PalArch's Journal of Archaeology of Egypt / Egyptology

Speech-to-text conversion using GRU and one hot vector encodings

Manpreet Singh Bhatia,[1] Alok Aggarwal,[2] Narendra Kumar[3]

[1,2] School of Computer Science, University of Petroleum & Energy Studies, Dehradun, India
[3] Icfai University, Jaipur, India
saby4911@gmail.com,[1] alok.aggarwal@ddn.upes.ac.in,[2] drnk.cse@gmail.com[3]

Manpreet Singh Bhatia,[1] Alok Aggarwal,[2] Narendra Kumar[3] : Speech-to-text conversion using GRU and one hot vector encodings - Palarch's Journal Of Archaeology Of Egypt/Egyptology 17(9). ISSN 1567-214x.

Abstract

In this work an RNN based model with the gated recurrent unit (GRU) embedding is proposed to convert a raw speech audio into the speech. The method involves clearing of noise if any present in the audio and then extracting the speech from the audio and converting it to the text. Data collection is done manually by creating audio clips from microphone and taking samples of audio clips at a sampled frequency of 44100 Hertz with a sampling size of 1375. The proposed model is able to convert into text from various speakers with the different accents (acoustics). The input voice of the speaker can be from single speaker or multiple speakers and it is compared with the datasets of the voice of multiple speakers in the model. Proposed approach with GRU and RNN model produces good results having 87% accuracy on the test set which is better than the approaches like Char2Wav with 85% accuracy, GST with 82%, Listen, attend and spell with 85.9% and CTC attention based model with 85.4% accuracy. Though compared to Deep Voice 2 with Tecatron and RNN with EESEN based WFST accuracy of the proposed approach is somewhat lower level which gives accuracy of 93% and 90% respectively compared to accuracy of the proposed approach which is 87%.

Keywords: Speech-to-text conversion; Recurrent neural network (RNN); Gated recurrent unit (GRU); one hot vector

1. Introduction

Speech-to-text conversion has been one of the latest trends in the industry and many methods have been tried so far to convert speech to text for better understanding of what the speaker has said. Speech-to-text conversion is very useful where speaker and listener don't have knowledge of each other's native language. Speech-to-text recognition is important in the field of translations. Many companies have come up with their translation engines to translate the speech from a particular language to the language familiar to the user. This has made the life of the people easy who are often

travelling and they are now better able to understand and communicate through these powerful translation engines. Speech-to-text conversion is used in building translation engines. In acoustic modeling Hidden Markov Model (HMM)/Gaussian Mixture Model (GMM) are traditionally used for automatic speech recognition. To normalize the temporal variability HMM's are used. In recent years, introduction of deep neural networks (DNNs) has improved ASR performance dramatically as acoustic models [1-3]. In conversion of text-to-speech models major challenge of the algorithms is to produce a natural speaking style like sound from a given text input and vice versa. DNN's have led to improve the various components of this problem. Components of the speech recognition are; acoustic, pronunciation, language models. End-to-end speech recognition technique [1-6] converts speech to text without the requirement of pre-defined alignments between acoustic frames and characters and it produces better results than the traditional methods like DNN-HMM (Deep neural network Hidden Markov Models).

Artificial speech synthesis uses the text-to-speech conversion technique and has numerous applications in technology interfaces, entertainment, media and accessibility. It is commonly known as TTS. In most of these systems, a single speaker voice is used to build while multiple speaker voices are used to have the distinct model parameters or speech databases. Compared to a single voice system, developing an artificial speech synthesis system with multiple voices more data and development efforts are required. TTS approach has two major advantages; one being faster and secondly it is performed in a single step to convert the speech into text. Speech recognition has also been done by speaker adaptation techniques in [7]. The trained model parameters are modified to match the features of the speaker and vice versa. In most of the cases, matching of the features is done by the minimal amount of adaptation data. Various methods have been proposed so far like MAP [8], MLLR [9] which are designed to perform better with the small amount of data. However these techniques are limited to single speaker frequency and cannot be used for a database where multiple speakers voices are needed to be modulated.

