Text to Speech Synthesis: A Systematic Review, Deep Learning Based Architecture and Future Research Direction

Fahima Khanam[1], Farha Akhter Munmun[1], Nadia Afrin Ritu[1], Aloke Kumar Saha[2], and Muhammad Firoz Mridha[1]
[1] Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh
[2] Department of Computer Science and Engineering, University of Asia Pacific, Dhaka, Bangladesh
Email: {fahima, farha.akter, nadira.afrin, firoz}@bubt.edu.bd, aloke@uap-bd.edu

***Abstract*—Text to Speech (TTS) synthesis is a process of translating natural language text into speech. Pieces of recorded speech generate synthesized speech and a database is maintained for storing this synthesized speech. A speech synthesizer's output is determined through its resemblance to the person utter and its capacity to be implied. In recent years between the two main subsections: machine learning and deep learning of Artificial Intelligence (AI), deep learning has achieved huge success in the domain of text to speech synthesis. In this literature, a taxonomy is introduced which represents some of the deep learning-based architectures and models popularly used in speech synthesis. Different datasets that are used in TTS have also been discussed. Further, for evaluating the quality of the synthesized speech some of the widely used evaluation matrices are described. Finally, the paper concludes with the challenges and future directions of the text-to-speech synthesis system.**

***Index Terms*—Text to Speech (TTS), deep learning, acoustic features, parametric synthesis, concatenative synthesis, text analysis**

I. Introduction

Text is one of the most basic forms of computer interaction with humans. Most of the time, we expect this interaction to feel as natural and as smooth as the interaction we experience with other humans. For providing this naturalness, text-to-speech conversion can be more effective. Speech Synthesis, formally known as Text-to-Speech (TTS), allows any computing system to convert a written text to a voice message via a microphone or telephone [1]. A speech synthesizer is an information processing system.

The advancement in the quality of speech synthesizers increases gradually with the expanding applications of speech synthesis. For example, to support blind persons in reading and efficient communication, different aids are the most necessary and useful application field in speech synthesis. Synthesized speech can also be used for particular functions like spelling and pronunciation teaching for various languages. Nowadays, most smartphones are capable of listening to questions from end-users and answering back through an intelligent personal assistant—Cortana (Microsoft), Siri (iPhone), or Google Assistant (Android) [2]. Speech synthesis has been the mainstream in research on Artificial Intelligence (AI). The main goal of a TTS system is to automatically produce speech output from new, arbitrary sentences. Texts are converted to speech in two main steps. The first step is text analysis, in which a text-input string is transformed into a symbolic or phonetic representation used to build acoustic and prosodic models. The second step is to create the speech waveforms. The techniques used for speech synthesis can be partitioned into two broad categories: (1) Traditional machine learning-based techniques and (2) Deep machine learning-based techniques. In traditional machine learning, two specific methods are used for TTS: concatenative speech synthesis [3] and parametric speech synthesis [4], [5]. Speech synthesis has been the mainstream in research on Artificial Concatenative synthesis is performed based on the concatenation of segments of the recorded voice. It is distinguished by selecting, storing, and smoothly concatenating human voices (phonemes, syllables, or longer units) [3]. There are different schemes for concatenative synthesis [6] like Epoch Synchronous Non-Overlap and Add (ESNOLA) [7], Pitch Synchronous Overlap and Add (PSOLA) [8], Time Domain Pitch Synchronous Overlap and Add (TDPSOLA) [9], [10], EMBROLA [11]. The parametric synthesis approach can also be regarded as a kind of concatenative synthesis [12]. The main dissimilarities lie in the units that are saved in the database and the signals restoration procedure. The most popularly used terms in this field have been demonstrated in Fig. 1. We have classified TTS based on architecture and models. The models used for TTS can be further divided into two categories: autoregressive (AR) and non-autoregressive (NAR). The autoregressive-based models along with the architectures have been discussed in Section III and the non-autoregressive models are also discussed in the last part of Section III in the form of a table. Fig. 2 represents the

Manuscript received February 12, 2022; revised June 6, 2022.

doi: 10.12720/jait.13.5.398-412

overall taxonomy of the deep learning-based architectures and the models used for TTS.

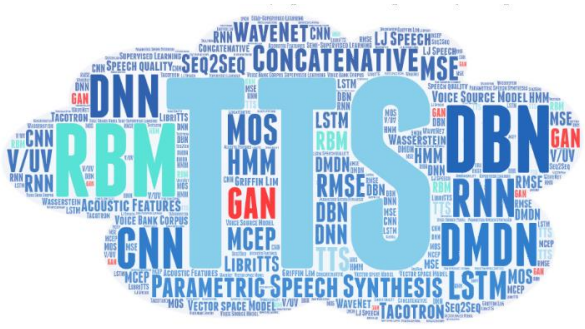

Figure 1. Word cloud with the most used terms in the field of speech synthesis.

In the last decade, a few review papers have been published on speech synthesis with Deep Machine Learning. But all of these articles have focused on deep learning-based techniques for some specific languages. We have studied several survey papers. Although these papers have presented a good literature survey, they have either discussed only deep learning-based speech synthesis methods and technologies [6], [12]-[14] or surveys based on a specific model or language. For example, R. A. Khan *et al*. [3] presented a review on concatenative-based speech synthesis. While Kayte *et al*. [6] focused on Hidden Markov Model (HMM) based speech synthesis [15]. Kayte *et al*. [14] also discussed Marathi's Speech Synthesis. A large number of works with deep learning-based speech synthesis have been published. The article mainly focuses on deep learning-based speech synthesis architectures including their advantages and disadvantages, models. The main contributions of the survey paper include:

- The paper presents an overall taxonomy and analyzes the architectures, and models including different types of learning techniques for speech synthesis.
- The paper finds out the evaluation matrices for measuring the performance of the architectures.
- The paper summarizes the datasets used for TTS.
- Finally, the paper highlights the limitation of the existing architecture along with the future research directions for TTS.

The rest of the paper is organized as: Section II provides a complete description of the survey methodology, Section III presents supervised learning-based speech synthesis, Section V represents a full description of the datasets, and Section VI depicts the most popularly used evaluation metrics. Finally, Section VII reveals the discussion and future research direction. the architectures along with the models that have been used for speech synthesis, Section IV provides unsupervised and semi-supervised learning based speech synthesis.

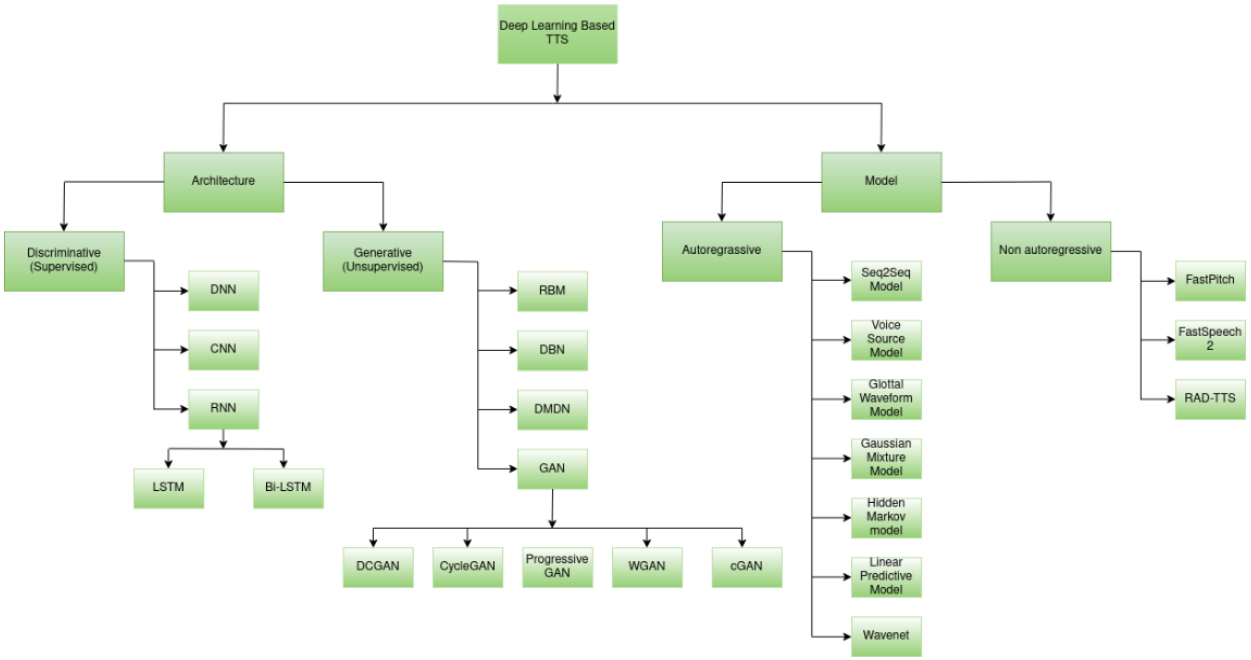

Figure 2. Overall taxonomy of TTS system.

II. Survey Methodology

A systematic Literature Review (SLR) selects experimental data by using a defined plan or protocol [16]-[18] that offers readers a comprehensive knowledge of the literature in different research areas [19]-[21]. An SLR also provides a complete idea about the gaps in specific research topic and directs the future research direction for that topic. The recent study in SLR was performed in three main stages: 1) Planning for the review, 2) Performing the review, and 3) Reporting the review [22].

