Efficient Human-Quality Kannada TTS using Transfer Learning on NVIDIA's Tacotron2

Anil Kumar K K
RaGaVeRa Indic Technologies Pvt Ltd
Bengaluru, 560012
Email: anil@ragavera.com

Shiva Kumar H R
RaGaVeRa Indic Technologies Pvt Ltd
Bengaluru, 560012
Email: shivahr@ragavera.com

Ramakrishnan Angarai Ganesan
Department of Electrical Engineering
Indian Institute of Science
& RaGaVeRa Indic Technologies Pvt. Ltd.
Bengaluru, 560012
Email: agr@iisc.ac.in

Jnanesh K P
RaGaVeRa Indic Technologies Pvt Ltd
Bengaluru, 560012
Email: officialjnan@gmail.com

***Abstract*—Very good quality, speech synthesis systems exist for languages like English and Chinese. However, only in the recent past, increased attention has been paid for developing TTS for Indian languages. There have been several reasons for the same in the past: 1) lack of adequate market, 2) non-availability of quality training data. In this work, we have developed a human-like quality Kannada text-to-speech conversion system using about 44.8 hours of training data recorded from a studio from a Kannada teacher with good diction. We have used the transfer learning technique to continue training over the Tacotron2 and WaveGlow checkpoints pre-trained on English. Evaluation by thirty five Kannada natives resulted in an overall MOS of 4.51 ± 0.52, whereas the original speech of the speaker was given an MOS of 4.62 ± 0.53. In another independent testing, where another set of 25 human evaluators were given ten pairs of the original utterances of the speaker and the synthesized speech of the same sentences, some of the synthesized speech samples were judged to be better than the original! In a final round of evaluation, five sentences were synthesized by our TTS, Google's Wavenet TTS and also Nuance's TTS. Kannada natives were presented these outputs in a random order and asked to choose one of them as their most preferred output. Based on 55 human evaluators, RaGaVeRa's Kannada TTS obtained a mean preference score of 78.2%, whereas Google's and Nuance's TTS got scores of 13.1% and 5.1%, respectively. Thus, to the best of the knowledge of the authors, this is the best quality TTS that has ever been achieved for Kannada so far.**

***Index Terms*—Kannada, English, speech synthesis, Tacotron2, WaveGlow, vocoder, end-to-end TTS, deep learning, transfer learning, RaGaVeRa, Nuance, Google.**

I. Introduction

The task of generating natural sounding speech from text remains a challenging problem to be solved. Deep learning based TTS systems are the current state-of-the-art in terms of producing natural sounding speech. Traditionally, concatenative [1]–[4] and parametric synthesis techniques were prevalent and required complex preprocessing [5] and pipelines and resulted in muffled speech. Besides, the speech output may have glitches and instabilities. They also employed special modules to improve naturalness [6]–[9].

Rapid development in deep learning based methods has shown immense success in this field. End-to-end generative models such as Tacotron2 [10], [11] and Deep Voice [12] have been proposed, which have replaced traditional pipelines. These models have demonstrated state-of-the-art performance by confining the entire pipeline involving spectrogram prediction and speech synthesis into a single pipeline. However, these end-to-end models require tens of hours of speech data and a lot of computational power. However, a TTS for Sanskrit was attempted on the Tacotron2 [13] plus WaveGlow model with limited training data using transfer learning [14].

Kannada is one of the classical languages of India. With eight Jnanpith awardees, it has a good literature, and good scope exists for creating audio books of Kannada works by popular authors. However, even though there has been a lot of development work on TTS for Indian languages in the past two decades [15]–[17], there has been less attempt in developing Kannada TTS and the only significant work in this regard has been reported by Shiva Kumar et al. [3], which was ranked second in the International Blizzard Speech Synthesis Challenge for Indian Languages in 2013.

In the current work, we perform transfer learning [18]–[21] on a model pre-trained on English. The existing Tacotron2 model with WaveGlow decoder, pre-trained with LJ speech corpus (English), is tuned with our curated high quality Kannada speech data.