For small vocabulary tasks, hybrid NN-HMM model [7] has been a choice of many researchers. The models where the requirement is to work on larger vocabulary, deep neural networks are becoming much more popular than other counterparts. The whole structure is replaced by the neural networks with simpler features with the use of traditional text-to-speech (TTS) pipeline. The text part is first converted to phoneme and then audio synthesis model is used to convert the linguistic features present in text for converting it into speech. Converting textual transcription to high quality speech required days to week time prior to the use of DNN's. The advantage of using DNN in speech synthesis has reduced the timeline of converting the text-to-speech from weeks to few hours of manual effort with just training the model. Due to the overall decrease in time and efforts for training the model, DNN's have become a success story and many researchers have started using DNN for statistical machine translations [10]. These models are also used in predicting the conditional probabilities of the target sequence. The prediction of the sequence also reduces the error in the speech

conversion and in predicting the occurrence of a word after a given word. In this way, nature of predicting the sequence after a given sequence makes TTS model a robust mechanism using DNN. The model uses encoder-decoder technique for predicting sequence. Various encoder-decoder systems have used LSTM's in the past but due to the problem of vanishing gradients, its use could not be very popular in the recent past.

In this work an RNN based model with the GRU embedding is proposed to convert a raw speech audio into the speech. The method involves clearing of noise if any present in the audio and then extracting the speech from the audio and converting it to the text. Data collection is done manually by creating audio clips from microphone and taking samples of audio clips at a sampled frequency of 44100 Hertz with a sampling size of 1375. The proposed model is able to convert into text from various speakers with the different accents (acoustics). The input voice of the speaker can be from single speaker or multiple speakers and it is compared with the datasets of the voice of multiple speakers in the model.

Rest of the paper is organized as follows. Section 2 gives related work done by earlier researchers in the said domain. Section 3 describes the one hot vector and GRU (Gated Recurrent Unit) which is the main core of our approach and proposed methodology. Results are discussed in section 4. Finally, section 5 concludes the work with future extension of the work.

2. Related Work

Various solutions to the speech recognition problem have been proposed for improving the efficiency both for single-speaker and multi-speaker speech synthesis. For single-speaker TTS system various approaches have been proposed like based on fundamental frequency prediction [11], duration prediction [12], sample audio waveform generation [13-14], Char2Wav [2], DeepVoice1 [15], Tacotron [16] etc. For multi-speaker TTS systems, various approaches have been proposed like extending neural artificial speech synthesis systems to handle multiple speakers [1], DNN-based systems [17], HMM based TTS synthesis [18], shared hidden representation among various speakers [19], DNN-based multi-speaker modeling [20].

Various solutions have been proposed to the speech recognition problem and to improve the efficiency. Andrew Gibiansky et.al [1] used DNN with post processing vocoder and improved Tacotron pipeline. A significant improvement is shown in audio quality by introducing an improved Tacotron with a post-processing neural vocoder. Demonstration of the proposed approach has been done on both Tacotron and Deep Voice 2 on two multi-speaker artificial speech synthesis system datasets. It is shown that within less than half an hour of training, a single neural artificial speech synthesis system can learn hundreds of unique voices and this is achieved with high quality of audio synthesis. Speaker identity is preserved almost perfectly. Jose Sotelo et al. [2] have proposed Char2Wav technique having two components namely, a reader and a neural vocoder. The reader contains an encoder-decoder model with attention mechanism. The encoder is a bidirectional RNN in which input is a text or

phonemes, while the decoder produces vocoder acoustic features. Neural vocoder generates raw waveform samples from intermediate representation which refers to a conditional extension of sample RNN. Yajie Miao et.al [3] proposed RNN model with EESEN based on WFST (weighted finite state transducers). It uses a recurrent neural network (RNN) for prediction of context independent targets in text (phonemes or characters). EESEN uses a generalized approach for decoding based on weighted finite state transducers (WFSTs).