A. Planning the Survey

First, the significance of doing this survey was pointed out. Then, suitable inclusion and exclusion criteria were selected for searching the most relevant papers, articles, and studies. Our survey mainly focused on the databases: ACM, IEEE, Springer, and Elsevier. Finally, research questions related to this survey were described.

B. Search Strategy and Syntax

Some major databases: Google Scholar, Springer Link, ACM digital library, IEEE Xplore, and ScienceDirect were used for searching the related articles. For searching the papers several keywords were used related to "Deep learning" AND "Speech synthesis" OR "Text to speech synthesis". Fig. 3 illustrates the graphical representation of the percentage of our reviewed papers.

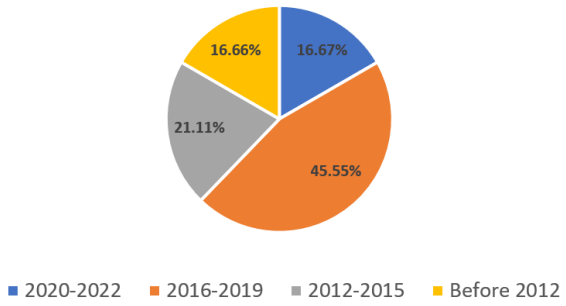

Figure 3. Year wise view of literature.

C. Inclusion and Exclusion Criteria

The criteria for selecting an article were as follows: (a) The studies published anytime on or before 2022; (b) The journal articles including conference papers; (c) Full-text availability in digital databases; (d) The articles that proposed model or framework; and (e) The research written in English. In addition, the following exclusion criteria were used: (a) duplicate articles found in multiple academic databases; (b) papers based on quality evaluation criteria; and (c) review, book chapters, magazine articles, theses, interview-based articles, and monographs.

III. Deep Learning-Based Speech Synthesis

In this section, we have discussed the previous work on speech synthesis. Throughout this literature, we found out a total of eight basic architectures for TTS system: Restricted Boltzmann Machine (RBM), Deep Belief Network (DBN), Deep Mixture Density Network (DMDN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), Deep Neural Network (DNN), and Generative Adversarial Network (GAN) are popularly used for TTS.

From this review, we have found some meaningful insight: the majority of the articles have implemented their work based on AR and NAR models. Where 91% of the papers used AR model with different types of architectures such as RBM, DBN, DMBN, DNN, CNN, RNN, LSTM, and GAN and 9% of the papers are based on NAR model. All of these architectures with AR models solved the three prime factors: vocoders, acoustic modeling accuracy, and over-smoothing that hinder the quality of synthetic speech [23]. The acoustic modeling accuracy and the over-smoothing problem has been solved partially by LSTM [23]. Moreover, CNN and RNN both are used for TTS systems but considering the training time, CNN takes less time than RNN [24]-[26]. However, recently NAR shows better performance than AR with different architectures [27]-[33].

A. Restricted Boltzmann Machine (RBM)

An RBM is an undirected bipartite model which is used for modeling speech recognition [34]-[36] and spectrogram coding [16]. This RBM is used as a strategy for pre-training a deep autoencoder or a DNN and is represented in Fig. 4.

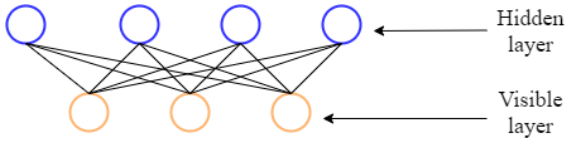

Figure 4. Architecture of a restricted Boltzmann machine.

Z. H. Ling *et al*. [37] proposed an RBM-based spectral envelope modeling method (SPE-RBM) for statistical parametric speech synthesis. For representing the distribution of the spectral envelopes, they adopt RBM at each HMM state rather than using single Gaussian distributions. Their aim was to describe the high dimensional spectral envelopes more strongly and reduce the problem caused by over-smoothing. Their experimental results show that SPE-RBM can significantly upgrade the conventional HMM-based speech synthesis [38] system's naturalness using Mel-cepstral (MCEPGaussian).

B. Deep Belief Networks (DBNs)

DBNs are a class of DNNs that use probabilities and unsupervised learning to produce outputs. The network is like a stack of RBMs, unlike RBMs, nodes in a deep belief do not communicate laterally within their layer. Fig. 5 represents the architecture of a DBN.

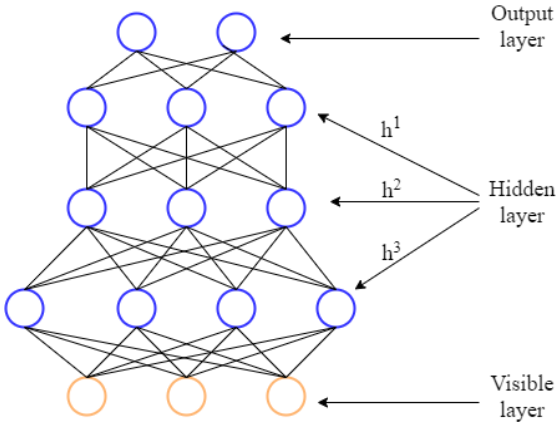

Figure 5. Architecture of deep belief network

Z. H. Ling *et al*. [35] extended their previous work [37] by extending RBM to DBN. They proposed a DBNHMM system for improving the naturalness of synthesized speech and reducing the over-smoothing effect instead of using traditional HMM-based speech synthesis systems.

Another model was proposed by S. Kang *et al*. [39] to fully make use of the generative nature of DBN. They tried to represent the speech parameters together with the spectrum and F0 concurrently and produce these parameters from DBN for synthesized speech. Their experimental results confirm that the spectrum created from DBN has significantly less distortion, and the general quality is far better than that of context independent HMM.

R. Fernandez *et al*. [40] proposed a combined model that mixes up the Gaussian process and DBN to predict prosodic contour, F0 aims from textual features by using nonparametric, exemplar-based regression in a speech synthesis system. In their work, they examined non-linear features extracted via DBNs [41]. They compared their proposal with the ideal clustering-tree methods implemented in parametric synthesis for predicting the prosodic target.

Y. J. Hu *et al*. [42] developed an HMM-based parametric speech synthesis method using DBN. A DBN was used to estimate the spectral envelopes at the training phase and then these spectral envelopes were transformed into binary codes. As per experimental results, their proposed method provides better naturalness than conventional methods using melcepstra. Table I represents the DBN-based models, advantages, and disadvantages.

C. Deep Mixture Density Networks (DMDNs)

Combine DNN and a mixture of distributions. Deep Mixture Density Networks (DMDNs) can be used for speech generation, artificial hand writing generation and are applicable to a broad variety of business-relevant tasks. Fig. 6 represents the architecture of a DMDN.

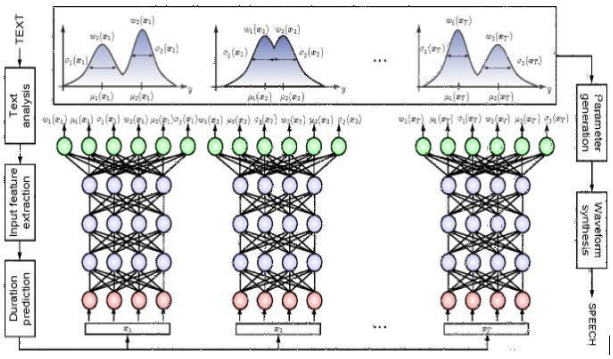

Figure 6. Overview of a deep MDN (DMDN) [43]. The red circles represent the input, the blue circles represent hidden units, and the green circles represent the output units.

H. Zen *et al*. [43] has extended DNN based Statistical Parametric Speech Synthesis (SPSS) by inaugurating MDNs. One of the fundamental problems in SPSS using DNNs is fewer variances and unimodal nature. To solve these problems, the authors used a mixture density output layer. The authors proved that by using a mixture density output layer, they predicted the acoustic features with better accuracy, and the naturalness has also been improved in the synthesized speech.

D. Deep Neural Network (DNN)

A Deep Neural Network (DNN)is an Artificial Neural Network (ANN) with several layers between the input and output levels. Neurons, synapses, weights, biases, and functions are all included in neural networks, which come in a range of shapes and sizes. Fig. 7 represents a DNN based architecture for the TTS system.

A novel neural network-based approach has been explored by S. H. Chen *et al*. to synthesize prosodic and spectral knowledge for Mandarin text-to-speech [44]. Firstly, a text is analyzed and engaged in pulling out some relevant contextual characteristics. Secondly, utilizing these contextual characteristics, many MLPs are involved in synthesizing prosodic and spectral parameters. All of these prosodic parameters synthesizing MLPs are trained by the Back-propagation (BP) algorithm. They found that this method worked reasonably well, contrasting these synthesized parameters with the real ones.

T. Falas *et al*. [45] employed a neural network-based multilayer feed-forward architecture for the transformation of the Greek text to speech. It is hierarchically organized into three-unit layers: an input layer, an output layer, and an intermediate or “hidden” layer. From input to output, information flows across the network. To choose the most suitable neural network to be used, they tried different feed-forward neural network architectures. They discovered that a neural network with 60 to 80 hidden neurons seems to be the best configuration for both training and testing, with the maximum classification efficiency.