II. Development details of the Kannada TTS

TTS systems using deep neural networks normally use different stages, but systems such as Tacotron2 [11] com-

prise two stages only, namely an acoustic model and a vocoder. Such end-to-end synthesis models are trained on speech and matching text without cumbersome phoneme-level annotation of the training speech corpus. In Tacotron2, the acoustic model is a recurrent network, which predicts a sequence of mel spectograms from an input character sequence. The WaveGlow vocoder [22] generates time-domain waveforms from the sequence of mel spectograms fed to it by the network.

The details of the Tacotron2-Waveglow architecture are given in Section II-A. The relevant details of the Kannada training speech collected by us are explained in Sec. II-B. Section II-C explains the preparation of the speech corpus for training the Tacotron2 model. Section II-D deals with the processes of pretraining and transfer learning.

A. Details of NVIDIA's Tacotron2 code used for synthesis

A number of standard models exist for speech synthesis, all of which make use of deep learning. After a reasonable literature survey, we have chosen Tacotron2 and WaveGlow. The Tacotron2 architecture makes use of 'location sensitive attention' and is an encoder-attention-decoder model. To begin with, an encoder converts the input character string into a word embedding vector. The decoder predicts the corresponding spectrograms from the embedding vector. From the spectograms generated by Tacotron2, the actual speech waveform is created utilizing the WaveGlow vocoder.

In this work, we use the implementations provided in [23], [24]. Tacotron2 and Waveglow networks are trained separately. No explicit duration or other models were used, apart from the intrinsic learning offered by Tacotron2. A WaveGlow model, pretrained with the same LJ speech corpus, is further trained using the Kannada speech corpus. The WaveGlow vocoder has been shown to work on unseen languages and speakers [25].

B. Speech Data Used for TTS Development

The dataset was recorded from a middle aged lady, whose voice was selected from among 6 speakers by about ten human evaluators based on the pitch, diction, and evaluation of variations in pitch and amplitude. This dataset consists of 20000+ utterances of read speech, which are in .wav format. The transcriptions were edited and matched to the actual utterances. The audio files and their transcripts of individual utterances are used as they are. The details of the Kannada dataset curated are given in Table I.

TABLE I: Details of the Kannada speech dataset utilized

Statistics	Value
Total No. of utterances	20,397
Net duration of data	44h 49min
Length of shortest utterance	0.37 sec
Length of longest utterance	23.2 sec
Mean utterance length	7.85 sec

C. Preparation of the speech corpus for training

Proper preprocessing of the voice data before using it for training results in better synthesis [21]. Our preprocessing consisted of:

- Trimming the silences at the beginning and end of the utterances, leaving out a 50 ms of silence uniformly in all of them.
- Amplitude of all audio files were normalized.
- The training speech data was downsampled to 22 KHz from 48 KHz.
- Text normalization: All the numbers and abbreviations in the text were expanded [26].

Tacotron2 employs location-sensitive attention; thus, long silences in the training data slow down the attention learning. Hence, both the silences in the start and end of the training utterances were pruned to 50 ms. The abbreviations, numerals etc. in the transcripts were normalized and converted to UTF-8 format unicode.

D. Pretraining the architecture using LJ speech corpus

The freely available LJ Speech dataset comprises 23.9 hours of speech collected from a female speaker and its text transcript. Tacotron2 checkpoint [27] pretrained with this dataset is published by Nvidia, on which we continue training with our Kannada corpus. This is known as transfer learning, since the network benefits from the pretraining. Similarly, the pretrained WaveGlow model [28] is also fine tuned for Kannada.