Daisy Stanton et al. [4] have used global style tokens (GST) with Tacotron. In the proposed system weights or style embedding are used as virtual style labels with Tacotron. The proposed architecture consist of Text-Predicted Global Style Token (TP-GST) in which GST combination weights or style embedding are treated as virtual" speaking style labels within Tacotron. It is shown that the proposed model generates audio which has better energy variation as well as pitch compared to the two state-of-the-art baseline models. It is concluded that TP-GST models are able to preserve the identity of multi-speaker successfully and factorize speaker identity and speaking style. William Chan et al. [5] used two components namely listener and speller. The proposed model learns all the components of a speech recognizer jointly compared to traditional DNN-HMM models. System has two components; a listener and a speller where the former accepts filter bank spectra as input and later emits characters as outputs. No assumption is made between the characters independently and sequence of characters are produced by the network which improves LAS systems over previous end-to-end CTC models. Rescoring over the top 32 beams, it is shown that the proposed approach achieves 10.3% with language model while 14.1% without a dictionary or language model in terms of word error rate. For mitigating the alignment issue Suyoun Kim et al. [6] proposed an approach in which a joint connectionist temporal classification attention model is used within the multi-task learning framework. It is shown that the proposed approach has its advantage over both the CTC and attention-based encoder-decoder baselines as it gives 5.4-14.6% relative improvements with respect to character error rate. For speech recognition a hybrid NN-HMM model is proposed by Ossama Abdel-Hamid et al. [7]. The proposed model can adapt all speakers and small speaker codes based on joint learning in neural network. For updating NN speaker codes and weights it uses standard back propagation algorithm. Size of speaker code used for learning has been kept as a smaller one and a separate code is used for learning each speaker. Results show an overall reduction in phone error rate by 10%.

Jean Luc Gauvain et al. [8] proposed a framework in which HMM is used with GMM (Gaussian mixture models).In this framework segmented k-means algorithm and forward backward algorithm have been expended to estimate the MAP functions.

This algorithm was adaptive in nature. C.J. Leggetter et al. [9] proposed continuous density HMM for speaker adaptation. This method also uses forward backward algorithm for maximizing the likelihood of the adaptation. For supervised mode error reduction of 37% and for unsupervised mode 32% error deduction has been reported.

An RNN model based on encoder-decoder consisting of two neural networks is proposed by Fethi Bougares et al. [10]. Existing log linear models are used for computation of conditional probability which is used as a measure for improving the performance of the machine translation system. Sergey Ioffe et al. [21] proposed a method in which training time of deep neural nets is reduced considerably for speech recognition. During the training of the model non linearity saturates the learning of the parameters at an early stage and slows down the learning process of the model referred as internal covariate shift. To overcome this problem normalization of the mini batches for training is proposed. Another method for improving learning rate of the model was proposed by Diederik P. Kingma et al. [22] known as ADAM optimizer. Proposed approach is invariant to diagonal rescaling of the gradients and computationally efficient. Empirical results showed that ADAM works better than the other stochastic optimization methods. Most of the Neural nets nowadays uses ADAM optimizer for tuning hyper-parameters [23-24].

3. Proposed Approach

One hot vector:

One hot encoding is a method of converting categorical variables into a form for the ML algorithms as an input to perform better prediction from the model. The categorical value is the numerical value in the dataset. One hot encoder is in a binary form (0 or 1) of the categories and included as a feature vector to train the model where 0 represents absence of the feature while 1 as presence of the feature. One-hot encoding often indicates the state of a machine. A decoder is used for determining the state of the machine in case of binary or gray code. However, it is not the case with one-hot machine which does not need a decoder to predict the state of the machine. It is in the n^{th} state if and only if the n^{th} bit is high.

In the corpus or dictionary, One hot vector is same as the feature vector with features labeled as 0 or 1 with the corresponding word number in the vector from the dictionary. Feature vector contains all the words from the dictionary with their corresponding indexes in the dictionary and are stored at the same index in the feature vector. One hot vector is extracted from dictionary embedding from the feature vector.

