INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

A Hybrid HMM-Waveglow based Text-to-speech Synthesizer using Histogram Equalization for Low resource Indian Languages

*Mano Ranjith Kumar M *, Sudhanshu Srivastava * , Anusha Prakash, Hema A Murthy*

Indian Institute of Technology, Madras

mano1381997@gmail.com, srivastava.rishabh4321@gmail.com, anushaprakash@smail.iitm.ac.in, hema@cse.iitm.ac.in

Abstract

Conventional text-to-speech (TTS) synthesis requires extensive linguistic processing for producing quality output. The advent of end-to-end (E2E) systems has caused a relocation in the paradigm with better synthesized voices. However, hidden Markov model (HMM) based systems are still popular due to their fast synthesis time, robustness to less training data, and flexible adaptation of voice characteristics, speaking styles, and emotions.

This paper proposes a technique that combines the classical parametric HMM-based TTS framework (HTS) with the neural-network-based Waveglow vocoder using histogram equalization (HEQ) in a low resource environment. The two paradigms are combined by performing HEQ across mel-spectrograms extracted from HTS generated audio and source spectra of training data. During testing, the synthesized mel-spectrograms are mapped to the source spectrograms using the learned HEQ. Experiments are carried out on Hindi male and female dataset of the Indic TTS database. Systems are evaluated based on degradation mean opinion scores (DMOS). Results indicate that the synthesis quality of the hybrid system is better than that of the conventional HTS system. These results are quite promising as they pave way to good quality TTS systems with less data compared to E2E systems.

Index Terms: Speech synthesis, Histogram Equalization, HMM- based speech synthesis, Waveglow, Hybrid systems

1. Introduction

Through time there have been many parametric and non-parametric approaches to train text-to-speech synthesizers. One of the first popular approaches in building a TTS system is the Unit Selection Synthesis (USS) [1, 2]. In USS, an ample amount of clean and labeled speech is spliced into sub-word units (phonemes/syllables) and stored in the database. During synthesis, appropriate units are selected based on the context and concatenated together. A large amount of data is required for this approach, with any emotional context being hard to intercalate during training. Following USS, the parametric hidden Markov model (HMM) based speech synthesis system [3, 4] was popular. Mel-cepstral coefficients and excitation parameters are extracted from the audio files and then modeled by context-dependent HMMs. The advantage of HMM-based systems is robustness to less training data, flexibilty to incorporate speaking styles, emotions, and other characteristics by tuning the HMM parameters. Most recently, non-parametric neural network based end-to-end (E2E) systems have become popular as they achieve speech quality as good as human speech. Despite this, E2E training is computationally expensive. The high sensitivity to data may lead to overfitting, furthermore, a substantial amount of data is required for training E2E based system. India has 1652 languages[1] with 22 being official languages. Major language families being Dravidian, Aryan, Sino-Tibetan. This diversity results in lack of data, making Indian languages digitally low-resource. HTS is robust to low-resource scenarios and hence, more suitable for Indian languages.

A hybrid TTS system combines one or more TTS techniques to enhance the existing system. In [5], the interweaving of natural and statistically generated segments is performed by iteratively finding a dynamic path containing as many natural segments as possible. Another work in the literature makes use of sequence-to-sequence (S2S) model such as Tacotron2 [6] to enhance USS based synthesis [7]. [8] explores the possibility of training multilingual systems for Indian languages in an E2E framework. Here, a multi-language character map (MLCM) is proposed to reduce the vocabulary size by mapping similar characters across languages together. The focus of the current work is on improving the quality of the synthesized audio. Some works in this domain include [9], this is a work on voice conversion. It uses postfilters to modify the modulation spectrum (MS) to make MS of synthetic voices close to natural ones. It also presents a way to overcome over-smoothing effect which makes statistical based voices muffled up.

In this paper, we propose a Hybrid TTS system, which makes use of the HTS framework and Waveglow vocoder. In the HTS framework, context-dependent pentaphones are modeled by HMMs. A unified parser [10], and a common label set [11] is used to process the text. Next, context-dependent labels are used to generate the acoustic features using a set of concatenated HMMs. The generated features are fed as input to a Mel Log Spectrum Approximation (MLSA) filter for synthesis. The mapping between the mel-spectra of HTS synthesized waveforms from training, and source waveforms are learned using HEQ. During test, the mel-spectra of the output generated by the HMM based synthesizer is mapped to training data spectra using the learned HEQ mapping. The mapped spectra are then fed to the Waveglow synthesizer. While HTS based system can be trained with small amounts of data, the Waveglow vocoder enables good quality synthesis.

