Interspeech 2018
2-6 September 2018, Hyderabad

An Investigation of Convolution Attention Based Models for Multilingual Speech Synthesis of Indian Languages

Pallavi Baljekar, SaiKrishna Rallabandi and Alan W. Black

Carnegie Mellon University

pbaljeka@cs.cmu.edu, srallaba@cs.cmu.edu, awb@cs.cmu.edu

Abstract

In this paper we investigate multi-speaker, multi-lingual speech synthesis for 4 Indic languages (Hindi, Marathi, Gujarathi, Bengali) as well as English in a fully convolutional attention based model. We show how factored embeddings can allow cross lingual transfer, and investigate methods to adapt the model in a low resource scenario for the case of Marathi and Gujarati. We also show results on how effectively the model scales to a new language and how much data is required to train the system on a new language.

Index Terms: End-to-end Synthesis, Convolutional Model, Neural Networks, Indic Languages.

1. Introduction

In the past year, we have seen great performance gains in the use of sequence-to-sequence attention based models for speech synthesis in English [1, 2, 3], and Japanese [4]. From these, convolutional based models as described in [5] and [6] have been gaining traction in synthesis because of shorter convergence times as shown in [2, 4]. Thus, in this paper, our goal is to study the performance of these convolution attention based models on multi-lingual speech synthesis. Our goal is two-fold. First, we would like to understand, given a trained multi-lingual model, what is the minimum amount of data required to train this model for the target speaker and target language. Second, we would like to understand how we can use these models in data scarce situations. Specifically, we would like to understand how best to augment these models with external data. This data could either be another higher resource language or data from other speaker(s) of the same language.

1.1. Previous Approaches

Previous approaches in using cross-lingual techniques in HMM based speech synthesis have been primarily explored in the direction of cross-lingual speech synthesis [7] or polyglot speech synthesis [8], where the goal is to map the speaker's characteristics from one language to another. Most of these techniques assume parallel corpora from the source and target speaker in order to learn a phone or state mapping as in [9, 10]. However, the goal in this work is to use cross-lingual resources to augment low-resource data in order to give better intelligible TTS. In this regard our goal is similar to the unsupervised cross-lingual techniques as in [11], and polyglot synthesis techniques mentioned in [7] as well as the work done in factored modelling approaches in [12] and [13].

In [11], they assume adaptation data to be coming from the same language as the language used to train acoustic models and so the new adaptation acoustic data is mapped into the phone-set of the source language. It then uses a two-pass decision tree clustering stage to add specific features in the adaptation data in the second stage. On the other hand, in [7], they build polyglot synthesis models without the need for parallel corpora. However, unlike both of these approaches, our primary goal is not to retain speaker characteristics, but rather to be able to synthesize intelligible speech in our target low-resource language, by augmenting with data from a more easily available higher-resource language.

Previous work in exploring factorized multilingual, multi-speaker neural models has been proposed in [12] and [14]. The model in [12] which in turn builds on top of the model proposed in [13], where data is factorized across speakers by having a separate speaker partitioned layer. In the model in [12], in addition to factorizing across speakers, it factorizes across languages as well. The model consists of language towers which represent different input languages as well as a mean tower, similar to the bias cluster in cluster adaptive training framework [15]. In their model, two sets of speaker-specific basis are learned, the language basis weights and the speaker basis weights for each input speaker. This model is attractive in a data scarce scenario since it allows us to share weights across speakers as well as making it easier to extend the model so that one can factorize across all the different input conditions. In addition, it can be extended to accommodate more languages and dialectical variations easily by adding a new tower. Thus, it allows us to explore various tying and weight sharing schemes between similar languages in case of lesser data, such as tying Assamese, Nepali and Bengali, while also allowing the use of a single model across all languages. Similarly the model in [14] also builds on this same multi-lingual, multi-speaker (MLMS) model for low-resource languages. However, both of these models still rely on frame-based linguistic features.

In this paper, we explore newer sequence-to-sequence neural style attention models for multi-lingual synthesis, which take as input a sequence of graphemes and produce a sequence of Mel-Spectral frames. To our knowledge, this has not been explored yet for multilingual synthesis of Indian languages.