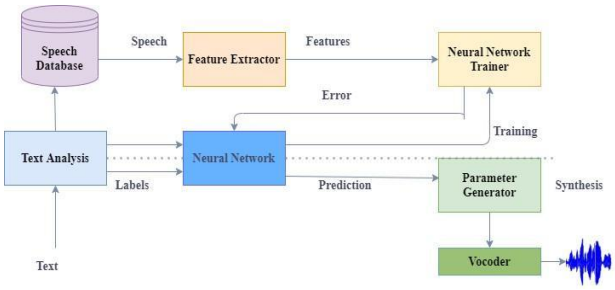

Figure 7. Illustration of Deep Neural Network (DNN) in speech synthesis system.

I. Rebai *et al*. proposed a TTS synthesis system for the Arabic language based on a statistical parametric approach [46]. They presented two subsystems. The first is the diacritization method that predicts the vowelization of the input text, and the second is the system of speech synthesis to generate a highquality speech based on statistical parameter synthesis. Both of these systems used a multilayer perceptron neural network with a fully connected backpropagation algorithm. The reason behind using this method is they can do a nonlinear mapping.

T. Raitio *et al*. [47] studied a DNN based voice source modeling method in speech synthesis with varying vocal effort. The proposed voice source model is compared with a robust and high-quality excitation modeling approach based on manually selected mean glottal flow pulses for each stage of vocal effort and a spectral matching filter to fit the desired styles’s voice source spectrum correctly. Subjective analyses show that the proposed DNN-based approach is comparable to the baseline approach but avoids manual pulse selection and is faster computational than a device using a spectral matching filter. Conventional SPSS depends on decision trees to cluster related contexts together, resulting in hidden HMM tied to context-dependent parameters. Decision tree clustering, however, has a significant disadvantage: it uses rigid division and subdivides the model space established on one characteristic at a time, fragmenting the data and failing to exploit associations between linguistic background features. So, H. Lu *et al*. took an attempt to

use DNNs to substitute decision tree parameter clustering, motivated by the ability to take full advantage of the continuously tested functionality that their new VSM-based front end provides [48]. The HMM benchmarking systems still exceed the DNNs from the results obtained. An investigation had been done on how to use the neural network in SPSS [49]. A process generation of SPSS based on generative models may be split into several components and DNNs representing those components. In this paper, the consequence of DNNs is investigated for each component by comparing DNNs with generative models. Experimental findings found that using a DNN as an acoustic model is successful and the generation of parameters together with a DNN increases the naturalness of the synthesized voice.

H. T. Luong *et al*. [50] considered the inclusion of standard text-based inputs of DNN-based acoustic models with auxiliary input features-collectively indicated as input codes-like speaker codes as well as other identified features such as gender and age encoding.

A. H. Ali *et al*. [51] proposed a deep neural network for Arabic speech synthesis. In this study, the authors used two models such as Tacotron 2 [52]-[54] and Tacotron because of their significant advantages and high efficiency. For speech synthesis, Tacotron uses the Grifflim method, while Tacotron 2 uses the WaveNet model. According to the collected experimental data, Tacotron 2 achieves a MOS of 4.38, while Tacotron 1 produces a MOS of 4.01. The obtained results showed that Tacotron 2 outperformed the concatenative system.

S. Takamichi [55]. used Fast Fourier Transform (FFT) spectra, to find the influence of Modulation Spectrum (MS) based preprocessing for DNN based speech synthesis. Preprocessing speech from training data is an efficient way to improve acoustic model training accuracy. The authors suggested an MS-based preprocessing method for DNN-based speech synthesis utilizing vocoder parameters titled "speech parameter trajectory smoothing" and verified that the method increases training accuracy by reducing components that are difficult to describe with the acoustic model.

L. Chen *et al*. [56] presented a method for synthesizing the Dungan language and compared their proposed method with conventional HMM-based Dungan speech synthesis. The synthesis was done by training a collection of DNN based acoustic models. The language was then synthesized by mapping the language features of the Dungan language with the acoustic features. As per experimental results, their proposed method provides a high naturalness and better synthesis effect for the Dungan language.

S Suzie *et al*. [57] proposed expressive DNN-based TTS with Limited Training Data. They used three methods for expressive speech synthesis: style codes, architecture with shared hidden layers and model retraining. Three architectures for producing expressive voice using deep neural networks are shown, each of which can reach a reasonable quality of synthesized speech with only 5 minutes of training data. Table II represents DNN based models, advantages and disadvantages.

E. Convolutional Neural Network (CNN)

CNN is a category of DNN. It is a neural network of many layers designed to examine visual inputs and carry out different tasks like classification, segmentation, and object detection for autonomous vehicles. It is also commonly used for computer vision/image recognition. CNN's architecture consists of three different layers such as convolution layers, pooling layers, and fully connected layers. An Overview of Convolutional Neural Network is shown in Fig. 8.

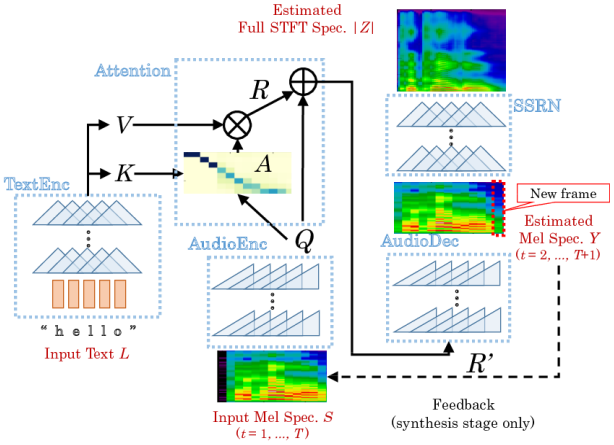

Figure 8. Overview of convolutional neural network. The DCTTS model consists of two networks: (1) Text2Mel, which synthesizes a Mel spectrogram from an input text, and (2) Spectrogram Super-resolution Network (SSRN), which converts a coarse Mel spectrogram to the full STFT spectrogram.

H. Tachibana *et al*. [25] presented a novel Text-to-Speech (TTS) technique based on deep CNNs without any recurrent units. In this paper, they proposed a Deep CNN-based TTS system rather than RNN-based systems [58], [59] because RNN usually needs a lot of time for training. The CNN-based TTS system aims to relieve the economic costs of training. They demonstrated in this paper that Deep Convolutional TTS systems can only be trained in one night (15 hours) while the synthesized speech's sound quality was almost appropriate.

H. Choi *et al*. [60] investigated multi-speaker emotional speech synthesis systems for Convolutional Neural Networks (CNN), according to the speaker modeling method and emotion modeling method. The Convolutional Neural Network (CNN) based speech synthesis system learns the mapping between linguistic and acoustic space by taking linguistic features as input and acoustic features (Mel spectrograms) as output. Compared to previous approaches in terms of naturalness, speaker similarity, and emotion similarity, the obtained experimental results showed that the multi-speaker emotional speech synthesis approach using trainable speaker embedding and emotion representation from Mel spectrogram exhibits greater performance. Table III represents CNN based models, advantages and disadvantages.

F. Recurrent Neural Network (RNN)

An RNN consists of three layers such as input layer, hidden layer, and output layer. A general diagram of the RNN is shown in Fig. 9. At each step, output builds upon

not only the current computations but also the previous computations. All of the inputs and outputs are independent of one another in standard neural networks, however in some circumstances, such as when predicting the next word of a phrase, the prior words are necessary, and so the previous words must be remembered. The hidden layer which remembers certain information about a sequence is the most essential element of RNN.

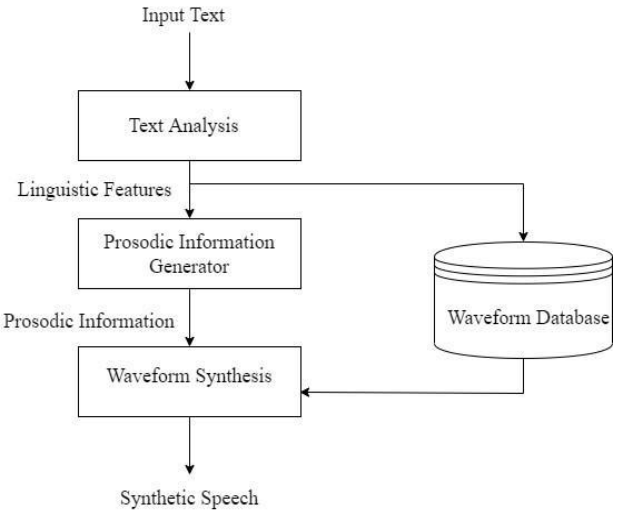

Figure 9. RNN based TTS architecture.

Authors in [61] proposed two different speech enhancement approaches based on RNN. In one approach, they used the features properly engaged in training TTS acoustic models, i.e., Mel cepstral (MCEP) coefficients. In the other technique, they trained an RNN using only the MCEP coefficients, adopting traditional speech enhancement methods. The purpose of the authors is to improve TTS voice quality. In this study, the proposed RNN is to create enhanced vocoder parameters to train an acoustic model [62]. They found that the second approach results in higher Mel cepstral distortion, and the synthetic voices trained with data-enhanced using RNN were rated higher and similar to voices trained with clean speech. Adversarially trained variational recurrent neural network (AdVRNN) based methods for designing and creating speech parameter sequences have been proposed in [63]. One of the key issues with speech synthesis is the issue of over smoothing. The authors applied an adversarial approach in this paper to solve the problem of over-smoothing. It has an increased dynamic range for synthesized speech data.