III. Experimental Results

A. Training and tuning the Model

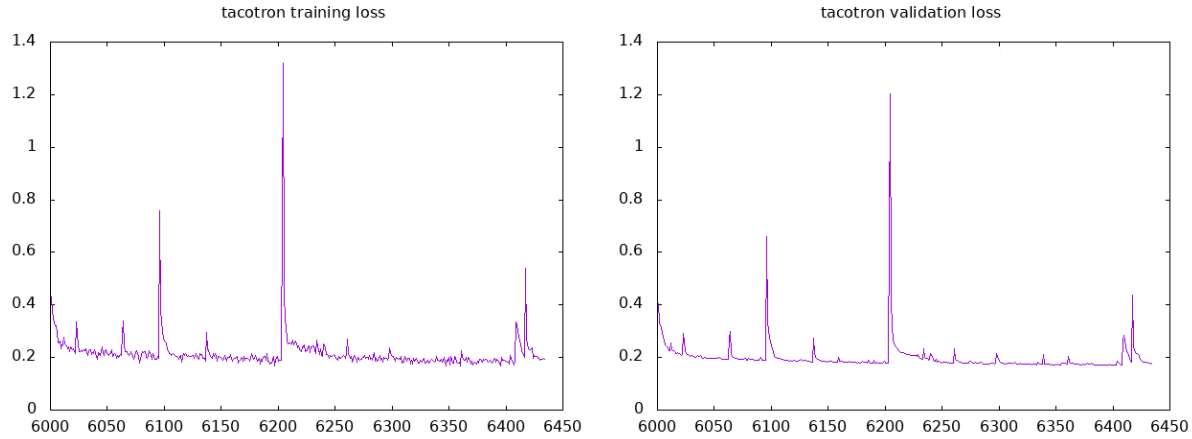
The Tacotron2 model pretrained for 6000 epochs with the LJ speech dataset is trained for 425 more epochs using our Kannada dataset. Tables II and III list the final set of hyperparameters.

TABLE II: Hyperparameters used for training Tacotron2

Hyperparameter	Value
η, Learning rate	$1 * 10^{-3}$
Size of batch	32
No. of epochs	7701
ϵ, Decay of weight,	$1 * 10^{-6}$
Factor of annealing	0.1
Anneal steps	6700 7200 7700

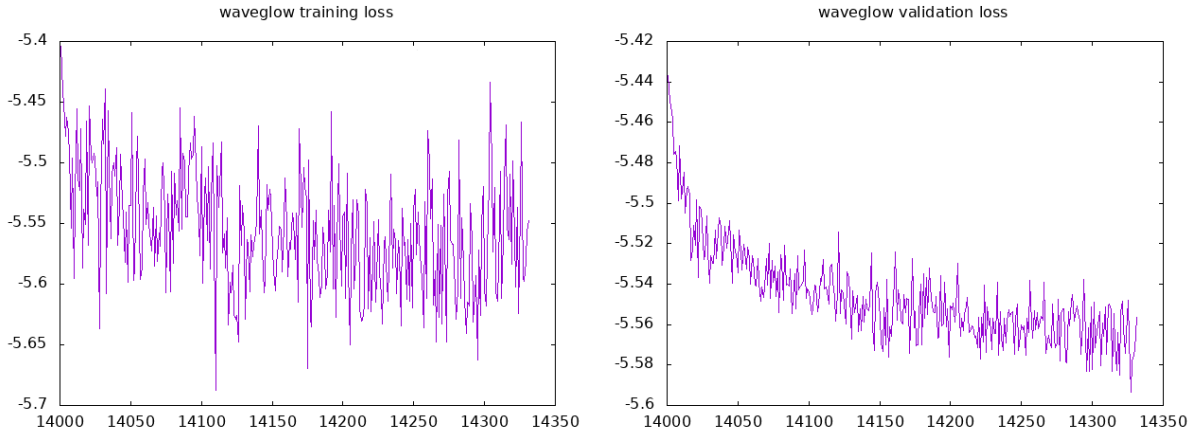
TABLE III: Hyperparameters for Waveglow (continued training from pre-trained model from NVIDIA)

Hyperparameter	Value
No. of Epochs	15500
Size of batch	10
Length of segment	16000
Decay of weight	0
grad-clip-thresh	65504.0



(a) Training Loss Curve (b) Validation Loss Curve

Fig. 1: Convergence of the Tacotron2 loss as a function of the number of iterations.



(a) Training Loss Curve (b) Validation Loss Curve

Fig. 2: Convergence of the loss (in Waveglow) as a function of the number of iterations.

Images (plots) for training and validation loss for Tacotron2 and Waveglow training are shown in Figs. 1 and 2. These were plotted using GNUPlot tool by extracting values from the log files.

B. Evaluation of the TTS Output by Kannada Natives

The quality of the synthesized speech was evaluated by thirty five Kannada natives, who can read and write in Kannada. Each of them evaluated twelve utterances out of which eight were synthesized speech samples and four were original utterances of the speaker. However, the evaluators were told that all of them were synthesized outputs. All the evaluators were adequately trained on the MOS scale using example English sentences [29]. The evaluation results are tabulated in Table IV. We have obtained a MOS of 4.62 ± 0.53 for the original speech and 4.51 ± 0.52 for the synthesized speech. Thus, the output of our TTS system can be considered as state-of-the-art quality for Kannada. Figure 3 summarizes these evaluations as a bar chart.