In the next section, Section 2 we briefly describe the model used. Section 3 details the various experiments and their results and finally Section 4 ends with a conclusion and discussion for future experiments.

2. Model Details

The model used in this paper is a modified version of the convolutional model, with multi-step attention (dot-product attention at each decoder layer), as described in [5] which was used for machine translation. This model was adapted by Ping *et. al* in [2] for multi-speaker speech synthesis of English speech. Similar to most sequence-to-sequence models used for speech synthesis, it consists of four main blocks. The encoder block, which converts the input sequence of text units (phones or characters) into a sequence of linguistic embeddings, the attention block which produces a context vector at each time step in the sequence, a compact weighted representation of the linguistic

10.21437/Interspeech.2018-1869

input which is given to the decoder and post-net. The third block is the decoder, which uses the output from the encoder and the context vector of the attention block to predict a sequence of Mel frames, which are then converted to the waveform using the post-net block, which converts the sequence of Mel frames to linear spectrograms and synthesizes them using Griffin-Lim algorithm for phase reconstruction [16].

Our model is based on the model described in [2]. In addition to using position embeddings for each speaker, the factored version of our model also includes global embeddings such as speaker, language, and gender which is given to each decoder layer. In addition, similar to the implementation in [17], we use a guided attention multi task loss as described in [4]. Since ours is a grapheme based system, we use the UniTran SAMPA Table [18], to map Unicode characters to the SAMPA symbol-set and we use this set of input symbols as the input to the system.

3. Experiments

In this set of experiments we seek to answer mainly the following questions:

- **Factored Embeddings:** Is there any advantage to factoring the data across the global attributes and more importantly whether factoring these global attributes will help share data across speakers and languages in data scarce scenarios.
- **Scaling to a new language:** What is the minimum amount of data required in a new language to be able to synthesize from a trained multilingual model and does adding other speakers from the same language improve performance.
- **Transferability:** Are these global embeddings for gender, language and speakers transferable, i.e., can we make *AXB*, an Indian female speaker speak like a man and can we make *AUP*, a Marathi male speaker talk in a different language such as Hindi or Bengali.

3.1. Factored *vs.* Un-factored Embeddings

We first investigate the case for factoring global attributes such as speaker, language and gender as opposed to assigning each speaker-language pair its own embedding.

Data: The data used to train the multilingual model consisted of multilingual datasets provided by Hear2Read and available on Festvox [19] as well as the Blizzard datasets provided by the Deity project from government of India [20]. These datasets range from 30 mins to 9 hours per speaker recorded in relatively clean conditions. We have listed the amount of each speaker's data in the results table, Table 1.

Training: First we trained a multi-lingual, multi-speaker model on 7 speakers comprising Hindi, English, Marathi and Gujarathi. The model was initialized with weights trained from the VCTK corpus [21], since we found that this makes the model converge faster as opposed to starting each training cycle from scratch. For the unfactored model, each language-speaker pair was treated as a separate embedding, while, for the factored model, we factored global attributes of speaker identity, gender and language separately, so it could share similar speaker characteristics across languages, as well as use common language features across speakers.

Results: Table 1 describes the amount of training data used per language and speaker as well as the results obtained on a held-out test set (10%) of each speaker's data using the factored and unfactored model. The results are measured using Mel Cepstral Distortion after making each sequence equal length using

Table 1: *DTWMCD results on small multi-lingual multi-speaker Indic Datasets (Marathi, Hindi, Gujarathi) and English*

Language Speaker (Gender)	Time (hh:mm)	Unfactored (DTWMCD)	Factored (DTWMCD)
English AXB (F)	00 : 30	8.20 ± 0.44	8.75 ± 0.64
English SLP (F)	00 : 30	9.05 ± 0.58	9.90 ± 0.43
Gujarathi IITM2 (M)	09 : 06	6.54 ± 0.60	7.03 ± 0.49
Hindi AXB (F)	01 : 55	7.51 ± 0.40	8.57 ± 0.91
Hindi IITF (F)	04 : 35	7.09 ± 0.40	7.57 ± 0.65
Hindi IITM (M)	04 : 30	6.62 ± 0.20	7.52 ± 0.25
Marathi AUP (M)	00 : 27	6.38 ± 0.28	7.35 ± 0.33
Marathi IITM1 (M)	03 : 03	6.48 ± 0.21	7.09 ± 0.35
Marathi SLP (F)	00 : 31	9.19 ± 0.40	10.23 ± 0.40