In paper [64], to produce the Mel spectrogram from the document, the authors used an RNN-based Seq2Seq model. The total loss on the model was measured as the sum of three component losses such as Mean-Squared-Error (MSE) [65], Linear Spectrogram MSE, and Binary Cross-Entropy Loss. The authors aim to improve model training speed and reduce model parameters. A recurrent network-based F0 model for TTS has been proposed in [66]. The proposed F0 model was trained to produce smooth F0 contours with relatively better perceived quality using a dropout strategy. Table IV represents RNN based models, advantages, and disadvantages.

G. Long-Short Term Memory (LSTM)

LSTM network is a form of an RNN capable of relying on sequence prediction issues to learn order. The backpropagation process can handle the remarkable sequence of time steps and the constant error flow. It is primarily about classifying, processing, and creating forecasts based on data from time series. A general diagram of the LSTM is shown in Fig. 10. Each memory cell has four units: input gate, forget gate, output gate, and self-recurrent unit.

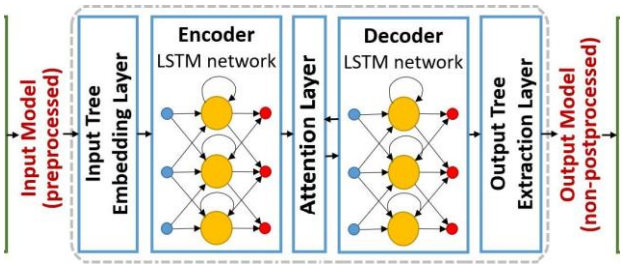

Figure 10. Overview of LSTM based neural network architecture [67].

Y. Fan *et al.* [67] proposed RNN for TTS Synthesis based on Bidirectional LSTM. This research uses objective and subjective methods to determine the test data's efficiency of three TTS systems, such as the Hybrid BLSTM-RNN, DNN, and HMM. They found that because of its ability to capture in-depth information in a sentence, the speech synthesized by the Hybrid system is significantly favored to the best HMM and DNN systems.

A flexible emphatic prosody generation model based on deep Bidirectional LSTM for controllable word-level emphasis realization has been presented by S. Shechtman *et al*. [68]. They trained a BiRNN-LSTM model for emphatic sentence prosody prediction. Their findings demonstrated that synthesized speech based on this model was evaluated as empathic while maintaining the original's quality and naturalness. C. Batista *et al.* [69] proposed an LSTM-based system for predicting the input parameters of the Klatt formant based speech synthesizer for utterance copy. Their system was compared the WinSnoori baseline software that is generated by the DECtalk TTS system and natural target ones. As per experimental results, our method outperforms the baseline for synthetic voices based on the parameters of PESQ, SNR, RMSE, and LSD. Table V represents LSTM based models, advantages and disadvantages.

H. Generative Adversarial Networks (GANs)

GAN is an algorithmic structure using two neural networks, dividing one against another (thus the "adversarial") to produce the new, simulated data that can move on to real data. They are commonly applied in generating pictures, video, and voices.

GANs have recently started appearing in TTS applications [70]. Fig. 11 represents a GAN based architecture for the TTS system.

Statistical Parametric Speech Synthesis (SPSS) with GANs under a multitask learning framework has been proposed by S. Yang *et al.* [71]. The authors aim to improve the output of synthesized speech in SPSS based

on GANs. To address the perceptual deficiency problem in the acoustic model, they suggested GANs.

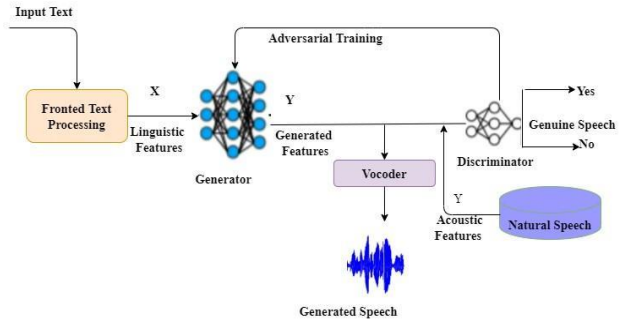

Figure 11. Overview of GAN in speech synthesis system.

Y. Saito *et al*. [72] proposed a method for SPSS incorporating GANs. They introduced a GAN composed of two neural networks: a discriminator for separating natural and produced samples and a generator for deceiving the discrimination. Acoustic models of speech synthesis are used to distinguish between natural and synthetic speech. Since the goal of the GANs is to minimize the divergence between the natural and produced speech parameters, in addition to reducing the loss of generation, the acoustic models are trained to close the parameter distribution of the generated voice parameters to that of natural speech. They also detected that the proposed algorithm incorporating the Wasserstein GAN mostly enhanced the synthetic speech quality compared to different GANs.

T. Kaneko *et al*. suggested a GAN-based post filter to minimize the distinction between natural speech and synthesized speech [73]. The consistency degradation involves three key factors: the precision of acoustic models, vocoding, and over smoothing. In this article, they concentrate on the issue of over-smoothing. The author's goal is to recreate from the synthesized one a "spectral texture" like the natural one. The results objective evaluation showed that the detailed spectral structure, including modulation spectra, can be reproduced by the proposed model and subjective evaluation showed that the quality of speech created is much closer to natural speech.

Recent advances in deep learning have led to the achievement of near-human naturalness through TTS systems. One of the big progress in TTS is the inauguration of waveform generation methods, such as WaveNet [74], [75] that have been embraced to use as "neural vocoder" [76].

A novel multi-scale GAN structure to generate pitch-synchronous waveforms is proposed by L. Juvela *et al*. [77]. The proposed generator works multi-time-scale progressive upsampling of characteristics maps and outputs waveforms, while the discriminator assures the waveforms last accurate at each time scale. They worked for producing glottal excitation (GlotGan) and speech waveforms (WaveGAN) pitch synchronously. For an SPSS system, the proposed model is tested as a neural vocoder, and the listening test results proved that GlotGAN could acquire comparable performance to a WaveNet vocoder.

B. Bollepalli *et al*. have done the first investigation to design the glottal waveform as an excitation waveform in SPSS using GANs [23]. They compared the synthetic speech generated by using both DNNs and GANs glottal waveforms. The results proved that, without using an additive noise part, the newly proposed GANs acquire synthesis efficiency compared to commonly used DNNs.

Y. Saito *et al*. [78] proposed two training algorithms using Short-Term Fourier Transform (STFT) spectra to integrate GANs into vocoder-free speech synthesis. To minimize the mean square error between natural and induced STFT spectral amplitudes at the original resolution and the variations in the distribution of their low-resolution distributions, acoustic models are trained in the proposed algorithm using a low-resolution GAN. This algorithm can be extended to one using multi-resolution GANs, which also minimizes the variations in the distribution of the natural and generated STFT spectra at the original resolution. Experimental results showed that the algorithm using the original GAN resolution and the proposed multi-resolution GAN algorithm decreased synthetic expression efficiency. Still, the suggested low-resolution GAN algorithm successfully improved it. To minimize the inconsistencies between natural and induced acoustic characteristics, a new framework integrating either a conditional GAN or its variant, Wasserstein GAN with Gradient Penalty (WGANGP), into multi-speaker speech synthesis using the WaveNet vocoder is proposed by Y. Zhao *et al*. [64]. As an acoustic model, the GAN generator works, and its outputs are used as WaveNet's local condition parameters. They also expand the GAN frameworks apply the Discrete-Mixture-of Logistics (DML) loss of a well-trained WaveNet as part of objective functions besides, mean squared errors and adversarial losses. Experimental findings revealed that acoustic models trained with WGAN-GP system using back-propagated DML loss achieve maximum consistency and speaker similitude subjective evaluation scores. Table VI represents GAN-based models, advantages, and disadvantages.

IV. Unsupervised and Semi-supervised Learning Based Speech Synthesis

Nowadays, TTS and ASR are two widespread tasks in speech processing. Y. Ren *et al*. [79] by using a few paired speech and text data and extra unpaired data, suggested the nearly unsupervised approach for both TTS and ASR. The system proposed consists of three elements for TTS and ASR, such as Denoising Auto-Encoder (DAE), Dual Transformation (DT), and Bidirectional Sequence Modeling (BSM). They found that by adding more paired data for PER on ASR and MOS on TTS gradually, the proposed approach would achieve higher accuracy. O. Watts *et al*. [80], the authors represent Text to Speech (TTS) systems that rely on universal resources (such as the Unicode character database) and unsupervised learning from unannotated data to ease system development. They develop their strategies in a way that avoids the need for language-specific expert knowledge. The paper explains how the techniques are applied to the 14 languages of the

Tundra corpus of 'found' audiobook data. Initial segmentation of the audiobooks into utterances chunks was performed using the lightly supervised GMM based VAD. Given concatenating waveforms, producing speech from a model has several potential advantages. An adaptation of the model is the most exciting. It has been shown that supervised voice adaptation can yield synthetic voices of high quality with an order of magnitude less data than needed to train a speaker-dependent model or to construct a simple unit-selection system. These supervised methods allow the target speaker to be marked with adaptation data. S. King *et al*. [81], the authors present a process that can be adapted without supervision, using only speech from the target speaker without any labeling. The authors aimed to compare unsupervised adaptation to supervised adaptation. It is already established that supervised adaptation can produce equivalent or better output than HMMbased synthesis, depending on the speaker. They wanted to find out how much degradation the unsupervised adaptation would produce. Y. A. Chung *et al*. [82] suggested semi-supervised learning to improve data efficiency in end-to-end speech synthesis. The authors' objective is to enhance Tacotron's data efficiency. Experimental results indicate that semi supervised Tacotron [83] achieves less MCD than the baseline, showing benefits beyond improved data quality. Table VII presents the unsupervised and semi supervised learning techniques and their advantages.