HEQ is a popular technique that is widely used in image processing [12]. It is used to effectively spread out the high-intensity elements that are present in the image, thus bringing out the global contrast in the image. HEQ is applied to colored components separately in the case of a colored image. In speech recognition, HEQ provides a transformation from noisy acoustic features to clean reference features [13]. This results in better recognition accuracy. The current work proposes to adapt HEQ for speech synthesis. To the best of our knowledge, this is the first attempt at improving the quality of synthesis output

*Equal contribution
[1] 1961 Census report of India

http://dx.doi.org/10.21437/Interspeech.2020-3180

using HEQ.

There has always been a trade-off between how flexible the model can be and the synthesis quality [14]. It is to be noted that the non-parametric TTS systems are highly data sensitive. Moreover, data being low-resourced adds to the challenge of training a TTS synthesizer. The current work proposes to develop a technique to improve synthesis quality and also retain the flexibility provided by classical approaches. Subjective evaluations suggest that the synthesized speech using the proposed approach produce better quality over that of the HTS system and a vanilla HTS+Waveglow system.

The rest of the paper are organized as follows. System description is given in Section 2. Histogram equalization and the proposed approach are presented in Section 3. Experiments and results are discussed in Section 4. The work is concluded in Section 5.

2. System Description

The system that is proposed in the paper consists of three components, namely, the hidden Markov model based synthesis framework (HTS), Waveglow vocoder and histogram equalization across features. The first two modules are described in this section, while the histogram equalization technique is explained in Section 3.

2.1. HTS framework

HTS is an HMM-based speech synthesis system that has been developed using HTK [15] toolkit. It uses the FESTIVAL [16] framework as a text analyzer. In the training part, the linguistic specification is extracted from the speech corpus. This part is known as the front end of the HMM synthesis. Given the linguistic specification, the waveform is being generated from that specification. It is to be noted that the linguistic specification is stored as parameters of the HMM model which can be retrieved later to synthesize the waveform.

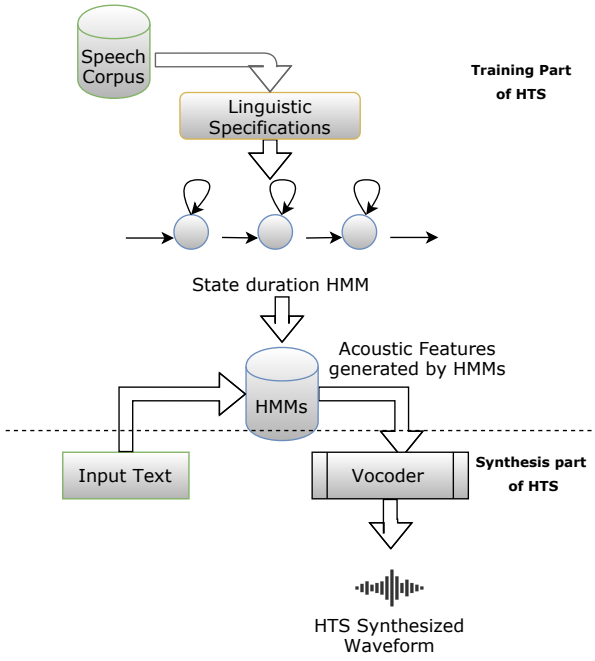

Figure 1: *Working pipeline of HTS*

Figure 1 depicts the training and synthesis parts of the HMM synthesizer. The training phase of HTS consists of three tasks:

1. Parsing words to constituent phone sequences
2. Segmenting Indic speech data
3. Training HMMs

The training text is parsed into phoneme units using the unified parser for Indian languages [10]. The unified parser converts words to syllable and phoneme sequences. After the generation of phoneme and syllable sequences, speech waveforms are aligned at the phone level using a hybrid HMM-deep neural network (DNN) based approach proposed in [17]. Phone boundaries are modeled by HMMs. Neural networks are trained with these phone bounadries as initial boundaries. An iterative training of DNN-HMM gives accurate final phone boundaries. Syllable boundaries in the speech waveforms are obtained through a group delay (GD) based approach which uses short-term energy of the waveform. By empirically setting the window scale factor (WSF), spurious syllable boundaries are generated. The syllable boundary closest to a phone boundary is considered as the correct syllable boundary. Phone boundaries are re-estimated within the syllable boundaries. The resolution of segmentation being set using the empirically set parameter, WSF. The alignment is then made using the GD corrected boundary. The HTS model is then trained on these alignments.