GRU (Gated Recurrent Unit):

GRUs are designed for solving the vanishing gradient problem of a standard recurrent neural network. An update gate and a reset gate is used its operation. GRUs exhibit better performance than LSTM on certain smaller datasets. GRUs are faster to train and need fewer data to generalize [25]. GRU architecture is shown in figure 1.

Proposed Methodology:

Model architecture of the proposed speech-to-text conversion is shown in figure 2. Following approach have used for speech-to-text conversion.

1. Gather the dataset with the speech.

2. Generate some negative dataset from the given dataset by adding noise into the background.
3. Pass the dataset into the GRU model to label the dataset. GRU is used for labeling as well and to remove the vanishing gradient problem. Pass the labeled dataset to the RNN model to create One Hot vector.

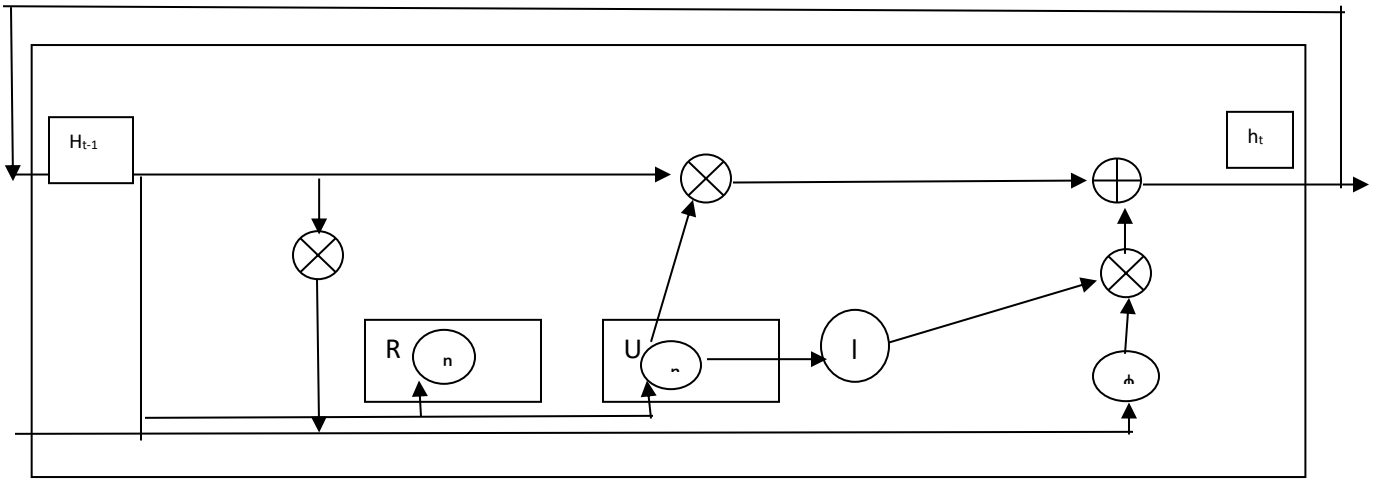

Fig. 1. GRU architecture

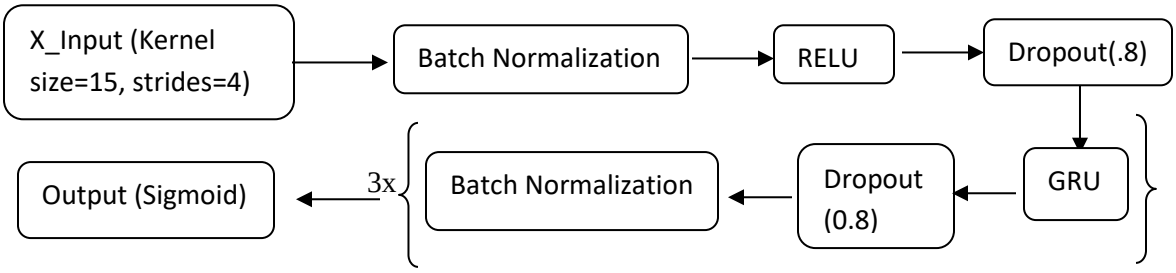

Fig. 2. Model architecture for the speech-to-text conversion

Once one hot vector is created, dataset is passed to the decoder part of the RNN model to convert the speech into the text format. This is the last step of the model and this will also separate noise from the speech and extract the speech and convert into text. Following points are noteworthy while observing and deriving the results:

1. The dataset containing both positive and negative examples are then tested for the prediction and found to be more accurate than the approaches discussed above.
2. The error rate of the characters reduced to 13%.
3. The model was able to convert into text from various speakers with the different accents (acoustics).
4. The input voice of the speaker can be from single speaker or multiple speakers and it is compared with the datasets of the voice of multiple speakers in the model.
5. The content of the speech delivered by the speakers is not related to any of the content in the database in the model.

Dataset Collection:

Data collection is done manually by creating audio clips from microphone and taking samples of audio clips at a sampled frequency of 44100 Hertz. Total length of the dataset is 1375. Dataset is not presented in the proposed work but however can be made available on the request to the author. The clip size is 10 seconds therefore, the total length of the dataset is (10 * 44100) numbers.

4. Results and Discussion

Proposed approach with GRU and RNN model produces good results having 87% accuracy on the test set which is better than the approaches like Char2Wav [2] with 85% accuracy, GST [4] with 82%, Listen, attend and spell [5] with 85.9% and CTC attention based model [6] with 85.4% accuracy. Though compared to Deep Voice 2 with Tecatron [1] and RNN with EESEN based WFST [3] accuracy of the proposed approach is somewhat lower level which gives accuracy of 93% and 90% respectively compared to accuracy of the proposed approach which is 87%.

Figure 3 shows the one hot encoding plot of sample audio. Frequency vs time plot for audio sample is shown in figure 4. Spectogram of audio sample and One hot encoding graph for audio sample are shown in figure 5 and 6 respectively. In the proposed approach, dataset is collected manually by creating recordings for 10 seconds audio clip with all the positive and negative samples. Positive samples are those where word is detected and negative samples where there is no word detected. The total samples collected are 1375 and sampled frequency is 44100 Hertz. In figure 3 & 6 one hot encoding graph spike in the graph indicates that there is a word detected in the audio clip and zero (flatten area) indicates that there is no word detected.