V. Dataset Description

The popular datasets used for speech synthesis are: i) the Japanese dataset, ii) the LJ Speech dataset, iii) the Chinese dataset, and iv) the Professional British English voice talents dataset. Among these datasets, the Japanese dataset is mostly used and presented in Table IX. The LJ speech and the Chinese dataset are equally used and the number comparatively less than the Japanese dataset. The detailed description of these two datasets are shown in Table X and Table XI respectively. The Professional British English dataset is least used and Table XII represents this dataset. Besides these four datasets, some datasets like ARCTIC corpus, Voice Bank corpus, LibriSpeech corpus, etc. [34], [39], [40], [51], [44]-[47], [61], [63], [64], [66], [67], [81], [82], [84]-[86] have also been used.

TABLE I. DBN Based Models, Advantages and Disadvantages

Architecture	Model	Reference	Advantages	Disadvantages
DBN	HMM	Z. -H. Ling *et al*. [35], Y. J. Hu *et al*. [42]	1. Quick and efficient because greedy learning algorithms are used to train DBN. 2. Help to optimize the weights at each layer.	1. Training efficiency is very low. 2.DBNs have a problem of catastrophic forgetting. 3. DBN is blind to logic and reasoning due to lack of knowledge representation within itself.
	Gaussian Process Regression model	R. Fernandez *et al*. [40]		
	Combination of mixed Gaussian Bernoulli RBMs, mixed Categorical- Bernoulli RBMs,and Bernoulli RBMs	S. Kang *et al*. [39]		

TABLE II. DNN Based Models, Advantages and Disadvantages

Architecture	Model	Reference	Advantages	Disadvantages
DNN	Linear predictive model	S.H. Chen *et al*. [44]	1. Simple, fast, and easy to program. 2.No need to have prior knowledge about the network. 3. Only the number of inputs is tuned and not another parameter. 4. Models high-dimensional acoustic parameters, Replace decision tree in HMM.	1. Possibly be sensitive to noisy data and irregularity. 2. The performance is highly dependent on the input data. 3. Needs excessive time for training. 4. Decision trees model complex context dependencies.
	N/A	T. Falas *et al*. [45]		
	MFCC neural network	I. Rebai *et al*. [46]		
	Voice source model	T. Raitio *et al*. [47]		
	Vector space model	H. Lu *et al*. [48]		
	HMM	K. Hashimoto [49], L. Chen *et al*. [56]		
	Acoustic model	H. T. Luong *et al*. [50], S. Takamichi [55]		
	Tacotron model	A. H. Ali *et al*. [51]		
	style-dependent shared hidden layer model (SDSM)	S. Suzie *et al*. [57]		

TABLE III. CNN BASED MODELS, ADVANTAGES AND DISADVANTAGES

Architecture	Model	Reference	Advantages	Disadvantages
CNN	Acoustic model	H. Choi *et al*. [60]	1. Easily extract relevant information at a low cost. 2. Robust to noise.	1. Cannot predict future behavior. 2. Due to its overfitting problem, the computational cost is high.
	DCTTS	H. Tachibana *et al*. [25]		

TABLE IV. RNN BASED MODELS, ADVANTAGES AND DISADVANTAGES

Architecture	Model	Reference	Advantages	Disadvantages
RNN	Acoustic model	C. Valentini-Botinhao *et al*. [61], J. Y. Lee *et al*. [63]	1. Can predict future behavior. 2. Can efficiently work where the length of the inputs does not matter. 3. Help to optimize model size.	1. Computation is slow, caused by its recurrent behavior. 2. Hard to train caused by its vanishing or exploding gradient problems. 3. Cannot process the long sequence of time steps.
	Seq2Seq model	X. Wang *et al*. [66]		
	F0 model	X. Wang *et al*. [66]		

TABLE V. LSTM BASED MODELS, ADVANTAGES, AND DISADVANTAGES

Architecture	Model	Reference	Advantages	Disadvantages
LSTM	HMM	Y. Fan *et al*. [67]	1.Vanishing gradient problems can be overcome and easier to train. 2.Can process the long sequence of time steps. 3.Prevent back propagated errors from vanishing or exploding problems.	1. Limited memory bandwidth.
	Deep RNN (DRNN) model	S. Shechtman *et al*. [68].		

TABLE VI. GAN BASED MODELS, ADVANTAGES, AND DISADVANTAGES

Architecture	Model	Reference	Advantages	Disadvantages
GAN	Acoustic model	S. Yang *et al*. [71], Y. Saito *et al*. [72], Y. Saito *et al*. [78], Y. Zhao *et al*. [64]	1. GANs don't need labeled data; they can train using unlabeled data while they know the data's internal representations. 2. Can generate data that is similar to real data. 3. GANs can learn messy and complicated distributions of data. 4. The discriminator network is a classifier and can be used to classify objects.	1. Hard to train, unstable. 2. Mode Collapse issue. The learning process of GANs that lack a pattern, the generator will start to degenerate and the same sample points will still be generated and the learning cannot be continued.
	N/A	T. Kaneko *et al*. [73]		
	Glottal excitation model	L. Juvela *et al*. [77]		
	Glottal waveform model	B. Bollepalli *et al*. [23]		
	Wavenet model	Y. Zhao *et al*. [64]		

TABLE VII. UNSUPERVISED AND SEMI-SUPERVISED LEARNING TECHNIQUES AND THEIR ADVANTAGES

Reference	Learning	Advantages
Y. Ren *et al*. [79], O. Watts *et al*. [80], S. King *et al*. [81]	Unsupervised	1. Easier to get unlabeled data, less complex. 2. Less complex.
Y. A. Chung *et al*. [82]	Semi-supervised	1. Uses both unlabeled and a very small amount of labeled data. 2. Less expensive. 3. To overcome supervised learning problems. 4. Use cheap and abundant unlabeled data.

TABLE VIII. A SUMMARY OF SOME PAPERS BASED ON NON-AUTOREGRESSIVE MODELS FOR TTS SYSTEM

References	Methodology	Dataset	Accuracy	Limitations
[27]	ParaNet, a non-autoregressive seq2seq model that converts text to spectrogram.	English speech	ParaNet performs 46.7 times faster than the lightweight autoregressive Deep voice 3 at synthesis. Also, the speech quality improves significantly.	The proposed model, ParaNet is less robust for parallel neural vocoder.
[28]	Nana-HDR, a new non-attentive non-autoregressive model with hybrid Transformer-based Densefuse encoder and RNN-based decoder for TTS.	Two Mandarin corpora.	Provides better naturalness and robustness than two well performing models: Tacotron and Fastspeech.	N/A
[29]	VARA-TTS, a non-autoregressive (non-AR) end-to-end text-tospeech (TTS) model using a very deep Variational Autoencoder (VDVAE) with Residual Attention mechanism.	LJ Speech corpus and Mandarin Chinese data	VARA-TTS performs better than bidirectional-inference variational autoencoder (BVAE-TTS) with almost same inference speed. Also, it is16x faster than Tacotron 2 with slightly poor performance in naturalness.	The model cannot generate intelligible waveform though it can clearly align text and waveform.

[30]	A hierarchical model to improve the performance of Transformer-based non-autoregressive text to-speech (TNA-TTS) model.	Korean female speech	The proposed model provides better performance than the baseline TNA-TTS.	Instead of using fixed window size, learnable window size can optimize the structure for the encoder and decoder.
[31]	propose a multi-scale time-frequency spectrogram discriminator to help Non-Autoregressive TTS(NAR-TTS) generate high-fidelity Mel-spectrograms.	English speech dataset, LJSpeech	Achieves significant improvement in the naturalness and fidelity.	N/A
[32]	A hierarchical prosody modeling framework that combines the word and phoneme-level prosody features.	LJSpeech dataset	Outperforms other prosody modeling framework in terms of naturalness and audio quality.	Phoneme level predicts less accurate prosody than word level and the quality degrades due to coarse granularity.
[33]	A hybrid TTS which combines the Transformer encoder and RNN based decoder architecture.	A corpus of a female, Mandarin Chinese, native speaker	The new hybrid system outperforms the previous hybrid system [87] (67% vs 11%).	Only one language has been used for the experiment.

