Proceedings of the EURALI Workshop @LREC2020, pages 70–74
Marseille, 20 June 2022

Exploring Transfer Learning for Urdu Speech Synthesis

Sahar Jamal, Sadaf Abdul Rauf and Qurat-ul-ain Majid
Fatima Jinnah Women University, Pakistan,
{sahar.syed,sadaf.abdulrauf}@gmail.com,quratulain@fjwu.edu.pk

Abstract

Neural methods in Text to Speech synthesis (TTS) have demonstrated momentous advancement in terms of the naturalness and intelligibility of the synthesized speech. In this paper we present neural speech synthesis system for Urdu language, a low resource language. The main challenge faced for this study was the non-availability of any publicly available Urdu speech synthesis corpora. Urdu speech corpus was created using audio books and synthetic speech generation.To leverage the low resource scenario we adopted transfer learning for our experiments where knowledge extracted is further used to train the model using a relatively smaller Urdu training data set. The results from this model show satisfactory results, though a good margin for improvement exists and we are working to improve it further.

Keywords: Neural TTS, Urdu text to speech synthesis, Deep Neural Networks, Transfer Learning

1. Introduction

Speech tools are imperative in the current era of voice driven interfaces. Speech synthesis is one such artificial speech production technology, which transforms text into intelligible speech (Holmes, 1973). With massive progress in high resource languages in speech synthesis, development of good quality TTS systems for low resource languages leaves much to be done.

The advent of neural approaches, especially Deep Neural Networks (DNN) proved to be revolutionary in overcoming limitations of previous approaches especially the inability to represent complex contextual dependencies (Ze et al., 2013). DNNs can model speech variations like speaking style and intonation even with limited data. (Qian et al., 2014a) report Deep Neural Networks outperforming HMMs with five hours of speech corpus by especially improving prosodical features. Multitask learning also considerably enhanced the efficiency of hidden representation, which in turn made the complex mapping possible. Prosody is the main feature which improves the performance of DNN in comparison to HMM (Qian et al., 2014b). The mapping is done directly between linguistic and acoustic features for each frame of the model (Wu et al., 2015).

A neural network based parametric system (Wang et al., 2016) eliminated the need of laborious alignment procedure by integrating the text and acoustic model. Mel-frequency spectrograms were used to connect two key modules of their Neural TTS. First, a network which acted as a predictor for a sequence of mel spectrogram frames for a given input and Wavenet with a few modifications(Shen et al., 2018). The MOS obtained for this experiment came quite close to the score reported by professionally synthesized speech. Standard methods used for TTS required considerable time and effort. The Neural methods took over this task of feature engineering by the use of self operating feature learning (Mametani et al., 2019).

Deep voice (Arik et al., 2017) based on neural approaches made things simpler by minimizing human intervention while training the data. DeepVoice2 (Gibiansky et al., 2017) used less than half hour speech per speaker to provide a lot of variations in the generated synthesized speech. This was achieved by using a post-processing neural vocoder. The focus was on using less speech data compared to single speaker models while maintaining high quality output. Deepvoice 3 (Ping et al., 2017) while keeping the synthesized speech as natural as possible, reduced the training time to ten times from the standard models. On the other hand, they built up the model using about eight hundred hours of training speech data which is quite high comparative to those used in standard Neural TTS.

The language selected for this research, Urdu, is a low resourced language in field of TTS. However, it is a widely spoken language in South Asia. Presenting for DNN based approach for Hindi-Telugu language pair promising results were credited to the ability of DNN in predicting spectral parameters. The prediction causes reduction in the noise and artificiality of synthesized speech (Reddy and Rao, 2018). Using a Hindi model for text normalization, models were built for Bangla language using Long Short Term Memory RNN (Gutkin et al., 2016).

Tacotron (Wang et al., 2017) by Google is an end to end model which does automatic feature extraction. Extending it, (Jia et al., 2018) presented a multiple speaker model based on three separate modules. Speaker verification module generates a vector extracted from few seconds of speech by seen or unseen speaker. The second module maps the text to phonemes while using pre-trained speaker embeddings. Lastly, the vocoder generates the wave-forms.

We discuss in section 2, the original model architecture along with the transfer learning approach we used.

Corpus	Hours	Lines	Size	Gender	Source
Urdu					
FS1	4.5	2807	694 MB	F	Google TTS
MR1	4.6	4841	732MB	M	Youtube
MR2	11.6	11,296	1.8 GB	M	Youtube
FR1	1	631	128 MB	F	AudioBook
English					
LJ Speech	24	13,100	2.7 GB	F	Public Domain
Arabic					
Nawar	3.7	1813	1.4 GB	M	Public Domain

Table 1: Corpus description

In section 3, we briefly list the Urdu corpora and other resources used in this research along with experiments conducted with these resources and their results. Section 4 discusses the evaluation metrics used for the analysis before concluding with the future prospects in Section 5.