TABLE IV: Human evaluation of the accuracy and naturalness of the synthesized Kannada speech by 35 natives of Kannada. Some of the original utterances of the speaker were also mixed with the test samples and given for blind evaluation. Mean scores for both types of speech are listed.

Evaluated speech	**Mean opinion score**
Original speech	4.62 ± 0.53
RaGaVeRa TTS	4.51 ± 0.52

In a completely different type of evaluation, we asked another set of twenty five human evaluators to compare the quality of synthesized speech against the original speech of the speaker for the same sentence. Each evaluator listened

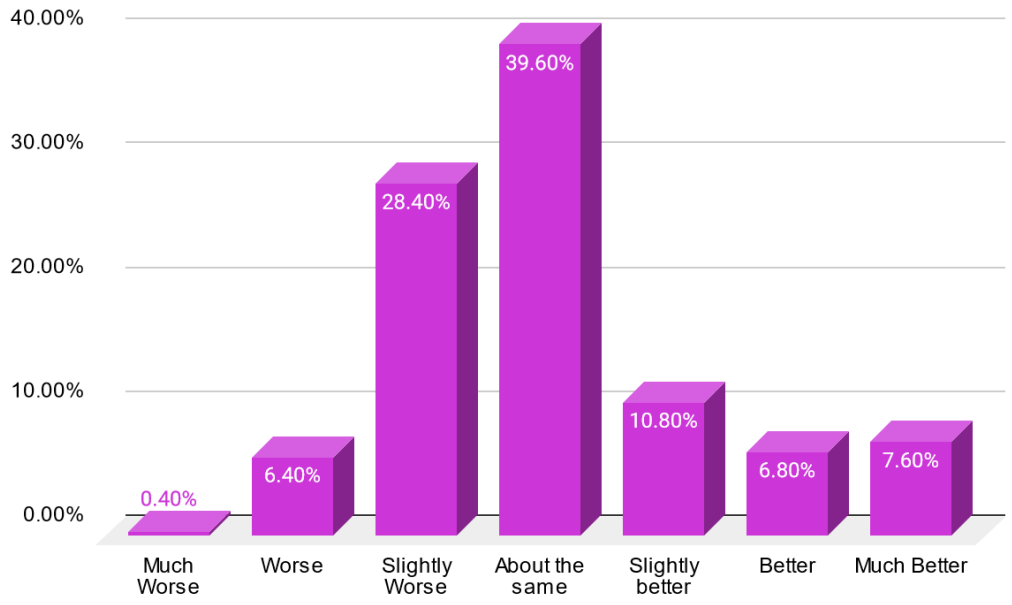


Fig. 3: Comparison of the synthesized speech with the ground truth (original speech of the speaker) by 25 natives of Kannada. 25% of the synthesized outputs have been rated to be better than the original utterances!

to ten pairs of original and synthesized speech. They tagged the synthesized utterances as (i) much worse, (ii) worse, (iii) slightly worse, (iv) about the same, (v) slightly better, (vi) better or (vii) much better than the actual speech of the same sentence. The labels were given scores from -3 to 3, with 'much worse' having a score of -3, to 'much better' having a score of +3.

The results of this interesting evaluation experiment are plotted in Fig. 3. It is clear that the quality of synthesized speech is adjudged as almost of the same quality as the original speech of the speaker. Based on the way the scoring was performed, if there were equal number 'better' and 'worse' evaluations, the mean score would be zero. Thus, a positive value of mean score implies that the synthesized speech is better than the original and vice versa. The mean score obtained from the 25 evaluators on the 10 synthesized sentences is 0.048 ± 1.282. Thus, the Kannada natives have clearly evaluated our TTS output as marginally better than the original utterances of the speaker! This could be due to the differences in the speaking rates. In fact, when some sample synthesized sentences were sent to the original speaker, she actually thought that they were part of the utterances recorded from her.

Some of the listeners opined that the synthesized speech is slightly fast, even though it does not affect the intelligibility, which has been rated as very good. However, the models we used do not provide any handle for modifying the speaking rate. Also, there was a feedback about the mispronunciation of the aspirated consonants, which is basically an issue of the speaker, which cannot be handled by the synthesis system.