Dynamic Time Warping (DTW), thus they are much higher than MCD results computed using ground-truth durations. From the results in Table 1 we see that the results of the unfactored model seem to be better in all cases than the factored model. This is interesting, and against our expectations. One reason the unfactored model might be better is that it allows the model to over-fit to the data better and this might be a good thing when synthesizing speech using end-to-end models.

Second, speaker *SLP* a female Marathi speaker has the worst performance in the multi-lingual model. Now it is surprising that male speaker *AUP*, who has the same amount of Marathi data as *SLP* performs much better. However, there is also another male Marathi speaker in the dataset with about 5 hours of data. So the question is, is it something about the speaker's characteristics and speaking style that make it a good voice for the neural network model to learn or can we improve the performance of *SLP*'s Marathi speech by augmenting it with another female Marathi speaker.

Finally, the performance on English speech is very bad for both *AXB* and *SLP*. This might again be because of the paucity of the English data in the dataset. The multilingual dataset has only about 1 hour worth of English data coming from two female speakers. However, since the model was transferred from a model trained on VCTK corpus, a multi-speaker English corpus, it begs the question, whether the neural network works better with similar languages and tends to forget its previous weights, as shown in [22] or can the problem be solved by adding more data to it. Thus to this end we look at one Indian female speaker *SLP*, and try to improve her performance on English speech.

3.2. Case Study: Marathi and English (*SLP*)

As mentioned in the previous section, we wanted to see if the model improves by adding additional external data from another speaker of the same language, or is it better to fine-tune the multilingual model on a single speaker of the target language. To this end, we present various experiments in Table 2 in trying to improve the performance of speaker *SLP*'s English as well as Marathi synthesized speech.

Data and Training: To the model described in Table 1, we either add more data from other speakers and retrain this model, or adapt it on our target data, which in this case is half hour of *SLP* Marathi speech, and half hour of *SLP* English speech. We also further fine-tune the data augmented model with *SLP*'s English and/or Marathi speech. To augment the models we use the 4 hours and 20 mins of female Marathi speaker (*IITF4*) and 6 hours of Indian English female (*IITF*).

Table 2: *DTWMCD results on SLP: Is adding more data better or can we get gains by adapting trained model on target data. Results on both Factored and Unfactored Models with and without Adaptation and Data Augmentation*

Adaptation / Data Added	**Unfactored (DTWMCD)**		**Factored (DTWMCD)**	
	SLP Marathi	SLP English	SLP Marathi	SLP English
Baseline Multingual Indic Model	9.19 ± 0.40	9.05 ± 0.58	10.23 ± 0.40	9.90 ± 0.43
Adapting base model on SLP	8.81 ± 0.43	8.46 ± 0.42	9.15 ± 0.35	8.77 ± 0.37
Adapting base model on SLP Marathi	8.80 ± 0.36	8.70 ± 0.38	9.28 ± 0.40	9.31 ± 0.42
Adapting base model on SLP English	8.90 ± 0.37	8.42 ± 0.39	9.16 ± 0.31	8.74 ± 0.34
External Marathi Female	9.36 ± 0.41	9.09 ± 0.49	10.49 ± 0.44	10.02 ± 0.53
External English Female	9.30 ± 0.43	9.13 ± 0.53	10.25 ± 0.36	10.08 ± 0.45
External Marathi + English Female	9.42 ± 0.40	9.29 ± 0.54	10.13 ± 0.44	9.88 ± 0.41
Adapting external Marathi female speech model on SLP Marathi	8.85 ± 0.42	8.74 ± 0.33	9.17 ± 0.41	9.07 ± 0.37
Adapting external English and Marathi female speech model on SLP	9.14 ± 0.42	8.70 ± 0.37	9.35 ± 0.40	8.84 ± 0.37

Results: Table 2 agrees with Table 3 in that the unfactored model's performance is much better than the factored model. Some observations from the results in Table 2 are as follows: Adaptation or fine-tuning the model on our target data seems to produce some improvements in quality of speech. Surprisingly, we did not see as much gain in adding more data. In fact, in some cases adding more other speaker's data actually degrades the performance very slightly. In addition, we see that fine-tuning the model after augmenting with more data is not as good as just fine-tuning alone. This again seems to show that over-fitting on our target dataset seems to be desirable for synthesis.

