonemes used in Malayalam discourses occur 100 times in the database at a minimum.

3.2. Speaker selection

People capable of correct pronunciation, pleasing rhythm, and consistent articulation with respect to the standard Malayalam dialect were subjectively selected for recording speech. The details of the speakers we engaged for our speech corpus generation are given in Table 1. Speakers were selected to achieve a diverse range of prosody and articulation. In the case of multi-speaker TTS systems, voices with different characteristics are preferable, since it gives the user a wide range of choices.

3.3. Recording of speech

The recording was done in a soundproof recording studio in the Phonetics Lab of the Department of Linguistics, University of Kerala. A RODE microphone was used as the recording device, and the recording was sampled at 48 kHz, mono, with 24-bit depth. The equipment configuration was tailored to the speaker and studio conditions, to produce high-quality audio. The distance between the speaker and the microphone was set in the range of 4-6 inches. The noise floor of the studio was kept below -60 dBFS, in contrast to the speech signals maintaining -20 dBFS or above.

A project assistant closely monitored the recordings and noted down discrepancies in comparison with the text, if any. We found that close monitoring and support during recording sessions are required to ensure the consistency of the read speech and to reduce discrepancies as much as possible. The voice quality was observed and intermittent breaks were given to the speakers as required to avoid fatigue. Mismatches that occurred were corrected by rerecording those sentences. For certain difficult sentences, a trial reading was carried out before the actual recording to ensure quality recordings. Particularly difficult sentences were discarded by each speaker when deemed necessary.

3.4. Post processing and speech corpus compilation

The recorded voice files were segmented into sentences automatically by detecting the silence between sentences. The automatically segmented speech was evaluated and corrected manually. Each audio file was examined in comparison with the corresponding text and corrections were made wherever required. All the sentences in the text corpus were uniquely labeled and the audio files carrying the corresponding speech were saved with a file name matching the label.

The compiled speech corpora for each of the artists were again vetted by language experts for mismatches between the articulated phonemes in the audio file and those in the corresponding text file. Sentences with substantial mismatches between text and audio were discarded. If the mismatch was slight, then the text was corrected to match the audio. Through this process, it was ensured that the audio files and the corresponding text matched up to the phonemic level.

4. Evaluation of the database using a deep learning TTS system

Traditional neural TTS architectures use two separate components for generative modeling, splitting the process into two stages. The first stage generates from text an intermediate representation of speech features, such as Mel-spectrograms, using an acoustic model. The second stage synthesizes a raw waveform from the intermediate representation via a neural vocoder. Those models are trained separately and then joined for inference. Two-stage pipelines, however, require a sequential and costly training procedure, and their dependence on predefined intermediate features such as Mel-spectrograms limits the potential of utilizing learned internal representations to improve performance further (Tan et al., 2021; Kim et al., 2021).

The VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) network is a parallel end-to-end architecture for TTS that outperforms traditional two-stage architectures and synthesizes natural-sounding speech extremely close to human quality (Kim et al., 2021). VITS circumvents the issues laden in two-stage pipelines by connecting the two modules of TTS systems through latent variables to enable efficient end-to-end learning. It is used as a baseline comparison model for advancements in recent TTS methods and applications. van Rijn et al. (2022) used a modified version of VITS for personalized voice generation, while Song et al. (2022) used multi-speaker VITS as a base architecture for talking face generation. Casanova et al. (2022) modified the network to achieve state-of-the-art results in zero-shot multi-speaker TTS and zero-shot voice conversion. Russell et al. (2022) trained VITS models for evaluating BU-TTS, a bilingual Welsh-English speech corpus.

Due to its simplified training procedure and improved performance, we decided to use VITS to evaluate the dataset over older but more common architectures for TTS systems such as Tacotron or FastSpeech. The model was separately trained for each of the 8 speakers. The trained models were then used for inference to generate synthetic speech to be evaluated via a mean opinion score (MOS) listening test (Streijl et al., 2016).

5. Results

5.1. Details of *IMaSC*

488,249 sentences were obtained from Wikipedia, from which 8,853 unique sentences were obtained after linguistic evaluation and processing as outlined in sections 3.1 and 3.4. The details of the speech corpus and the corresponding text of each speaker after quality check and data cleaning are given in Table 2. The reported times for each speaker in the table are rounded to the nearest second. The recorded speech of F4 is significantly higher than the rest of the speakers, in terms of the number of sentences and total duration of speech, as the speaker was engaged in experimental recording sessions as well.