Finally, we also compared the quality of our synthesized speech with that of Google's WaveNet TTS and Nuance's TTS. In a new experiment distinct from the two experiments described above, we synthesized 5 sentences using RaGaVeRa's, Google's and Nuance's Kannada TTS, which were evaluated by 55 Kannada natives. The three synthesized outputs were presented to them in a random order and they had to choose one of them as their most preferred rendition. The obtained mean preference scores are listed in Table V. The results clearly show the superiority of our TTS over those of Google and Nuance, as far as Kannada is concerned. Figure 4 illustrates these evaluations as a bar chart.

TABLE V: Comparison of the quality of the synthesized Kannada speech with that of Google's WaveNet TTS and Nuance TTS as assessed by 55 natives of Kannada. Mean preference scores for the outputs of the three synthesizers are listed.

TTS Engine	Mean preference score
Google TTS	13.1%
Nuance TTS	5.1%
RaGaVeRa TTS	78.2%
No Preference	3.6%

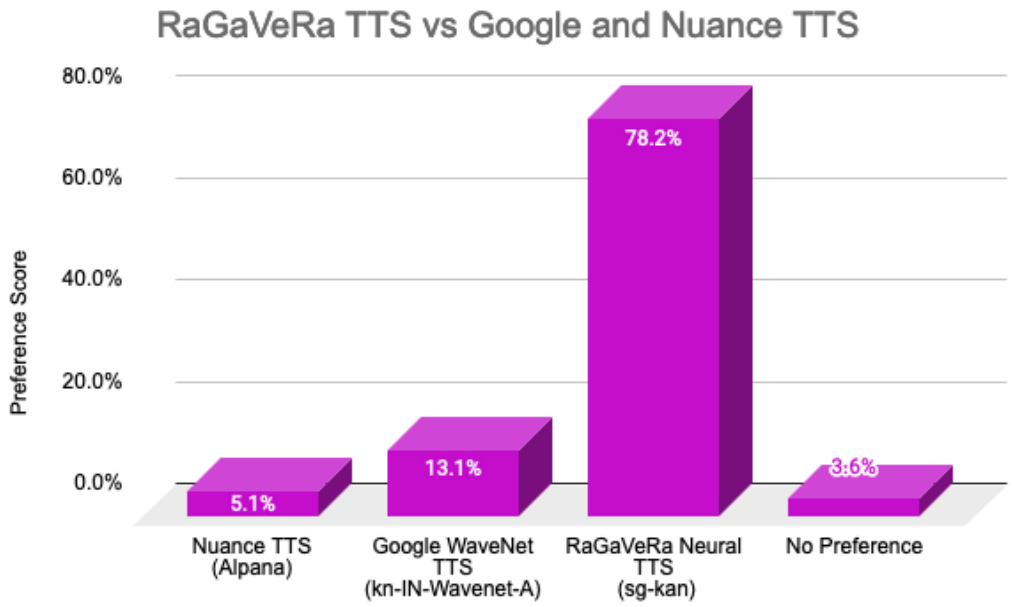


Fig. 4: Comparison of the quality of RaGaVeRa's TTS against Google's WaveNet and Nuance's Kannada TTS as assessed by 55 natives of Kannada. RaGaVeRa's TTS got a mean preference score of 78.2% in contrast to 13.1% for Google's TTS and 5.1% for Nuance's TTS.

IV. Conclusion

We have designed, developed and tested a state-of-the-art speech synthesis system for Kannada. Some of the speech samples synthesized with our TTS are available on the RaGaVeRa website [30]. The output speech generated by our TTS was rigorously evaluated for quality in three different types of experiments by over one hundred people with good reading, writing and speaking proficiency in Kannada. The mean preference scores obtained from fifty five human evaluators clearly establish that the output quality is felt superior to that of Google's WaveNet TTS and Nuance's Kannada TTS. as well as the original speech of the speaker. The MOS of 4.51 obtained by our TTS as against the score of 4.62 for the original utterances of the speaker makes RaGaVeRa's Kannada TTS eligible to be called human-like quality.