3.3. How does it Scale to a New Language and Many New Speakers

To the model mentioned in Section, 3.1 which produced results in Table 1, we added 7 more speakers and one more language Bengali. The details of the training data in terms of speaker, gender and hours of data are shown in Table 3. We wanted to see if the results degrade when adding more diverse data, showing the case for overfitting to improve results. The details of the data that have been added are shown in Table, 3.
]
Results: We see that the degradation in performance if at all is not much. In fact, some unfactored model results improve marginally for a few speakers. The results for most factored models degrade but only very marginally.

3.4. Case Study: *IITF4*-Gujarathi - How does the Model Perform on a New Language?

The next question we wanted to answer was how much data in a new language is good enough to adapt the multilingual model to the new target language and speaker. We also wanted to compare the performance improvements if any if we had other external speakers in the same language and whether similarity of the speakers in terms of gender mattered, if at all. First we looked at how augmenting with data in the same language but from other speakers affects performance and how much of a role gender of the external speaker plays in improving results. These are described in Table 5.

Data: The baseline single speaker model was trained on English audiobook data of 23 hours (*LJSpeech* dataset) [23] and then fine-tuned on our target Gujarathi female speaker *IITF4*. Our baseline multilingual model which we abbreviate MLMSI-NG or Multi-lingual, Multi-speaker, Indic model with non Gujarathi data includes all of the subsets from Table 3, without

Table 3: *DTWMCD results on larger multi-lingual multi-speaker Indic Datasets (Marathi, Hindi, Gujarathi and Bengali) and English*

Language Speaker (Gender)	**Time (hh:mm)**	**Unfactored (DTWMCD)**	**Factored (DTWMCD)**
Bengali IITF3 (F)	01 : 34	7.30 ± 0.39	7.64 ± 0.39
Bengali IITM3 (M)	08 : 01	6.79 ± 0.56	7.28 ± 0.64
English AXB (F)	00 : 30	8.08 ± 0.47	8.70 ± 0.72
English SLP (F)	00 : 30	9.46 ± 0.65	9.95 ± 0.54
English IITF (F)	06 : 14	7.78 ± 0.43	9.17 ± 0.42
English IITM (M)	06 : 00	7.93 ± 0.47	9.00 ± 0.49
Gujarathi AD (M)	00 : 54	6.74 ± 0.37	7.71 ± 0.77
Gujarathi DP (F)	00 : 48	7.45 ± 0.58	9.21 ± 0.87
Gujarathi IITM2 (M)	09 : 06	6.85 ± 1.02	8.27 ± 0.66
Gujarathi IITF2 (F)	05 : 07	6.79 ± 0.48	7.52 ± 0.63
Gujarathi KT (F)	00 : 26	6.44 ± 0.39	9.36 ± 1.44
Hindi AXB (F)	01 : 55	7.47 ± 0.39	8.65 ± 0.49
Hindi IITF (F)	04 : 35	6.96 ± 0.38	7.72 ± 0.58
Hindi IITM (M)	04 : 30	6.48 ± 0.22	8.32 ± 0.32
Marathi AUP (M)	00 : 27	6.34 ± 0.30	7.68 ± 0.49
Marathi IITF4 (F)	04 : 19	7.32 ± 0.44	8.41 ± 0.40
Marathi IITM1 (M)	03 : 03	6.38 ± 0.20	7.35 ± 0.54
Marathi SLP (F)	00 : 31	9.62 ± 0.55	10.37 ± 0.48

the Gujarathi speakers. To this model we add an external Gujarathi male (*IITM2*) dataset of 9 hours which we call (Big) and a smaller male dataset is that of speaker *AD* totalling just under 1 hour of data. Similarly the external female data which is not the target speaker is Gujarathi speaker *DP*, including just less than an hour (50 mins) of data. Finally, the multi-speaker Gujarathi corpus includes 4 other Gujarathi speakers, *AD*, *DP*, *IITM2*, and KT apart from the target speaker, totalling about 16 and half hours of total Gujarathi data. Table 5 presents these results.