Speaker	Synthesized	Ground Truth
M1	4.60 ± 0.09	4.75 ± 0.19
F1	4.48 ± 0.11	4.75 ± 0.19
M2	4.38 ± 0.11	4.58 ± 0.23
F2	4.55 ± 0.10	4.65 ± 0.26
M3	4.46 ± 0.12	4.68 ± 0.23
F3	4.58 ± 0.10	4.90 ± 0.14
M4	4.46 ± 0.12	4.55 ± 0.26
F4	4.54 ± 0.11	4.58 ± 0.23
Average MOS	$\mathbf{4.50 \pm 0.04}$	$\mathbf{4.68 \pm 0.08}$

Table 3: Mean Opinion Score of the speech synthesized from the recorded database of each of the 8 speakers and average MOS

The word cloud representation of the 2,25,781 words in the text corpus of *IMaSC* is given in Fig. 1. The word cloud indicates the most frequent words and provides a broad visual description of the words in the database. A close look into the word cloud reveals that the number of characters per word varies widely, indicating the agglutinative nature of Malayalam.

A histogram of the number of words per sentence read by the 8 speakers is given in Fig. 2. It shows that more than 30% of the sentences are 6 words long. It can be noted that F4 has spoken the longest sentences in terms of the number of words.

The distribution of the duration of the audio samples for each speaker is given in Fig. 3. It can be seen that the duration of the audio samples largely lies between 2s to 8s. The minimum variation is for speaker F1 (approximately 2s to 9s) and maximum for F4 (approximately 1s to 15s). Longer duration audio is expected for F4, as she has spoken comparatively longer sentences. The variation in the distribution of duration between the speakers is due to diversity in the manner of articulation and variation in the set of sentences read by the speakers.

The scatterplot of the text length (number of characters) in each sentence and the duration of its corresponding audio is given in Fig. 4. The text length for each sentence correlates linearly with the corresponding audio duration as per the scatterplot. It also shows that text length mostly ranges between 40 to 100 and the corresponding duration between 2s to 8s. F4 has the audio sample with the

Database	Sentences	Total Words	Hours	No. of speakers
Baby et al. (2016)	11,300	58,098	17.89	2
IndicSpeech (Srivastava et al., 2020)	19,954	109,245	29.1	1
He et al. (2020)	4,126	25,330	5.51	**42**
IMaSC	**34,473**	**225,781**	**49.63**	8

Table 4: Comparison of *IMaSC* with other TTS corpora

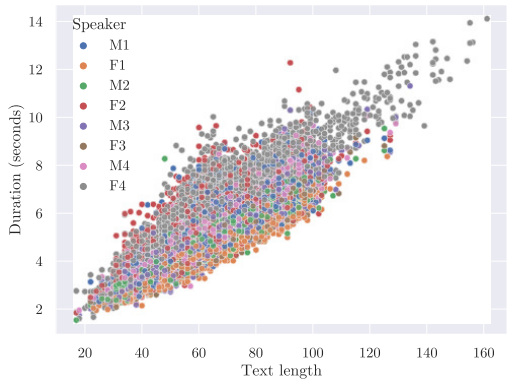

Figure 4: Scatterplot of the number of characters in each sentence and the duration of its corresponding audio

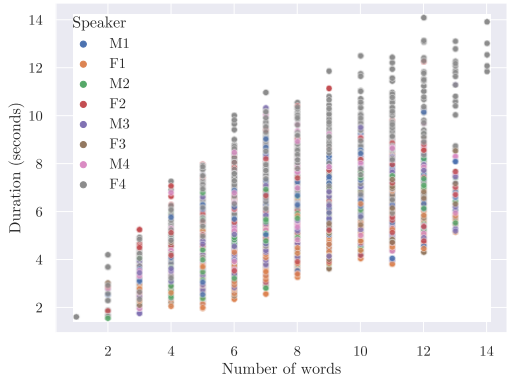

Figure 5: Scatterplot of the number of words in each sentence and the duration of its corresponding audio

maximum text length (160). The range of duration values as seen in this plot is in tune with that of the violin plot in Fig. 3.

The number of words in the sentence also correlates similarly with audio duration as per the scatterplot in Fig. 5, though with significantly more variation. This is again a consequence of agglutination in Malayalam, with the length of a word varying widely. For example, in the case of sentences with 6 words, the duration varies from 2s to 10s. It is noted that F4 has more instances of longer sentences in terms of duration and number of words. The same fact is also seen in Fig. 3 and Fig. 4. The analysis of the corpus shows that *IMaSC* has covered a large variety of words. This can make it a suitable candidate for training a deep learning TTS system with multiple speakers.