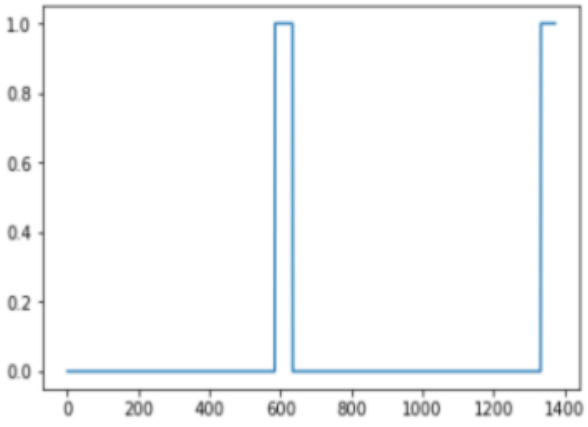

Fig. 3. One hot encoding plot of sample audio

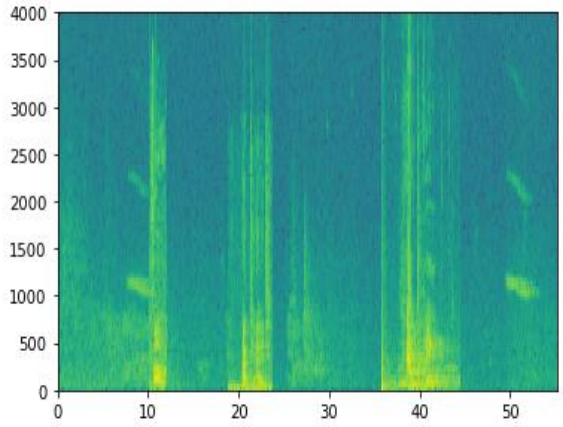

Fig. 4. Frequency vs time plot for audio sample

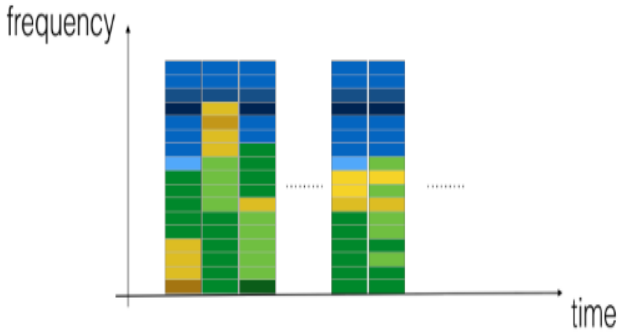

Fig. 5. Spectrogram of audio sample

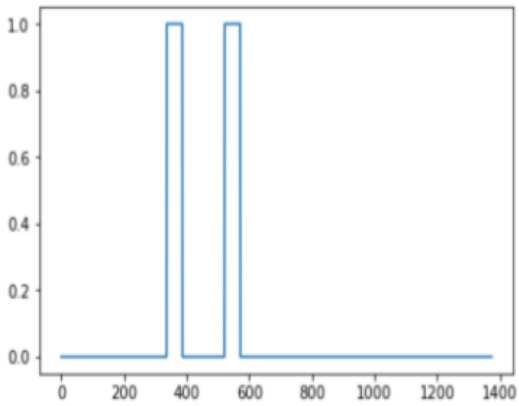

Fig. 6. One hot encoding graph for audio sample

In spectrogram of audio sample shown in figure 5, the color in the spectrogram shows the degree to which different frequencies are present (loud) in the audio at different points in time. Green means a certain frequency is more active or more present in the audio clip (louder). Blue squares denote less active frequencies. The dimension of the output spectrogram depends upon the hyper-parameters of the spectrogram software and the length of the input.

Table 1
Comparison in terms of efficiency of the proposed approach with other contemporary approaches.

Model	Efficiency
Proposed Approach (RNN with GRU model)	87%
Deep Voice 2 with Tecatron [1]	93%
RNN with EESEN based WFST [3]	90%
Char2Wav [2]	85%
GST [4]	82%
Listen, attend and spell [5]	85.9%
CTC attention based model [6]	85.4%

5. Conclusion

Proposed model works well with the voice of single speaker as well as multiple speakers. Model is able to distinguish users from different background having different accents (acoustics). RNN with GRU unit works well to recognize character from voices of the speakers saying it differently. Character error in recognition is reduced to 13% compared to some other contemporary approaches. Character recognition is a major step in any of the model while converting from speech-to-text

and most of the models discussed above have a significantly higher error in recognizing characters than the proposed model, making the proposed model a better one in recognizing characters from the speech and converting it back to the normal text from multi-lingual speakers.

For future purposes, this model can be extended with a working robot translating different languages to one language or this can be further extended by creating a real time live model by using sensors which works like a translation engine and helping the travelers to interact with the locals in their own language without learning their language. This can help in building a better relationship between locals and travelers and promote tourism industry all over the globe by creating an optimized product which benefits the society.

References

[1] A. Gibiansky, S. Ö. Arık, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, Y. Zhou, Deep voice 2: Multi speaker neural text to speech, 31st Conf. Neural Information Processing Systems (NIPS), Long Beach, CA, USA. (2017) 1-15.

[2] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, Y. Bengio, CHAR2WAV: End to end speech synthesis, Proc. Inter. Conf. Learning Representations (ICLR). (2017) 1-6.

[3] Y. Miao, M. Gowayyed, F. Metze, EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ. (2015) 167-174, https://doi.org/10.1109/ASRU.2015.7404790.

[4] D. Stanton, Y. Wang, R.J..S Ryan, Predictive Redicting Expressive Speaking style from text in End-to-end Speech Synthesis, Proc. IEEE Spoken Language Technology Workshop (SLT). (2019), https://arxiv.org/abs/1808.01410.