TABLE IX. DETAILED DESCRIPTION OF JAPANESE SPEECH DATASET USED FOR TTS

Data source	Sentences	Male/ female speaker	Training data	Testing data	Sampling rate	References
ATR Japanese speech database	503 Phonetically balanced sentences	Single male speaker	450	53	16 kHz	Y. Saito *et al*. [72]
Japanese speech data	7,000 utterances	Professional female narrator	6500	500	22.05 kHz	T. Kaneko *et al*. [73]
Japanese female speaker	4007 sentences	Female	3808	899	16 kHz	Y. Saito *et al*. [78]
Japanese speech database	503 utterances	N/A	450	53	48 kHz	K. Hashimoto *et al*. [49]
Japanese Voice Bank corpus	N/A	Both male and female	11,170 utterances	100 utterances	48 kHz	H. T. Luong, *et al.* [50]

TABLE X. DETAILED DESCRIPTION OF LJ SPEECH DATASET USED FOR TTS

Data source	Sentences	Male/ female speaker	Training data	Testing data	Sampling rate	References
LJ speech	20 sentences from Harvard Sentences List 1 and 2.	N/A	ADAM optimizer	N/A	22050 Hz	H. Tachibana et al. [25]
Open source LJ speech	N/A	Single female English speaker	300 epochs	N/A	N/A	X. Wang *et al*. [66]
LJ speech	13,100 English audio clips and the corresponding transcripts	N/A	12500 samples	300+300 samples	N/A	Y. Ren *et al*. [79]

TABLE XI. DETAILED DESCRIPTION OF CHINESE DATASET USED FOR TTS

Data source	Sentences	Male/ female speaker	Training data	Testing data	Sampling rate	References
Chinese speech corpus	10,000 utterances	Single female speaker	80%	10%	16kHz	S. Yang *et al*. [71]
1-hour Chinese speech database produced by a professional female speaker	1,000 (720 samples)	Single Female Speaker	520 samples	200 samples	16 kHz	Z. H. Ling *et al*. [37]
1-hour Chinese speech database produced by a professional female speaker	1,000	Single Female Speaker	800 samples	200 samples	16 KHz	Z. H. Ling *et al*. [35]

TABLE XII. DETAILED DESCRIPTION OF BRITISH ENGLISH DATASET USED FOR TTS

Data source	Sentences	Male/ female speaker	Training data	Testing data	Sampling rate	References
Professional British English voice talents	2542 utterances from male and 4314 utterances from female	Both male and female	Rest for training	100 utterances from both speakers	16kHz	L. Juvela *et al*. [77]
Professional British English voice talent	4314 utterances	Single female speaker	Rest for training	100 utterances for testing	16kHz	B Bollepalli *et al*. [23]
British English Speaker	1000 utterances	Male speaker	860 samples	140 samples	N/A	H. Lu *et al*. [48]

VI. Evaluation Metrics

For speech synthesis, several evaluation metrics such as Mean Opinion Scores (MOS), F0 Root-Mean-Square Error (RMSE), Voiced/unvoiced (V/UV), Mel cepstral (MCEP) are popularly used and Table XIII provides a detailed description of these evaluation metrics. Among these MOS has been used in maximum papers for objective evaluation. In the domain of speech synthesis, speech quality is the most popular way to compare the performance of any algorithm, architecture, or model. MOS calculates the average score of quality and for this reason, MOS has emerged as the most common and popular figure for synthesized speech quality.

VII. Discussion and Future Research Direction

SPSS and concatenative synthesis are two primary models in TTS technology. Compared with the concatenative speech synthesis method, SPSS systems have many strengths. They are much more flexible to transform the synthesized speech into different speech characteristics, emotions, and speaking ways. Three prime factors that hinder the quality of synthetic speech are the quality of vocoders, acoustic modeling accuracy, and over-smoothing. In SPSS systems, RBMs have been conveniently used for modeling voice signals. For SPSS, the Gaussian mixture model performs less accurately in modeling spectral envelopes' distribution over RBM.RBM replaces single Gaussian distributions by representing the spectral envelope distribution at each HMM state. This procedure significantly upgrades the conventional HMM-based speech synthesis system's naturalness using Mel-cepstra and reduces the over-smoothing problem at synthesis time. After that, DBN has been implemented for presenting the distribution of the spectral envelope at each HMM state and at the same time to include the dynamic features of spectral envelopes into RBM modeling. The experimental results and evaluation process shows that both DBN-HMM and RBM-HMM create better spectral envelope parameter sequences over traditional Gaussian-HMM with better generalization power DBN-HMM and RBM-HMM perform most likely due to the use of Gaussian distribution. Numerous deep learning-based approaches have been suggested in recent years for converting text to speech synthesis, however. Several techniques are mainly used in interpreting text-to-speech, such as CNN and RNN. The critical comparison between CNN and RNN occurs in the training period. Deep Convolutional TTS is only adequately trained at night (15 hours), while the quality of speech was almost appropriate. On the other hand, RNN usually takes a lot of time to train the model for several days or weeks, as it is suitable for a powerful machine and less suitable for GPU parallel computation. Many texts on speech synthesis methods use CNN instead of RNN to resolve this problem. Due to elevated parallelizability, CNN-based text to speech synthesis methods works much faster than RNN-based techniques. In recent times, the methods used for speech synthesis that performs better than traditional approaches have been improved.

TABLE XIII. Widely Used evaluation Metrices and Their Properties

Evaluation Matrices	Properties	References
MOS	MOS is a widely used metric for determining an audio signal's quality which is scored on a scale of 1 (bad) to 5 (excellent). With the increment of MOS score, the quality of generated speech increases. For humans, a score between 4.3 to 4.5 is considered excellent and a score below 3.5 is unacceptable.	[43], [51], [55], [56], [54], [63], [68], [83], [84], [88].
MCEP	Between the Mels, lower frequencies have a larger distance and higher frequencies have a smaller distance, supporting human-like characteristics.	[37], [43], [46], [71], [50], [61].
Voiced/unvoiced (V/UV)	The percentage of unvoiced speech signals that are misclassified as voiced is known as voiced error. Similarly, the percentage of voiced speech signals that are misclassified as unvoiced is known as unvoiced error.	[43], [71], [60], [61], [66], [67].
RMSE	Root-mean-square error (RMSE) is the most commonly used for measuring the differences between sample or population values expected by an estimator and the observed values. The value of RMSE is always positive and 0 indicates that the data is perfectly suited. A lower RMSE is often preferable to a greater one.	[43], [71], [50], [60], [61], [66], [67], [69].

The AdVRNN now implements the variability of natural speech for acoustic modeling in speech synthesis. However, the Adversarial Learning Scheme in AdVRNN training to solve the problem of over-smoothing AdVRNN performs better than traditional speech synthesis based on RNN. HMM-based SPSS has become more popular in the last few years. But the status of the synthesized speech does not reach the naturalness using HMM. So to upgrade the naturalness of synthesized speech, DNN has been used. The acoustic modeling accuracy and the over-smoothing problem has been solved partially by LSTM. But it was found that by using GAN, the over-smoothing problem can be solved fully. For reproducing the stochastic component in the glottal excitations, GAN performs better than DNN. Moreover, the DNN-based GANs perform significantly better than deep CNN-based GAN. While several problems have been solved in recent years, there is still a great potential for development. Here, we have included some of the future research directions:

- A Neural TTS with simple architecture [76] can be applied to other applications such as emotional/ nonlinguistic/ personalized speech synthesis.
- Some advanced vocoder-based models such as Wavenet can be implemented instead of Griffin-Lim for improved quality of the generated audio [79].

- As the MSE loss is still used for stabilizing the adversarial process, different architectures using Wasserstein GAN [89] and VAEGAN [90] can be implemented to directly evaluate the distribution of synthesized speech.
- To analyze the performance of different types of GANs such as Conditional GAN (CGAN), Mel Spectrogram GAN, Wasserstein GAN [89], VAEGAN, speech enhancement GAN (SEGAN), iterated SEGAN (ISEGAN), and deep SEGAN (DSEGAN) can be compared for further improvement of the naturalness of synthetic speech.

VIII. Conclusion

The research interest in speech synthesis has been changing from clarity and intelligibility to expressiveness and naturalness. In the beginning period of speech synthesis, parametric synthesis and concatenative-based methods have been mainly used, which hinders the naturalness of synthesized speech. Diversely, deep learning-based speech synthesis methods focus on the naturalness of speech. In this paper, we have explored deep learning-based speech synthesis methods. We have inclined a taxonomy of speech synthesis methods, showing a generic block diagram of major architectures and models popularly used to synthesize speech and highlighting their advantages and disadvantages. Different evaluation metrics and datasets with their advantages and disadvantages have been discussed. A comparison of the various experiments was presented. Even if deep learning-based speech synthesis methods have achieved tremendous success in recent years, there is still a big opportunity to produce high-quality speech from text. With the current advent of lightweight deep neural network architectures, speech can be synthesized automatically from the text in the near future.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

The research is conducted under the supervision of M. F. Mridha and A. K. Saha. The initial draft is written by F. Khanam, F. A. Munmun, and N. A. Ritu. The final version is written by F. Khanam and F. A. Munmun.

Acknowledgment

The authors would like to thank the Institute of Energy, Environment, Research, and Development (IEERD, UAP) and the University of Asia Pacific for financial support.

References

[1] H. Sak, T. Gung, and Y. Safkan, "A corpus-based concatenative speech synthesis system for Turkish," *Turkish Journal of Electrical Engineering Computer Sciences*, vol. 14, no. 2, pp. 209-223, 2006.

[2] G. Lopez, L. Quesada, and L. A. Guerrero, "Alexa vs. siri vs. Cortana vs. Google assistant: A comparison of speech-based natural user interfaces," in *Proc. International Conference on Applied Human Factors and Ergonomics*, May 2018, pp. 241-250.

[3] R. A. Khan and J. S. Chitode, "Concatenative speech synthesis: A review," *International Journal of Computer Applications*, vol. 136, no. 3, p. 0975-8887, 2016.