We shall be exploring transfer learning for adapting the developed system to the sister Dravidian languages, namely Tamil, Telugu and Malayalam. We also intend to generate voice of the same person in languages unknown to her, by first training the model on the target language speech by a native speaker of that language and then transfer learning using Kannada speech data from the intended speaker. Further, we will be exploring synthesis of emotional speech [31] by employing limited speech data recorded with different emotions. We also intend to explore the possible benefits of using a grapheme-to-phoneme converter [32] as a preprocessing module for the input text. Other work in the planned pipeline is high quality synthesizers for Hindi, English with Indian accent and Hinglish.

Acknowledgment

The authors thank the Department of Information Technology and Biotechnology, Government of Karnataka for partly funding the development of this TTS through a grant to RaGaVeRa Indic Technologies Pvt. Ltd. as one of the winners based on a state-wide, multilevel selection of hundred promising startups [33] under the the Elevate2019 initiative. We also thank the speaker, Mrs. Suma Gurumurthy for her clear voice and pronunciation. Thanks are also due to SID, IISc for the incubation. Of course, immense thanks to all the Kannada enthusiasts, who responded to our request over whatsapp and evaluated our TTS against other TTS and also the original voice. Finally, Ramakrishnan A G wants to record his thanks to Prof. D K Subramanian and Prof. N J Rao, who encouraged him two decades ago to actively pursue work on technologies for Indian languages. Thanks are also due to his many research and M Tech students, Ms. Kalika Bali, Dr. R N V Sitaram and Dr. Ksenia from whom he learnt different aspects of linguistics and speech processing.

References

[1] Rama, GL Jayavardhana, A. G. Ramakrishnan, R. Muralishankar, and R. Prathibha. "A complete text-to-speech synthesis system in Tamil," In Proc. 2002 IEEE Workshop on Speech Synthesis, pp. 191-194. IEEE, 2002.

[2] K. Partha Sarathy, A.G.Ramakrishnan, “A research bed for unit selection based text to speech synthesis”, Proc. II IEEE Spoken Language Technology (SLT) workshop, Goa, India, Dec 15 - 18, 2008.

[3] Shiva Kumar H R, Ashwini J K, Rajaram B S R and A G Ramakrishnan, “MILE TTS for Tamil and Kannada for blizzard challenge 2013,” Proc. of Blizzard Challenge Workshop, Barcelona, Spain, Sept. 3, 2013.

[4] B S R Rajaram, H R Shiva Kumar, and A G Ramakrishnan, ”MILE TTS for Tamil for Blizzard challenge 2014”, In Blizzard Challenge Workshop, vol. 2014.

[5] R. Muralishankar, A. Vijay Krishna and A. G. Ramakrishnan, “Subspace based Vowel Consonant Segmentation,” Proc. IEEE Workshop on Statistical Signal Processing, Sept 28 – Oct 1, 2003, St. Louis, Missouri, pp. 589- 592.

[6] R. Muralishankar, A. G. Ramakrishnan and P. Prathibha, ”Modification of Pitch using DCT in the Source Domain,” Speech Communication, 2004, Vol. 42/2, pp. 143-154.

[7] Vikram Ramesh Lakkavalli, Arulmozhi P. and A. G. Ramakrishnan, “Continuity metric for unit selection based text-to speech synthesis,” IEEE International Conference on Signal Processing & Communications (SPCOM 2010), 2010.

[8] R. Murali Shankar, A. G. Ramakrishnan and Lakshmish N Kaushik, “Time Scaling of Speech using Independent Subspace Analysis”, Proc. INTERSPEECH 2004 – 8th Intern. Conf. Spoken Language Processing, Oct 4 – 8, 2004, Vol 3, pp. 2465 – 2468.

[9] Sridhar Krishna N, Partha P Talukdar, Kalika B, AG Ramakrishnan, ”Duration Modeling for Hindi Text-to-Speech Synthesis”, Proc. VIII Intern Conf Spoken Lang. Processing (INTERSPEECH 2004 - ICSLP), Jeju Island, Korea, Oct 4-7, 2004.

[10] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous, ”Tacotron: Towards end-to-end speech synthesis,” in INTERSPEECH, 2017.

[11] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui. Wu, ”Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018.

[12] Sercan O Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al., ”Deep voice: Real-time neural text-to-speech,” arXiv preprint arXiv:1702.07825, 2017.