5.2. Evaluation of *IMaSC* using VITS

We used the public VITS implementation available at Coqui TTS[5]. The character set for Malayalam, including punctuation, was directly tokenized, as Malayalam characters have close to a one-to-one correspondence with phonemes. Tokenized raw text and the corresponding speech were used for training. A separate model for each speaker was trained to evaluate their datasets individually. Each model was trained for 60k steps on a Tesla P100 GPU.

5.2.1. Mean Opinion Score

We conducted a crowd-sourced Mean Opinion Score (MOS) listening test with 20 native language speakers to evaluate the models (Streijl et al., 2016). 10 sentences were randomly selected from the test dataset to be synthesized, and 2 sentences, along with their corresponding audio, were chosen for evaluating ground truth. In the test, each evaluator thus had 12 text-audio pairs to be evaluated, repeated for all 8 speakers for a total of 96 text-audio pairs. The different audio samples were each scored on a 5-point scale for naturalness, with 5 being excellent, 4 being good, 3 being fair, 2 being poor, and 1 being bad.

Table 3 gives the score obtained for each of the 8 speakers. The MOS for ground truth and synthesized speech for each speaker is detailed in the table. We note that each of our models performs close to ground truth.

5.3. Comparison of *IMaSC* with other TTS corpora

Table 4 provides a comparison of *IMaSC* with other publicly available Malayalam TTS corpora. Given that IndicVoices-R is a dataset derived from automatic speech recognition corpora with notably

[5] https://github.com/coqui-ai/TTS

lower quality recordings, we do not include it in this comparison (Sankar et al., 2024). We note that *IMaSC* is significantly larger than other previous corpora both in the number of sentences and hours of recorded audio. IndicSpeech employs a single female speaker while Baby et al. (2016) employ one male and one female, as compared to our work which employs 4 male and 4 female speakers. He et al. (2020) use crowdsourcing to record audio and thus have 42 speakers (18 male and 24 female), but with only 5.51 hours of total speech, on average each speaker has less than 8 minutes of speech. A single-speaker TTS system is thus infeasible. In comparison, each speaker in our database has a minimum of 4.37 hours and an average of 6.20 hours. We also note that in contrast to the crowdsourcing approach, our speaker selection is deliberate and intended to represent a range of prosody and articulation styles while maintaining clear and comprehensive speech, as detailed in Section 3.2.

6. Conclusion

In this work, we presented *IMaSC*, a Malayalam text and speech corpora that aims to address the lack of publicly available data for TTS applications. With 49 hours and 37 minutes of audio, 8 speakers and 34,473 text-audio pairs, *IMaSC* is larger than any other public Malayalam text and speech corpus by a wide margin, with proper care taken for quality control. We trained end-to-end TTS models for each speaker based on the VITS architecture to evaluate the database and conducted a subjective MOS listening test with 20 participants. We reported an average MOS of 4.50 for speech synthesized with our models, close to the ground truth of 4.68. With an average of more than 6 hours per speaker across 8 speakers, *IMaSC* will enable the development of multi-speaker text-to-speech synthesis systems, which provide the user with a choice of selecting synthesized speech with prosody and other voice qualities of their choice.

7. Acknowledgements

We would like to thank the International Centre for Free and Open Source Software (ICFOSS), an autonomous organization set up by the Government of Kerala for promoting free and open source software, for funding this project. We also acknowledge the support rendered by Shijith S. and other linguists at Department of Linguistics, University of Kerala in different stages of the database creation.

8. Bibliographical References

Ronald Asher. 2013. *Malayalam*. Routledge.

William Bright. 1999. RE Asher & TC Kumari, Malayalam.(Descriptive grammars.) London & New York: Routledge, 1997. Pp. xxvi, 491. *Language in Society*, 28(3):482–483.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

Narayan Choudhary. 2021. LDC-IL: The Indian repository of resources for language technology. *Language Resources and Evaluation*, 55(3):855–867.

Laila Dybkjær, Holmer Hemsen, and Wolfgang Minker. 2007. *Evaluation of text and speech systems*, volume 38. Springer Science & Business Media.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Sreeja Manghat, Sreeram Manghat, and Tanja Schultz. 2020. Malayalam-English Code-Switched: Grapheme to Phoneme System. In *INTERSPEECH*, pages 4133–4137.