2. Model architecture

We used Tacotron (Wang et al., 2017) for building our systems. Tactoron is an end to end system which synthesizes speech directly from the text. We built Urdu standalone systems and transfer learned system using pre-trained models of English and Arabic as parent models. The models use a training batch size of 32 with an initial learning rate of 0.002. We used feed forward neural networks with input as a predictor for the vocoder parameters using multiple layers of hidden units (Wu et al., 2016).

The original model has three basic components. The *encoder* gives sequential representations of the input text and uses the scheme by (Lee et al., 2017) for feature extraction. A bottleneck layer is used for better generalization and convergence by using a compressed representation with reduced dimensions. The bottleneck layer is basically a pre-net with dropout mechanism. It helps to focus more on the input text. The CBHG (1-D convolution bank + highway network + bidirectional GRU) module is the building block used for feature extraction in the encoder. The *decoder* models the mel-scale spectrogram representing the relation between text and speech. The *post processor* not only fixes errors of decoder predictions but also brings it into a form which can be then converted into wave-forms. Griffin-Lim is used to generate the final outputs in form of waveform audios.

Pre-trained models for English built on LJ Speec data [1] was used to to initialize our models for transfer learning. For Arabic we built our own model using the same parameters as our Urdu models.

[1] `https://data.keithito.com/data/speech/tacotron-20180906.tar.gz`

2.1. Transfer learning

Transfer learning uses knowledge obtained from performing a task to achieve another related task. The knowledge transfer from one task to another task can contribute to learning by improving the target model, accelerating learning, increasing efficiency or decreasing required time (Torrey and Shavlik, 2009). Use of large existing data as an additional set of data was suggested by (Dai et al., 2007) to augment new smaller set of data. The boosting algorithm promised good model accuracy using only small new data set while extracting possible knowledge from parent models. We used a similar approach to train our models by using Arabic and English as the parent models.

3. Experimental Setup and Corpora

We used deep neural networks for building our text to speech synthesizer. We employed transfer learning to cater for the data scarcity and also built standalone models to establish a comparison. The details of the experiments are given in Table 1.

3.1. English Corpus

We used LJ Speech Data (Keith Ito and Linda Johnson, 2017) to build the parent model for English. It consists of 13k single speaker utterances totalling to twenty four hours of speech in female voice. The sample rate used in this corpus is 22.05 kHz.

3.2. Arabic Corpus

Nawar (Halabi, Nawar and Wald, Mike, 2016) Arabic corpus , based on 3.7 hours of professionally recorded speech was used to train the Arabic model based on transfer learning.

3.3. Urdu Corpora

There was no publicly available Urdu speech corpus for TTS research. Due to unavailability of corpora, we built our own Urdu speech corpora using audio books and synthetic speech generation. The sampling frequency was fixed at 22.05 kHz for each corpus. A brief description of these corpora is given in Table 1 and details are given as follows:

Model	Lines	Duration	MOS naturalness	MOS intelligibility
Transfer Learning English ⟶ Urdu				
M1(LJSpeech ⇒ FS1)	13,100+2807	24+4.5 hrs	3.15	3.45
M2(LJSpeech ⇒ MR1)	13,100+4841	24+4.6 hrs	3.30	3.10
M3(LJSpeech ⇒ MR2)	13,100+11,296	24+11.6 hrs	3.40	3.30
Transfer Learning Arabic ⟶ Urdu				
M4 (Nawar ⇒ FS1)	1813+2807	3.7+4.5 hrs	3.00	2.80
Urdu Standalone				
M5 (FS1)	11,296	11.6 hrs	2.90	2.60

Table 2: Model description

3.3.1. FS1-Synthetic speech corpus

FS1 Urdu corpus (Sahar Jamal, 2020) includes transcriptions collected using multiple sources which including manual annotation, news web sources and texts from some publicly available corpus. The corresponding audio data sets were generated using Google Text-to-speech system (Google, 2017). The final data set consisted of 2807 utterances on single speaker female voice.

3.3.2. MR1-Male Audio book1

This corpus (Sahar Jamal, 2021c) was made by using an Urdu audio book. It uses native Urdu speaker male voice with a duration of 4.6 hours.

3.3.3. MR2-Male Audio book2

This Urdu (Sahar Jamal, 2021a) 11.6 hours speech corpus was created by splitting multiple Urdu audio books. It is a single speaker male voice.

3.3.4. FR1-Female Audio book

This Urdu speech corpus (Sahar Jamal, 2021b) is in female voice and consists of one hour of speech data. This corpus was created using Urdu audio book.