Results: We see that augmenting the model with external data at least for the case of Gujarathi, does not give expected gains in synthesis quality. As expected, there are some gains when augmenting with a speaker of the same gender, but if we compare the quality of synthesis obtained from adding 9 hours of external male Gujarathi data *vs.* adding only 1 hour of external male Gujarathi data, we do not see much difference in

Table 4: *Performance of the model computed in terms of DTWMCD, on subsets of a new language (Gujarathi-IITF4) indicated in each column as No. of utterances (spoken time of each corpus)*

Data	**100 (00:37:21)**	**250 (01:26:40)**	**500 (02:52:42)**	**1104 (05:07:29)**
Model adapted from single speaker English model	8.10 ± 0.42	8.39 ± 0.43	8.13 ± 0.43	7.90 ± 0.38
Model Adapted from non Gujarathi Multilingual Indic Model	7.20 ± 0.76	6.97 ± 0.57	6.90 ± 0.54	7.96 ± 0.50
Model Adapted from Multilingual Indic Model including multiple Gujarathi speakers	8.50 ± 0.74	8.17 ± 1.26	8.14 ± 0.84	7.52 ± 0.63

improvement. One reason for this might be that 5 hours of target speaker data is enough to train the model and it is not getting any gains from the other speakers of the same language. Thus, in the next experiment we wanted to see whether this had any effect if we reduced the data.

Table 5: *DTWMCD results introducing a new speaker (IITF4) in a new language (Gujarathi), and augmenting with different types of external data in another language or the same language but from an external speaker*

External Data Added	**Hours-Guj. Data (hh:mm)**	**DTWMCD**
Adapt on English model (Baseline)	05 : 07	7.90 ± 0.38
MLMSI-NG	05 : 07	7.96 ± 0.50
MLMSI-NG + Guj. male (Big)	14 : 13	7.96 ± 0.53
MLMSI-NG + Guj. male (Small)	06 : 01	7.90 ± 0.55
MLMSI-NG + other Guj. female	05 : 55	6.96 ± 0.69
MLMSI-NG + Multispeaker Guj.	16 : 34	7.52 ± 0.63

Data and Training: To understand the effect of the amount of data needed to make a model synthesize a new speaker's voice in a new language, we made 4 subsets of the female Gujarathi (*IITF4*) data. These included subsets of 100, 250, 500 and 1104 utterances each with durations as mentioned in Table 4, going from about half hour of data to almost five hours, and doubling in size with each subset. We assumed three different adaptation/initialization conditions in training these models. In the first case, we directly adapted the model on the target speaker, by transferring weights from English audiobook data trained model on *LJSpeech*. In the second case, we trained a multilingual model with all of the speakers mentioned in Table 3 except the other Gujarathi speakers (*AD*, *DP*, *KT* and *IITM2*). For the last model, we trained with all of the speakers from Table 3.

Evaluation: For calculating the DTWMCD, we use a 10% held out test-set from the original speaker. Table, 4 lists these results. Table, 4 lists these results. From Table 4 we see that we obtain some marginal improvements in increasing the data from 100 utterances to 500, however, we get a very surprising result in that, in some cases, increasing data seems to hurt performance especially when using the entire dataset. One reason might be that the dataset includes some noisy sentences on which the model fails. In the future, we would like to investigate this more, by getting rid of the worst 10% of data and re-running these set of experiments to understand why we see a performance degradation when we double the data.