Gary Marcus. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

S Kunjan Pillai. 1965. *Malayalam Lexicon*. University of Kerala.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*.

Abhijit Pradhan, Anusha Prakash, S Aswin Shanmugam, GR Kasthuri, Raghava Krishnan, and Hema A Murthy. 2015. Building speech synthesis systems for Indian languages. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6. IEEE.

B Premjith, KP Soman, and M Anand Kumar. 2018. A deep learning approach for Malayalam morphological analysis at character level. *Procedia computer science*, 132:47–54.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Hyoung-Kyu Song, Sang Hoon Woo, Junhyeok Lee, Seungmin Yang, Hyunjae Cho, Youseong Lee, Dongho Choi, and Kang-wook Kim. 2022. Talking Face Generation with Multilingual TTS. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21425–21430.

Robert C Streijl, Stefan Winkler, and David S Hands. 2016. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Pol van Rijn, Silvan Mertes, Dominik Schiller, Piotr Dura, Hubert Siuzdak, Peter Harrison, Elisabeth André, and Nori Jacoby. 2022. VoiceMe: Personalized voice generation in TTS. *arXiv preprint arXiv:2203.15379*.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Zhizheng Wu, Oliver Watts, and Simon King. 2016. Merlin: An open source neural network speech synthesis system. In *Proc. 9th ISCA Speech Synthesis Workshop (SSW9)*, Sunnyvale, CA, USA.

9. Language Resource References

Arif Ahmad, Md Reza Selim, Md Zafar Iqbal, and M Shahidur Rahman. 2021. SUST TTS Corpus: A phonetically-balanced corpus for Bangla text-to-speech synthesis. *Acoustical Science and Technology*, 42(6):326–332.

Arun Baby, Anju Leela Thomas, NL Nishanthi, TTS Consortium, et al. 2016. Resources for Indian languages. In *Proceedings of Text, Speech and Dialogue*.

Edresson Casanova, Arnaldo Candido Junior, Christopher Shulby, Frederico Santos de Oliveira, João Paulo Teixeira, Moacir Antonelli Ponti, and Sandra Aluísio. 2022. TTS-Portuguese Corpus: a corpus for speech synthesis in Brazilian Portuguese. *Language Resources and Evaluation*, pages 1–13.

Roberts Darǧis, Peteris Paikens, Normunds Gruzitis, Ilze Auziņa, and Agate Akmane. 2020. Development and Evaluation of Speech Synthesis Corpora for Latvian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6633–6637.

Tingwei Guo, Cheng Wen, Dongwei Jiang, Ne Luo, Ruixiong Zhang, Shuaijiang Zhao, Wubo Li, Cheng Gong, Wei Zou, Kun Han, et al. 2021. Didispeech: A large scale mandarin speech corpus. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6968–6972. IEEE.

Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, and Knot Pipatsrisawat. 2020. Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6494–6503, Marseille, France. European Language Resources Association (ELRA).

Saida Mussakhojayeva, Aigerim Janaliyeva, Almas Mirzakhmetov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2021. Kazakhtts: An open-source kazakh text-to-speech synthesis dataset. *arXiv preprint arXiv:2104.08459*.

Kyubyong Park and Thomas Mulc. 2019. Css10: A collection of single speaker speech datasets for 10 languages. *arXiv preprint arXiv:1903.11269*.

Kishore Prahallad, E Naresh Kumar, Venkatesh Keri, S Rajendran, and Alan W Black. 2012. The IIIT-H Indic speech databases. In *Thirteenth annual conference of the international speech communication association*.

Stephen John Russell, Dew Bryn Jones, and Delyth Prys. 2022. BU-TTS: An Open-Source, Bilingual Welsh-English, Text-to-Speech Corpus. In *LREC 2022 Workshop Language Resources and Evaluation Conference 20-25 June 2022*, page 104.

Ashwin Sankar, Srija Anand, Praveen Srinivasa Varadhan, Sherry Thomas, Mehak Singal, Shridhar Kumar, Deovrat Mehendale, Aditi Krishana, Giri Raju, and Mitesh M. Khapra. 2024. Indicvoices-r: Unlocking a massive multilingual multi-speaker speech corpus for scaling indian TTS. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*.

Nimisha Srivastava, Rudrabha Mukhopadhyay, KR Prajwal, and CV Jawahar. 2020. Indicspeech: text-to-speech corpus for Indian languages. In *Proceedings of the 12th language resources and evaluation conference*, pages 6417–6422.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.