3.4. Experiments and Results

All the models were trained until convergence following (Wang et al., 2017). We started the training using our Urdu training data `FS1`, constituting four hours of speech, the model had bad alignments (as expected). This led us to explore transfer learning techniques.

Starting with pretrained English model trained on 24 hours LJ speech corpus (Keith Ito and Linda Johnson, 2017) `M1` was built by initialising the Urdu model using the synthetic `FS1` corpus. The learned parameters of pretrained English model were used to initialize training with Urdu speech corpus `FS1`. Several learning rates were tested before setting the final learning rate to 0.02, which gave the best performance. At 477k steps, the model started to show uniform alignments.

We proceeded with using the created corpora to train the model and observe their effect on the resulting speech. `M1`, `M2` and `M3` (Table 2) were trained through the transfer learning from the English model using different Urdu Corpora for fine tuning.

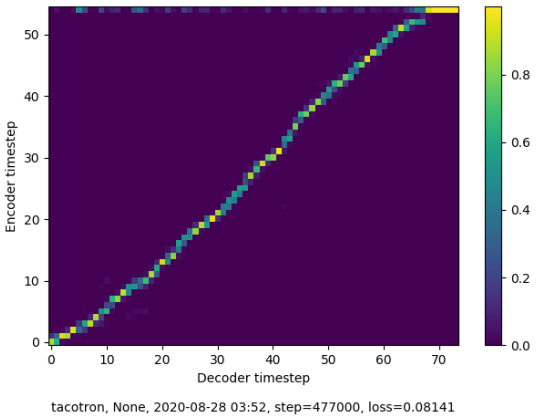

Figure 1: Attention alignment of Urdu TTS Model M1

Exploring the second language as parent, we used transfer learned from Arabic model with the synthetic `FS1` as fine-tuning corpus.

Thirdly, we built 11.6 hours Urdu stand alone TTS model `M5` following the original procedure.

4. Human Evaluation

Quality in TTS is a multidimensional term (Mariniak, 1993). The quality of synthesized speech is dependent on variable factors and parameters. The output quality is usually judged by a group of listeners (Jekosch, 1993).The evaluation is performed by observing the variation between the natural and synthesized speech.

For our study, forty subjects were selected who were native speakers of Urdu language. The subjects chosen for this experiment were provided an online form for submitting their opinions. Fifteen sentences were provided as sample and each participant was asked to listen to at-least five sentences before ranking the intelligibility and naturalness of speech. Each subject rated the synthesized speech from the rank of one to five for intelligibility and naturalness separately, these parameters are listed in table 4 with 1 being lowest and 5 being the highest quality.

Naturalness is a parameter which is hard to quantify. Different listeners participating in the experiment may have different preferences in selection of the most natural voice (Ojala, 2006). Mean opinion score (MOS)

was used to asses the opinion scores . MOS is used to evaluate the quality of speech. Five categories were created for the assessment of MOS. The categories are mentioned in table 4 and results are reported in table 2.

Rank	Intelligibility Criteria
1	Very Low intelligibility, No clarity
2	Low intelligibility,few parts are comprehensible
3	Average intelligibility
4	Overall intelligible with few distortions
5	Highly Intelligible
Rank	**Naturalness Criteria**
1	Highly robotic
2	Very robotic, few instances of naturalness
3	Average naturalness
4	Overall natural with traces of robotic instances
5	Very Natural Sounding

Table 3: Assessment categories to measure naturalness and intelligibility

Figure 2 demonstrates the results of the evaluation tests performed on the TTS Model M1. Out of the forty participants more than sixty percent found the synthesized speech of M1 intelligible.

For our model M1, the naturalness factor had more margin to improve as it was considered comparatively less satisfactory than intelligibility.This is also due to the synthetic voice used in the training corpus of M1. A better and bigger corpus may provide better results with the same approach. The main advantage of using the transfer learning approach is the elimination of hand engineered feature extraction.

This study tested five different model. English was used as a parent language for the first three models. Arabic was used as the parent language for the fourth model M4. The last model M5 was a standalone Urdu Model. The results show satisfactory naturalness for all the models with parent language as English. However, the intelligibility score was much better for model M1 which had English as parent language and Urdu Synthetic Speech Corpus as the second corpus. This was the most clean and noise free corpus we used in this research.

5. Conclusion

This research work presented approaches for building an effective Urdu Text to speech system which is a zero resource language in terms of corpus availability. We explored two approaches: 1) Building and using Urdu speech corpus of different sizes to train standalone models 2) Transfer learning from English vs. Arabic. Initializing the model by a larger non-Urdu corpus and further training it on a significantly Urdu corpus, we were able to get encouraging results. Although the results seem promising , however there is room for improvement. A larger and richer Urdu corpus will significantly contribute to better results.