[4] Y. Ning, S. He, Z. Wu, C. Xing, and L. J. Zhang, "A review of deep learning based speech synthesis," *Appl. Sci.*, vol. 9, no. 19, 2019.

[5] H. Zena, K. Tokuda, and A. W.Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039-1064, 2009.

[6] S. Kayte, M. Mundada, and J. Gujrathi, "Hidden markov model based speech synthesis: A review," *International Journal of Computer Applications*, vol. 130, no. 3, pp. 35-39, 2015.

[7] S. K. D. Mandal and A. K. Datta, "Epoch synchronous non-overlapadd (ESNOLA) method-based concatenative speech synthesis system for bangla," *SSW*, pp. 351-355, 2007.

[8] S. Norbert, G. Peeters, S. Lemouton, P. Manoury, and X. Rodet, "Synthesizing a choir in real-time using pitch synchronous overlap add (PSOLA)," *ICMC*, 2000.

[9] S. A. Toma, G. I. Tarsa, E. Oancea, D. P. Munteanu, F. Totir, and L. Anton, "A TD-PSOLA based method for speech synthesis and compression," in *Proc. 8th International Conference on Communications*, June 2010, pp. 241-250.

[10] W. Mattheyses, W. Verhelst, and P. Verhoeve, "Robust pitch marking for prosodic modification of speech using TD-PSOLA," in *Proc. IEEE Benelux/DSP Valley Signal Processing Symposium, SPS-DARTS*, June 2006, pp. 43-46.

[11] A. Gopi, T. Sajini, and V. K. Bhadran, "Implementation of malayalam text to speech using concatenative based TTS for android platform," in *Proc. International Conference on Control Communication and Computing*, December 2013.

[12] A. Shipilo, A. Barabanov, and M. Lipkovich, "Parametric speech synthesis and user interface for speech modification," in *Proc. International Conference on Speech and Computer*, December 2013, pp. 249-256.

[13] A. Indumathi and E. Chandra, "Survey on speech synthesis," *Signal Processing: An International Journal (SPIJ)*, vol. 6, no. 5, p. 140, 2012.

[14] K. Sangramsing, K. Waghmare, and B. Gawali, "Marathi speech synthesis: A review," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 3, no. 6, pp. 3708-3711, 2015.

[15] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," in *Proc. Sixth European Conference on Speech Communication and Technology*, 1999.

[16] S. Afrooz and N. J. Navimipour, "Memory designing using quantumdot cellular automata: Systematic literature review, classification and current trends," *Journal of Circuits, Systems and Computers*, vol. 26, no. 12, p. 1730004, 2017.

[17] B. A. Milani and N. J. Navimipour, "A systematic literature review of the data replication techniques in the cloud environments," *Big Data Research*, vol. 10, pp. 1-7, 2017.

[18] A. Luxton-Reilly, "A systematic review of tools that support peer assessment," *Computer Science Education*, vol. 19, no. 4, pp. 209-232, 2009.

[19] F. Aznoli and N. J. Navimipour, "Deployment strategies in the wireless sensor networks: Systematic literature review, classification, and current trends," *Wireless Personal Communications*, vol. 95, no. 2, p. 819-846, 2017.

[20] F. Aznoli and N. J. Navimipour, "Cloud services recommendation: Reviewing the recent advances and suggesting the future research directions," *Journal of Network and Computer Applications*, vol. 77, pp. 73-86, 2017.

[21] Z. Soltani and N. J. Navimipour, "Customer relationship management mechanisms: A systematic review of the state of the art literature and recommendations for future research," *Computers in Human Behavior*, vol. 61, pp. 667-688, 2016.

[22] R. K. Behera, P. KumarBala, and A. Dhir, "The emerging role of cognitive computing in healthcare: A systematic literature review," *International Journal of Medical Informatics*, vol. 129, pp. 154-166, 2019.

[23] B. Bollepalli, L. Juvela, and P. Alku, "Generative adversarial networkbased glottal waveform model for statistical parametric speech synthesis," arXiv preprint arXiv:1903.05955, 2019.

[24] H. Zen and H. Sak, "Implementation of malayalam text to speech using concatenative based tts for android platform," in *Proc. IEEE*

International Conference on Acoustics, Speech and Signal Processing, April 2015.
[25] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable textto-speech system based on deep convolutional networks with guided attention," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2018, pp. 4784-4788.
[26] G. Wang, "Deep text-to-speech system with seq2seq model," arXiv preprint arXiv:1903.05955, 2019.
[27] K. Peng, W. Ping, Z. Song, and K. Zhao, "Non-autoregressive neural text-to-speech," in *Proc. International Conference on Machine Learning*, 2020, pp. 7586-7598.
[28] S. Lin, W. Su, L. Meng, F. Xie, X. Li, and L. Lu, "Nana-HDR: A non-attentive non-autoregressive hybrid model for TTS," arXiv preprint arXiv:2109.13673, 2021.
[29] P. Liu, Y. Cao, S. Liu, N. Hu, G. Li, C. Weng, and D. Su, "Vara-TTS: Non-autoregressive text-to-speech synthesis based on very deep vae with residual attention," arXiv preprint arXiv:2102.06431, 2021.
[30] J. S. Bae, T. J. Bak, Y. S. Joo, and H. Y. Cho, "Hierarchical context-aware transformers for non-autoregressive text to speech," arXiv preprint arXiv:2106.15144, 2021.
[31] H. Guo, H. Lu, X. Wu, and H. Meng, "A multi-scale time-frequency spectrogram discriminator for GAN-based non-autoregressive tts," arXiv preprint arXiv:2203.01080, 2022.
[32] C. M. Chien and Y. H. Lee, "Hierarchical prosody modeling for non-autoregressive speech synthesis," in *Proc. IEEE Spoken Language Technology Workshop*, 2021, pp. 446-453.
[33] F. L. Xie, X. H. Li, W. C. Su, L. Lu, and F. K. Soong, "A new high quality trajectory tiling based hybrid TTS in real time," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 5704-5708.
[34] H. Zen, K. Tokudaa, and A. W.Blackc, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039-1064, 2009.
[35] Z. H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2129-2139, 2013.
[36] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7962-7966.
[37] Z. H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted boltzmann machines for statistical parametric speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7825-7829.
[38] M. Tamurat, T. Masukot, K. Tokudatt, and T. Kobayashil, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 805-808.
[39] S. Kang, X. Qian, and H. Meng, "Multi-distribution deep belief network for speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8012-8016.
[40] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "F0 contour prediction with a deep belief network-gaussian process hybrid model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6885-6889.
[41] A. R. Mohamed, G, E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14-22, 2011.
[42] Y. J. Hu and Z. H. Ling, "DBN-based spectral feature representation for statistical parametric speech synthesis," *IEEE Signal Processing Letters*, vol. 23, no. 3, pp. 321-325, 2016.
[43] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2014, pp. 3844-3848.
[44] S. H. Chen, S. H. Hwang, and C. Y. Tsai, "A first study on neural net based generation of prosodic and spectral information for mandarin text-to-speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1992, pp. 45-48.
[45] T. Falas and A. G. Stafylopatis, "Neural networks in text-to-speech systems for the greek language," In *Proc. 10th Mediter-ranean Electrotechnical Conference. Information Technology and Electrotechnology for the Mediterranean Countries*, 2000, pp. 574-577.
[46] I. Rebai and Y. BenAyed, "Arabic text to speech synthesis based on neural networks for mfcc estimation," in *Proc. World Congress on Computer and Information Technology*, June 2013, pp. 1-5.
[47] T. Raitio, A. Suni, L. Juvela, M. Vainio, and P. Alku, "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort," in *Proc. Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 1969-1973.
[48] H. Lu, S. King, and O. Watts, "Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis," in *Proc. Eighth ISCA Workshop on Speech Synthesis*, 2013, pp. 261-265.
[49] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "The effect of neural networks in statistical parametric speech synthesis," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.
[50] H. T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling dnn-based speech synthesis using input codes," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2017, pp. 4905-4909.
[51] A. H. Ali, M. Magdy, M. Alfawzy, M. Ghaly, and H. Abbas, "Arabic speech synthesis using deep neural networks," in *Proc. International Conference on Communications, Signal Processing, and their Applications*, 2021, pp. 1-6.
[52] A. Alastalo, *et al.*, "Finnish end-to-end speech synthesis with tacotron 2 and wavenet," 2021.
[53] W. Guo-liang, C. Meng-nan, and C. Lei, "An end-to-end Chinese speech synthesis scheme based on Tacotron 2," *Journal of East China Normal University (Natural Science Edition)*, no. 4, pp. 111-119, 2019.
[54] Y. Li, D. Qin, and J. Zhang, "Speech synthesis method based on tacotron2," in *Proc. 13th International Conference on Advanced Computational Intelligence*, 2021, pp. 94-99.
[55] S. Takamichi, "Modulation spectrum-based speech parameter trajectory smoothing for dnn-based speech synthesis using fft spectra," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2017, pp. 1308-1311.
[56] L. Chen, H. Yang, and H. Wang, "Research on dungan speech synthesis based on deep neural network," in *Proc. 11th International Symposium on Chinese Spoken Language Processing*, 2018, pp. 46-50.
[57] S. Suzie, T. Nosek, M. Secujski, D. Pekar, and V. Delie, "DNN based expressive text-to-speech with limited training data," in *Proc. 27th Telecommunications Forum*, 2019, pp. 1-6.
[58] S. H. Chen, S. H. Hwang, and Y. R. Wang, "An RNN-based prosodic information synthesizer for mandarin text-to-speech," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 226-239, 1998.
[59] S. Achanta, T. Godambe, and S. V. Gangashetty, "An investigation of recurrent neural network architectures for statistical parametric speech synthesis," in *Proc. Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
[60] H. Choi, S. Park, J. Park, and M. Hahn, "Multi-speaker emotional acoustic modeling for cnn-based speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6950-6954.
[61] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust textto-speech," *SSW*, 2016.
[62] M. Aso, S. Takamichi, N. Takamune, and H. Saruwatari, "Acoustic model-based subword tokenization and prosodic-context extraction without language knowledge for text-to-speech synthesis," *Speech Communication*, vol. 125, pp. 53-60, 2020.
[63] J. Y. Lee, S. J. Cheon, B. J. Choi, N. S. Kim, and E. Song, "Acoustic modeling using adversarially trained variational recurrent neural network for speech synthesis," in *Proc. INTERSPEECH*, 2018, pp. 917-921.
[64] Y. Zhao, S. Takaki, H. T. Luong, J. Yamagishi, D. Saito, and N. Minemats, "Wasserstein gan and waveform loss-based acoustic model training for multi-speaker text-to-speech synthesis systems