2.2. Waveglow vocoder

Waveglow [18] deals with the vocoder part. Waveglow samples from a distribution for the generation of audio samples, ultimately using a neural network with a single likelihood cost function as a generative model. Here, audio sample distribution is modeled on conditioning the mel-spectrogram [6]. The samples are generated from a zero mean Gaussian distribution that has the same dimension as that of the output. Mel-spectrograms are created by applying mel-filters to the spectrum. There are a total of 80 mel filters extracted using 80 frequency bins of librosa [19] mel filters. A group of audio samples is squeezed into a vector. The Glow part enables the follow-up 1 x 1 convolutions, thus mixing information across channels. The next step uses non-linear layers of convolutions to transform the input Gaussian distribution to the desired distribution. Waveglow uses 12 coupling layers, each with 8 layers of dilated convolution. The mel-spectrograms are included at each of the coupling layers.

3. Histogram Equalization (HEQ)

In statistics, the histogram is a graphical representation of the data. The term “equalization” refers to mapping one distribution to another distribution. So, histogram equalization maps one histogram to another.

In image processing, the accuracy of the transformation is proportional to the number of observations, which is pretty big for images (thousand/million pixels). At the same time, for speech, a signal is broken into several segments (in milliseconds), each segment is called a frame, represented by a feature vector. The number of frames is much less compared to that in image processing. In the context of automatic speech recognition (ASR), HEQ is applied in the front end for eliminating noise. Experiments in [20] and [21] show that it reduces the linear and non-linear distortions in the speech signal. There are many variants/categories of HEQ [22, 23].

In this paper, we describe how HEQ can be adapted to provide better speech synthesis, which is fairly simple and easy to

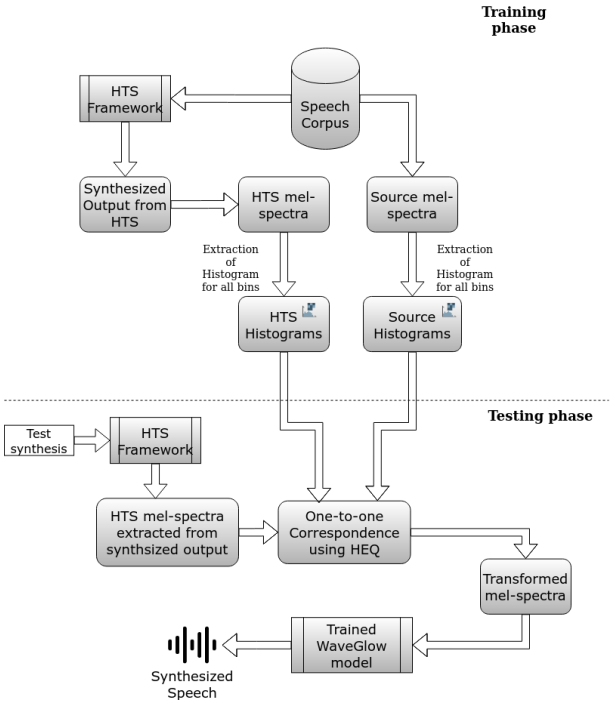

Figure 2: *Training and testing phases of the proposed approach using histogram equalization*

implement. Not only does it act as a feature compensator, it also comes out as a pretty good normalization technique. The proposed method is based on a hypothesis that the HTS generated utterances, and the spectrograms of the originally recorded utterances are highly related. The bottom panel of Figure 3 shows the original mel-cepstra histograms, and the top panel shows the corresponding mel-cepstra of HMM synthesized waveforms.

The mel-spectrograms of both source and the HTS synthesized audio files are extracted using mel-filters of librosa toolkit [19]. The new dimensions of individual files being *number of frames x number of mel filters*. We determine the histogram for each mel-filter after pooling all the data for that mel-filter across all training files. Using the same number of mel-filters for both classes (HTS and original waveforms), we get an equal number of histograms across both classes. Figure 2 give a formal and detailed view of the proposed approach. Each histogram corresponds to a particular mel-filter. The histograms are created in a fashion such that each bin of a particular histogram occupies almost an equal amount of area, and an equal number of feature vectors are present in each bin (Figure 3). This leads to a non-uniform bin width, the regions that are dense are represented by a narrow bin width. The number of bins across all histograms is the same. A one-to-one mapping transformation is applied between the histograms of HTS and source spectra, and this mapping is stored during training. This transformation is applied to HTS spectrogram features while testing. For test text, we use the HTS model and generate the HTS synthesized audio. We compute the mel-spectra of the speech signal. For each frame of the mel-filter, we find the bin in the HTS histogram and replace it with corresponding mean bin value in the source wave histogram.