[5] W. Chan, N. Jaitly, Q. Le O. Vinyals, Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, 2016 IEEE Inter. Conf. Acoustics, Speech and Signal Processing (ICASSP), Shanghai. (2016) 4960-4964, https://doi.org/10.1109/ICASSP.2016.7472621.

[6] Kim, Suyoun, Takaaki Hori, Shinji Watanabe, Joint CTC-attention based end-to-end speech recognition using multi-task learning, 2017 IEEE Inter. Conf. Acoustics, Speech and Signal Processing (ICASSP). (2017) 4835-4839, https://doi.org/10.1109/ICASSP.2017.7953075.

[7] O. Abdel-Hamid, H. Jiang, Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code, 2013 IEEE Inter. Conf. Acoustics, Speech and Signal Processing, Vancouver, BC. (2013) 7942-7946, https://doi.org/10.1109/ICASSP.2013.6639211.

[8] J. Gauvain, Chin-Hui Lee, Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, IEEE Trans. Speech and Audio Processing. 2 (1994) 291-298, https://doi.org/10.1109/89.279278.

[9] C. J. Leggetter, P. C., Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Computer Speech and Language. 9 (1995) 71–185.

[10] K. Cho, B. v. M.¨enboer, C. Gulcehre, D. Bahdanau, F. B. H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP). (2014) 1724-1734, https://arXiv:1406.1078.

[11] S. Ronanki, O. Watts, S. King, G. E. Henter, Median-based generation of synthetic speech durations using a non-parametric approach, 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA. (2016) 686-692, https://doi.org/10.1109/SLT.2016.7846337.

[12] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, P. Szczepaniak, Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers form mobile devices. (2016), https://arXiv:1606.06061.

[13] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu. Wavenet: A generative model for raw audio. (2016), https://arXiv:1609.03499.

[14] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, Y. Bengio, Sample RNN: An unconditional end-to-end neural audio generation model, Proc. Inter. Conf. Learning Representations (ICLR). (2017) 1-11, https://arXiv:1612.07837.

[15] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, M. Shoeybi, Deepvoice: Real-time neural text-to-speech, Proc, ICML'17 34th Inter. Conf. Machine Learning. 70 (2017) 195-204.

[16] Y. Wang, RJ S.-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, R.A. Saurous, Tacotron: Towards end-to-end speech synthesis, Proc. Interspeech. (2017), https://arXiv:1703.10135.

[17] S. Yang, Z. Wu, L. Xie, On the training of DNN-based average voice model for speech synthesis, 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju. (2016) 1-6, https://doi.org/10.1109/APSIPA.2016.7820818.

[18] J. Yamagishi, T. Nose, H. Zen, Z. -H. Ling, T. Toda, K. Tokuda, S. King, S. Renals, Robust speaker-adaptive hmm-based text-to-speech synthesis, IEEE Trans. Audio, Speech and Language Processing. 17 (2009) 1208-1230.

[19] Y. Fan, Y. Qian, F. K. Soong, L. He, Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis, 2015 IEEE Inter. Conf. Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD. (2015) 4475-4479, https://doi.org/10.1109/ICASSP.2015.7178817.

[20] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, S. King, A study of speaker adaptation for DNN-based speech synthesis, Proc. Interspeech. (2016) 879-883, https://doi.org/10.21437/INTERSPEECH.2017-1038.

[21] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, ICML'15: Proc. 32nd Inter. Conf. Machine Learning. 37 (2015) 448-456, https://arXiv:1502.03167.

[22] D. Kingma, J. Ba., Adam: A method for stochastic optimization, Proc. International Conference on Learning Representations (ICLR). (2015).

[23] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, M. Shoeybi, Deep voice: Real-time neural text-to-speech, Proc. 34th Inter. Conf. Machine Learning. 70 (2017) 195–204.

[24] J. Bradbury, S. Merity, C. Xiong, R. Socher, Quasi-recurrent neural networks, In ICLR. (2017), https://arxiv.org/abs/1611.01576

[25] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, H.-M. Wang, Voice conversion from unaligned corpora using variational auto encoding wasserstein generative adversarial networks, Interspeech. (2017), https://*arXiv:1704.00849*, 2017.