using a wavenet vocoder," *IEEE Access*, vol. 6, pp. 60478-60488, 2018.
[65] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7962-7966.
[66] X. Wang, S. Takaki, and J. Yamagishi, "An RNN-based quantized f0 model with multi-tier feedback links for text-to-speech synthesis." in *Proc. INTERSPEECH*, 2017, pp. 1059-1063.
[67] Y. Fan, Y. Qian, F. L. Xie, and F. K. Soong, "TTS synthesis with bidirectional lstm based recurrent neural networks," in *Proc. Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
[68] S. Shechtman and M. Mordechay, "Emphatic speech prosody prediction with deep lstm networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5119-5123.
[69] C. Batista, R. Cunha, P. Batista, A. Klautau, and N. Neto, "Utterance copy in formant-based speech synthesizers using lstm neural networks," in *Proc. 8th Brazilian Conference on Intelligent Systems*, 2019, pp. 90-95.
[70] H. Zen, "Generative model-based text-to-speech synthesis," in *Proc. IEEE 7th Global Conference on Consumer Electronics*, 2018, pp. 327-328.
[71] S. Yang, L. Xie, X. Chen, X. Lou, X. Zhu, D. Huang, and H. Li, "Statistical parametric speech synthesis using generative adversarial networks under a multi-task learning framework," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, December 2017, pp. 685-691.
[72] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, p. 84-96, 2017.
[73] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, June 2017, pp. 4910-4914.
[74] A. V. D. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.
[75] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4779-4783.
[76] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-Dependent wavenet vocoder," *Interspeech*, vol. 2017, pp. 1118-1122, 2017.
[77] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "Waveform generation for text-to-speech synthesis using pitch-synchronous multiscale generative adversarial networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2019, pp. 6915-6919.
[78] Y. Saito, S. Takamichi, and H. Saruwatari, "Text-to-Speech synthesis using stft spectra based on low-/multi-resolution generative adversarial networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2018, pp. 5299-5303.
[79] Y. Ren, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Y. Liu, "Almost unsupervised text to speech and automatic speech recognition," in *Proc. International Conference on Machine Learning*, 2019, pp. 5410-5419.
[80] O. Watts, A. Stan, R. A. Clark, Y. Mamiya, M. Giurgiu, J. Yamagishi, and S. King, "Unsupervised and lightly-supervised learning for rapid construction of tts systems in multiple languages from 'found'data: evaluation and analysis," in *Proc. Eighth ISCA Workshop on Speech Synthesis*, 2013.
[81] S. King, K. Tokuda, H. Zen, and J. Yamagishi, "Unsupervised adaptation for hmm-based speech synthesis," ISCA, 2008.
[82] Y. A. Chung, Y. Wang, W. N. Hsu, Y. Zhang, and R. Skerry-Ryan, "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6940-6944.
[83] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Tacotron-Based acoustic model using phoneme alignment for practical neural text-to-speech systems," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, 2019, pp. 214-221.
[84] S. Li, B. Ouyang, L. Li, and Q. Hong, "Lightspeech: Lightweight non-autoregressive multi-speaker text-to-speech," in *Proc. IEEE Spoken Language Technology Workshop*, 2021, pp. 499-506.
[85] M. Bi, H. Lu, S. Zhang, M. Lei, and Z. Yan, "Deep feed-forward sequential memory networks for speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4794-4798.
[86] K. Ito and L. Johnson. (2017). The LJ speech dataset. [Online] Available: https://keithito.com/LJ-Speech-Dataset/
[87] Y. Qian, F. K. Soong, and Z. J. Yan, "A unified trajectory tiling approach to high quality speech rendering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 280-290, 2012.
[88] K. Azizah, M. Adriani, and W. Jatmiko, "Hierarchical transfer learning for multilingual, multi-speaker, and style transfer dnn-based tts on lowresource languages," *IEEE Access*, vol. 8, pp. 179798-179812, 2020.
[89] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. International Conference on Machine Learning*, 2017, pp. 214-223.
[90] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed, "Variational approaches for auto-encoding generative adversarial networks," arXiv preprint arXiv:1706.04987, 2017.

F. Khanam completed her B.Sc. degree in Computer Science and Engineering from Military Institute of Science and Technology (MIST) and M.Sc. in Information and Communication Technology at Bangladesh University of Engineering and Technology (BUET). She is currently working as a Lecturer of department of Computer Science and Engineering at Bangladesh University of Business and Technology (BUBT), Dhaka, Bangladesh. Her research interest includes Machine Learning, Data Mining, Bioinformatics, Network Security, and Natural Language Processing (NLP).

Farha Akhter Munmun received her B.Sc. degree in Computer Science and Engineering from Rajshahi University of Engineering and Technology (RUET). She is currently doing her M.Sc. in Computer Science and Engineering at Bangladesh University of Engineering and Technology (BUET). She is working as a Lecturer in the Department of CSE, BUBT. She has experience working in RAXML, PAUP, Python, and Sklearn. Her research interest includes Bioinformatics, Network Security, Machine Learning, and Natural Language Processing (NLP).

Nadia Afrin Ritu completed her both Bachelor of Science (B.Sc) and Master of Science (M.Sc) in Computer Science and Engineering from Jahangirnagar University, Savar, Dhaka, Bangladesh. She is currently working as a Lecturer of department of computer science and engineering at Bangladesh University of Business and Technology, Dhaka, Bangladesh. She is very much interested in the field of artificial neural network, machine learning, data mining and image processing. She specially enjoys learning and her job of teaching.

Dr. Aloke Kumar Saha is Professor of Computer Science and Engineering (CSE) Department of University of Asia Pacific (UAP), Dhaka, Bangladesh. He joined UAP on March 1999 as a Lecturer. Before, he was a Lecturer at Queens University from June 1997 to March 1999. He completed B.Sc. (Hons.) in Applied Physics & Electronics from the University of Dhaka in 1995. He received his M.Sc. (Thesis) in Computer Science from University of Dhaka in 1997. Dr. Saha received his Ph.D. in Computer Science and Engineering from Jahangirnagar University, Savar, Dhaka, Bangladesh. He has thirty (30) Journal papers and twenty five (25) Conference papers. He usually teaches courses on Digital Logic & System Design, Numerical Methods, Data Structures, Discrete Mathematics and Computer Graphics. His current research interest includes Algorithm, Artificial Intelligence, Machine Learning, Deep Learning and Natural Language Processing. For more than 25 (Twenty five) years, he is working with the undergraduate and master's students as a supervisor or co-researcher of their project and thesis works. Dr. Saha has worked as the head of the CSE department, UAP from 2008 to 2018. He was the Chair of organizing committee of International Conference on Computer and Information Technology (ICCIT) 2017. He was the Contest Director of National Collegiate Programming Contest (NCPC) 2016. Under his leading UAP host the International Collegiate Programming Contest (ICPC) 2106 and 2017. He is Chief of Organizing Committee of International Journal of Computer and Information Technology (IJCIT), published by department of CSE, UAP. He is reviewer of different conferences and journals.

M. F. Mridha (Senior Member, IEEE) received the Ph.D. degree in AI/ML from Jahangirnagar University, in 2017. He joined the Department of Computer Science and Engineering, Stamford University Bangladesh, in June 2007, as a Lecturer, where he was promoted a Senior Lecturer and an Assistant Professor, in October 2010 and October 2011, respectively. Then, he joined UAP, in May 2012, as an Assistant Professor. He is currently working as an associate professor with the Department of Computer Science and Engineering, Bangladesh University of Business and Technology. He also worked as a faculty member of the CSE Department, University of Asia Pacific, and as a graduate coordinator, from 2012 to 2019. His research experience, within both academia and industry, has resulted in over 80 journal and conference publications. His research interests include artificial intelligence (AI), machine learning, deep learning, big data analysis, and natural language processing (NLP). For more than ten years, he has been with the master's and undergraduate students, as a supervisor of their thesis work. He has served as a program committee member of several international conferences and workshops. He also served as an associate editor in several journals.