$$\text{Let } H_{orig_i} \text{ and } H_{hts_i}\ i = 1 \Rightarrow n$$

$$\text{n = number of mel-filter}$$

be the stored histograms of original and HTS generated waveforms.

$$x_i = \text{transformation for } i^{th} \text{ mel-filter}$$

$$\text{y = test mel-spectra}$$

The transformed value to be replaced with for each entry in the test mel-spectra

$$x_i[y_{ij}] = \overline{H_{orig_i}}[H_{hts_i}[y_{ij}]]$$

$$j = 1 \Rightarrow\ k$$

$$\text{k = number of frames in test utterance}$$

The transformed mel-spectra are fed as input to Waveglow for synthesis.

4. Experiments and Results

4.1. Dataset used

For HTS and Waveglow training, 8 hours of Hindi male and female datasets are used. Datasets are part of the Indic TTS database [24]. Speech is recorded by professional speakers in a studio enviroment at 48 kHz sampling rate.

4.2. Experimental setup

HTS systems (male and female systems) are trained after obtaining the HMM-DNN based alignments. HTS synthesized utterances generated from training text are then downsampled to 22.05 kHz, making it compatible with the Waveglow environment. A pre-trained Waveglow model trained on LJ speech dataset [25] is used as the initial model. It is then re-trained on Hindi male data for Hindi male model and Hindi female data for Hindi female model.

Librosa mel-filters are used to extract the mel-spectra. Both source and HTS synthesized audio files are converted to mel-spectra. 80 mel-filters are used to extract mel-spectra, thus, each utterance corresponds to a size of number of frames times 80. A separate histogram is built for each frequency bin of source and HTS spectra. Thus, totally 160 histograms are built and stored. Samples of histogram plots are shown in Figure 3.

For learning the HEQ transformation, experiments were performed with different amounts of data to generate the histograms– 10 minutes, 30 minutes, 2 hours and 8 hours. It was observed that the system's performance was almost constant even as we increased the duration of data required to train the HEQ transformation beyond 10 minutes.

The source and HTS histograms used in the experiments are 80 bin histograms. Increasing the number of bins didn't affect the synthesis quality either. We did not attempt to directly convert the mel-generalised-cepstra (MGCs) to mel-spectra as there is incompatibility between the acoustic feature parameters of HTS and Waveglow.

4.3. Evaluation measures

Degradation mean opinion scores (DMOS) is used to assess the quality of synthesized speech. In this measure, the evaluators listen to the synthesized samples of the proposed system, HTS synthesized samples, and the source speech data in random order. The mean scores with respect to the source speech data is calculated as DMOS scores. Listeners are asked to rate each utterances on a scale of 1-5 (5 being human-like quality and 1 being poor quality) based on the naturalness of the synthesized utterances. 15 native Hindi speakers were chosen as listeners to evaluate the systems and a total of 25 sentences were used.