6. Bibliographical References

Arik, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., et al. (2017). Deep voice: Real-time neural text-to-speech. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 195–204. JMLR. org.

Dai, W., Yang, Q., Xue, G.-R., and Yu, Y. (2007). Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 193–200, New York, NY, USA. Association for Computing Machinery.

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*, pages 2962–2970.

Google. (2017). Google tts api. `https://cloud.google.com/text-to-speech/`.

Gutkin, A., Ha, L., Jansche, M., Pipatsrisawat, K., and Sproat, R. (2016). Tts for low resource languages: A bangla synthesizer. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2005–2010.

Holmes, J. (1973). The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer. *IEEE transactions on Audio and Electroacoustics*, 21(3):298–305.

Jekosch, U. (1993). Speech quality assessment and evaluation. In *Third European Conference on Speech Communication and Technology*.

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Moreno, I. L., Wu, Y., et al. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in neural information processing systems*, pages 4480–4490.

Lee, J., Cho, K., and Hofmann, T. (2017). Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Mametani, K., Kato, T., and Yamamoto, S. (2019). Investigating context features hidden in end-to-end tts. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6920–6924. IEEE.

Mariniak, A. (1993). A global framework for the assessment of synthetic speech without subjects. In *Third European Conference on Speech Communication and Technology*.

Ojala, T. (2006). Auditory quality evaluation of present finnish text-to-speech systems. *Helsinki University of Technology*.

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J., and Miller, J. (2017). Deep voice 3: Scaling text-to-speech with

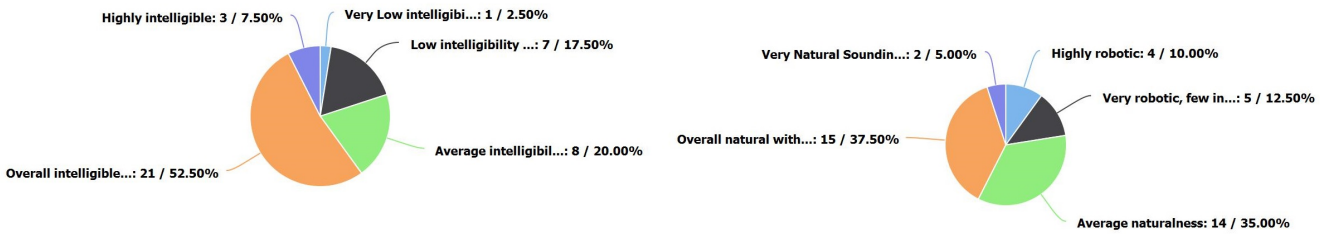

Figure 2: Left: Quality w.r.t Intelligibility for Model M1; Right: Quality w.r.t Naturalness for Model M1

convolutional sequence learning. *arXiv preprint arXiv:1710.07654*.

Qian, Y., Fan, Y., Hu, W., and Soong, F. K. (2014a). On the training aspects of deep neural network (dnn) for parametric tts synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3829–3833. IEEE.

Qian, Y., Fan, Y., Hu, W., and Soong, F. K. (2014b). On the training aspects of deep neural network (dnn) for parametric tts synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3829–3833. IEEE.

Reddy, M. K. and Rao, K. S. (2018). Dnn-based bilingual (telugu-hindi) polyglot speech synthesis. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1808–1811. IEEE.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., 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.

Torrey, L. and Shavlik, J. (2009). Chapter 11 transfer learning.

Wang, W., Xu, S., Xu, B., et al. (2016). First step towards end-to-end parametric tts synthesis: Generating spectral parameters with neural attention.

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*.

Wu, Z., Valentini-Botinhao, C., Watts, O., and King, S. (2015). Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4460–4464. IEEE.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Ze, H., Senior, A., and Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing*, pages 7962–7966. IEEE.

Halabi, Nawar and Wald, Mike. (2016). *Phonetic inventory for an Arabic speech corpus*. `http://en.arabicspeechcorpus.com/`.

Keith Ito and Linda Johnson. (2017). *The LJ Speech Dataset*. `https://keithito.com/LJ-Speech-Dataset/`.

Sahar Jamal. (2020). *SJ Urdu Synthetic Corpus*. `https://github.com/saharsyed/Sj-Urdu-Synthetic-Corpus`.

Sahar Jamal. (2021a). *Kutub Urdu Speech Corpus*. `https://github.com/saharsyed/kutub-Urdu-Speech-corpus`.

Sahar Jamal. (2021b). *SJ Kahani Speech Corpus*. `https://github.com/saharsyed/SJ-Kahani-Speech-Corpus`.

Sahar Jamal. (2021c). *Urdu Adab Speech Corpus*. `https://github.com/saharsyed/Urdu-Adab-Speech-Corpus`.