3.5. Transferability

One advantage that we get with using factored embeddings is that it allows more control on the gender, speaker and language. Thus, with this model it is possible to get a speaker to speak a language for which we do not have any training data. We did some cross-lingual and cross-gender synthesis. The results can be found at this link [1]. As you can hear, transfer from one male *IITM2* speaking Gujarathi, to another male of a close language AUP speaking Marathi give relatively good results when we synthesize *AUP* speaking Gujarathi or Bengali. However, the results are not so good if we try to transfer gender or to synthesize *AUP* as a Hindi female. It still sounds like AUP and only very slightly changes from the Male version. We found that it is easiest to transfer in similar languages across gender, while it is difficult to change gender, since it is difficult to factorize the gender from speaker-ID.

4. Discussion and Conclusions

In this paper, we present our first preliminary work exploring convolution attention based models for multi-speaker, multilingual speech synthesis of Indian languages. We explored factored embeddings and found that factoring embeddings across speaker, gender and language actually degrades performance. We also explored how the model scales to many more speakers and new languages and found that this degradation is pretty marginal if any. In addition, we explored how much data one needs to synthesize a new language, given a trained model and if augmenting it with other speakers from the same language helps. Counter intuitively we found that it is better to fine-tune the model on the target data rather than augment it with more data. In addition, we found that the results do not improve and in some cases degrade when adding more data for the case of one Gujarathi female speaker. Our hypothesis is that this might be due to some utterances which fail and it might be useful to explore data selection strategies to chose a good subset to train the model with.

Acknowledgement *This research was funded in part by the Google Faculty Research Award and the Amazon AWS award.*

5. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: A fully end-to-end text-to-speech synthesis model," in *Proceedings INTERSPEECH*, 2017.

[2] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," in *Proceedings International Conference on Learning Representations (ICLR)*, 2018.

[1] http://tts.speech.cs.cmu.edu/pbaljeka/www/IS2018_wavs

[3] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in *Proceedings International Conference on Learning Representations (ICLR)*, 2017.

[4] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," *Proceedings International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[5] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings International Conference on Machine Learning (ICML)*, 2017.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 6000–6010.

[7] X. Peng, K. Oura, Y. Nankaku, and K. Tokuda, "Cross-lingual speaker adaptation for HMM-based speech synthesis considering differences between language-dependent average voices," in *Proceedings IEEE 10th International Conference on Signal Processing*. IEEE, 2010, pp. 605–608.

[8] C.-P. Chen, Y.-C. Huang, C.-H. Wu, and K.-D. Lee, "Polyglot speech synthesis based on cross-lingual frame selection using auditory and articulatory features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1558–1570, 2014.

[9] Y.-J. Wu, S. King, and K. Tokuda, "Cross-lingual speaker adaptation for HMM-based speech synthesis," in *Proceedings 6th International Symposium on Chinese Spoken Language Processing,(ISCSLP)*. IEEE, 2008, pp. 1–4.

[10] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis." in *Proceedings INTERSPEECH*, 2009.

[11] M. Gibson, "Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.

[12] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis," in *Proceedings INTERSPEECH 2016*, 2016, pp. 2468–2472.

[13] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proceedings International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4475–4479.

[14] A. Gutkin, "Uniform multilingual multi-speaker acoustic model for statistical parametric speech synthesis of low-resourced languages," in *Proceedings INTERSPEECH*, 2017, pp. 2183–2187.

[15] V. Wan, J. Latorre, K. Yanagisawa, M. Gales, and Y. Stylianou, "Cluster adaptive training of average voice models," in *Proceedings International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 280–284.

[16] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[17] R. Yamamoto, "DeepVoice3 PyTorch Implementation," https://github.com/r9y9/deepvoice3_pytorch, December 2017.

[18] T. Qian, K. Hollingshead, S.-Y. Yoon, K.-Y. Kim, and R. Sproat, "A Python toolkit for universal transliteration." in *Proceedings Language Resources and Evaluation Conference (LREC)*, 2010.

[19] "Festvox Indic Voices provided by Hear2Read," http://festvox.org/cmu_indic/, 2017.

[20] A. Baby, A. L. Thomas, N. N. L, and T. Consortium, "Resources for Indian languages," in *CBBLR Workshop, International Conference on Text, Speech, and Dialogue - 2016*, vol. 978-80-263-1084-6. Tribun EU, 2016, pp. 37–43.

[21] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017.

[22] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[23] K. Ito, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.