A pairwise comparison test (PC test) was also conducted

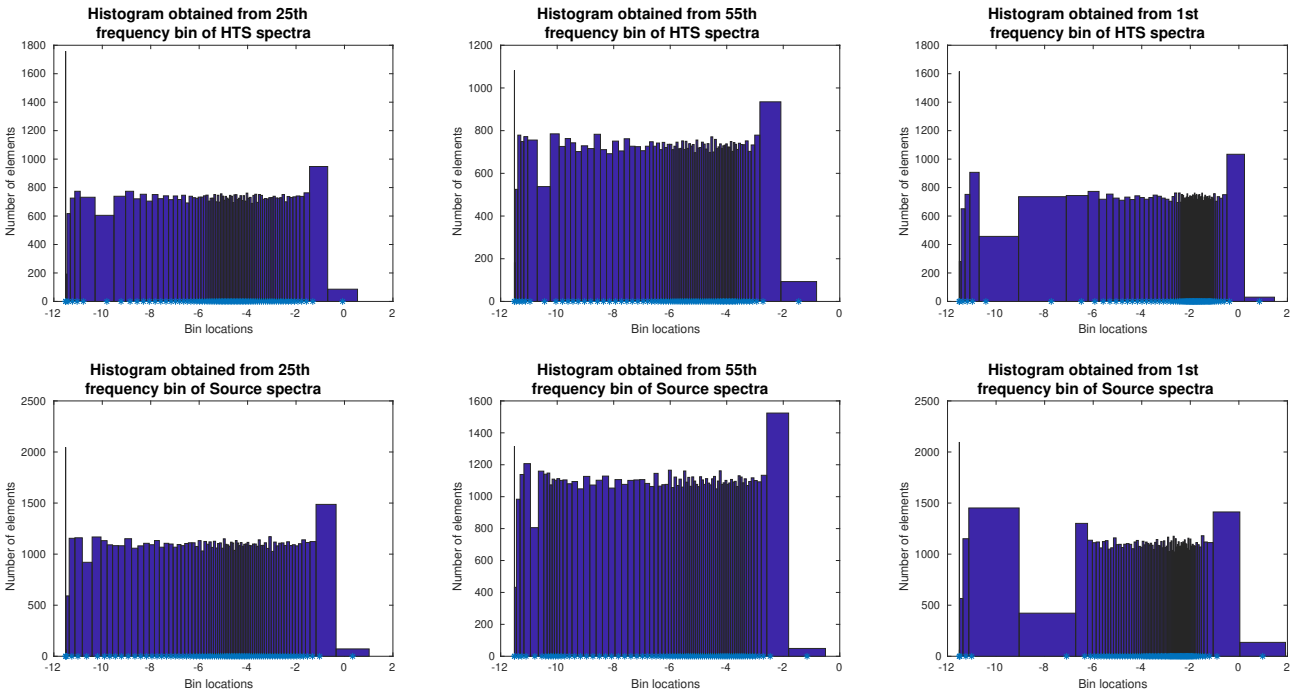

Figure 3: *Sample histograms of source and HTS mel-spectra*

Table 1: *Results of subjective evaluation comparing the proposed approach and the baseline system*

System Description	DMOS	
	Male	Female
HTS System	3.18	2.97
HTS + Waveglow using HEQ	3.40	3.39

Table 2: *PC test results: Hybrid system with HEQ vs. Hybrid system without HEQ (preference in %)*

Language	Hybrid system with HEQ	Hybrid system without HEQ	Equal
Hindi-male	56.29	21.46	22.25
Hindi-female	77.33	8	14.66

between the proposed hybrid system and the hybrid system without HEQ. In this measure, the evaluators listen to the same synthesized utterance generated by both systems and indicate their score as their preference. 15 native speakers were chosen as listeners to compare each of same 10 pairs of utterances synthesized by the above mentioned systems (HTS+Waveglow using HEQ and HTS+ Waveglow without HEQ). The utterances were presented in random order to the listeners.

4.4. Results

Table 1 presents the DMOS scores of the HTS system along with the proposed system. It is clearly seen that the proposed system exhibits a higher score than the HTS system in case of both male and female systems. From Table 2 it is evident that the proposed hybrid system using HEQ has higher preference. Based on listeners' feedback and informal assessment, it was observed that the hybrid system with HEQ removed the buzziness present in the HTS synthesised audio. The buzziness in the HTS synthesis is due to the averaging of parameters as HMM states are tied in tree-based clustering.

5. Conclusions

Although there have been many S2S based TTS systems that achieve better synthesis quality, HMM-based speech synthesis systems are robust, flexible, and adaptable for low-resource scenarios. State-tying is a problem with HTS, leading to a quality gap between natural and synthetic speech. There are cases of voice being shaky and over-smoothing. Our system presents an approach to overcome this by using HEQ, along with the advantage of using the Waveglow vocoder for high fidelity speech. This paper provides a way to enhance the quality of HMM-based synthesis and also paves the way to extend the work of HEQ in the domain of speech synthesis. HEQ can also be used for voice adaptation tasks in speech synthesis. This approach will be especially useful when adapting to new languages in the same language family with low digital resources. Similar to ASR, we would also like to reduce the data required for adaptation for new speakers.

6. Acknowledgements

We would like to thank the Department of Science and Technology (DST), the Ministry of Electronics and Information Technology (MeitY), Office of the Principal Scientific Adviser (PSA) to the Government of India, for funding the projects, "Text to Speech Generation with chosen accent and noise profile for Aerospace and Industrial domains" (CSE1819172MIMPHEMA), "Natural Language Translation Mission" (CS2021012MEIT003119), "Speech to Speech Machine Translation" (CS2021152OPSA003119), respectively.