[13] R. Prenger, R. Valle, and B. Catanzaro, ”PyTorch implementation of Natural TTS Synthesis By Conditioning WaveNet On Mel Spectrogram Predictions,” https://github.com/NVIDIA/tacotron2

[14] Ankur Debnath, Gangotri Nadiger, Shridevi S Patil, Ramakrishnan A. G., ”Low-Resource End-to-end Sanskrit TTS using Tacotron2, WaveGlow and Transfer Learning,” Proc. IEEE 17th India Council International Conference (INDICON 2020).

[15] GL Jayavardhana Rama, AG Ramakrishnan, M Vijay Venkatesh, R Murali Shankar, ”Thirukkural-A Text-to-Speech Synthesis System,” In Proc. Tamil Internet 2001, Kuala Lampur, Malaysia, Aug 26-28, 2001.

[16] Abhijit Pradhan, Anusha Prakash, S Aswin Shanmugam, GR Kasthuri, Raghava Krishnan, Hema A Murthy, ”Building speech synthesis systems for Indian languages,” In Proc. 21st National Conference on Communications (NCC), 2015.

[17] Majji Sreekanth, A. G. Ramakrishnan, ”Festival based maiden TTS system for Tamil language,” in Proc. 3rd Language and Technology Conference, 2007.

[18] Sinno Jialin Pan and Qiang Yang, ”A survey on transfer learning,” IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345-1359, 2010.

[19] Jia, Ye, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. ”Transfer learning from speaker verification to multispeaker text-to-speech synthesis.” In Advances in neural information processing systems, pp. 4480-4490. 2018.

[20] W. Fang, Y.-A. Chung, and J. Glass, “Towards transfer learning for end-to-end speech synthesis from deep pre-trained language models,” arXiv preprint arXiv:1906.07307, 2019.

[21] Tits, Noé, Kevin El Haddad, and Thierry Dutoit. ”Exploring transfer learning for low resource emotional TTS.” In Proceedings of SAI Intelligent Systems Conference, pp. 52-60. Springer, Cham, 2019.

[22] R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: A flow-based generative network for speech synthesis,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3617–3621.

[23] NVIDIA’s Tacotron2 code. https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/SpeechSynthesis/Tacotron2. Last accessed 22 April 2021.

[24] R. Prenger, R. Valle, and B. Catanzaro, ”WaveGlow: a Flow-based Generative Network for Speech Synthesis,” https://github.com/NVIDIA/waveglow

[25] P.-c. Hsu, C.-h. Wang, A. T. Liu, and H.-y. Lee, “Towards robust neural vocoding for speech generation: A survey,” arXiv preprint arXiv:1912.02461, 2019.

[26] K. Panchapagesan, Partha Pratim Talukdar, N. Sridhar Krishna, Kalika Bali, A. G. Ramakrishnan, ”Hindi Text Normalization,” Proc. Fifth International Conference on Knowledge Based Computer Systems (KBCS), pp. 19-22, 2004.

[27] Pretrained Tacotron2 model. https://ngc.nvidia.com/catalog/models/nvidia:tacotron2pyt_fp16.

[28] Pretrained WaveGlow model. https://ngc.nvidia.com/catalog/models/nvidia:waveglow256pyt_fp16/files?version=2.

[29] EH Rothauser, ”IEEE recommended practice for speech quality measurements,” IEEE Trans. on Audio and Electroacoustics, vol. 17, pp. 225-246, 1969.

[30] Samples of synthesized speech available at https://www.ragavera.com/tts/sg-kan-samples

[31] R Muralishankar and AG Ramakrishnan, ”Synthesis of Speech with Emotions”, Proc. Int. Conf Commn. Computers and Devices, Kharagpur, Dec. 14-16, 2000, pp. 767-770.

[32] A. G. Ramakrishnan and M Laxmi Narayana, ”Grapheme to Phoneme Conversion for Tamil Speech Synthesis,” In Proc. Workshop in Image and Signal Processing (WISP-2007).

[33] Elevate 2019 Startup winners: Full list of 100 winners announced by Karnataka Govt. https://indianexpress.com/article/cities/bangalore/elevate-2019-100-startups-full-list-of-winners-announced-karnataka-govt-5869522/