7. References

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Inter-*

national Conference on Acoustics, Speech and Signal Processing (ICASSP), 1996, pp. 373–376.

[2] A. W. Black and P. Taylor, "Chatr: a generic speech synthesis system," in *Proceedings of the 15th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 1994, pp. 983–986.

[3] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 3, pp. 1039–1064, November 2009.

[4] S. King, "A beginners' guide to statistical parametric speech synthesis," *The Centre for Speech Technology Research, University of Edinburgh, UK*, 2010.

[5] S. Tiomkin, D. Malah, S. Shechtman, and Z. Kons, "A hybrid text-to-speech system that combines concatenative and statistical synthesis units," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 5, pp. 1278–1288, 2010.

[6] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[7] X. Zhou, Z. Ling, and L. Dai, "Extracting unit embeddings using sequence-to-sequence acoustic models for unit selection speech synthesis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7659–7663.

[8] A. Prakash, A. Leela Thomas, S. Umesh, and H. A Murthy, "Building Multilingual End-to-End Speech Synthesisers for Indian Languages," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 194–199. [Online]. Available: http://dx.doi.org/10.21437/SSW.2019-35

[9] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," vol. 24, no. 4, 2016, pp. 755–767.

[10] A. Baby, N. N. L., A. L. Thomas, and H. A. Murthy, "A unified parser for developing Indian language text to speech synthesizers," in *International Conference on Text, Speech and Dialogue*, 2016, pp. 514–521.

[11] B. Ramani, S. Lilly Christina, G. Anushiya Rachel, V. Sherlin Solomi, M. K. Nandwana, A. Prakash, S. Aswin Shanmugam, R. Krishnan, S. Kishore, K. Samudravijaya, P. Vijayalakshmi, T. Nagarajan, and H. A. Murthy, "A common attribute based unified HTS framework for speech synthesis in Indian languages," in *Speech Synthesis Workshop (SSW)*, 2013, pp. 291–296.

[12] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. T. H. Romeny, and J. B. Zimmerman, "Adaptive histogram equalization and its variations," *Comput. Vision Graph. Image Process.*, vol. 39, no. 3, p. 355–368, Sep. 1987. [Online]. Available: https://doi.org/10.1016/S0734-189X(87)80186-X

[13] L. García, C. B. Ortúzar, A. De la Torre, and J. C. Segura, "Class-based parametric approximation to histogram equalization for ASR," *IEEE Signal Processing Letters*, vol. 19, no. 7, pp. 415–418, 2012.

[14] M. Bulut and S. S. Narayanan, "Chapter 10 - speech synthesis systems in ambient intelligence environments," in *Human-Centric Interfaces for Ambient Intelligence*, H. Aghajan, R. L.-C. Delgado, and J. C. Augusto, Eds. Oxford: Academic Press, 2010, pp. 255 – 277. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780123747082000103

[15] S. Young and P. Woodland, "HTK: Speech recognition toolkit," http://htk.eng.cam.ac.uk/.

[16] A. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system," `http://festvox.org/festival/`, 1998.

[17] A. Baby, J. J. Prakash, R. Vignesh, and H. A. Murthy, "Deep Learning Techniques in Tandem with Signal Processing Cues for Phonetic Segmentation for Text to Speech Synthesis in Indian Languages," in *INTERSPEECH*, 2017, pp. 3817–3821.

[18] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3617–3621.

[19] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 2015.

[20] A. De La Torre, J. C. Segura, C. Benitez, A. M. Peinado, and A. J. Rubio, "Non-linear transformations of the feature space for robust speech recognition," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2002, pp. I–401.

[21] J. C. Segura, M. C. Benítez, A. De La Torre, S. Dupont, and A. J. Rubio, "VTS residual noise compensation," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2002, pp. I–409.

[22] S.-H. Lin, Y.-M. Yeh, and B. Chen, "A comparative study of histogram equalization (HEQ) for robust speech recognition," in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 2, June 2007*, 2007, pp. 217–238.

[23] P. M. Martinez, J. C. Segura, and L. Garcia, "Robust distributed speech recognition using histogram equalization and correlation information," in *Eighth Annual Conference of the International Speech Communication Association*, 2007, pp. 1058–1061.

[24] A. Baby, A. L. Thomas, N. N. L, and H. A. Murthy, "Resources for Indian languages," in *Community-based Building of Language Resources (International Conference on Text, Speech and Dialogue)*, 2016, pp. 37–43.

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