SEYBOLD Report
ISSN: 1533 - 9211
https://seyboldreport.net/

SPEECH SYNTHESIS METHODS USING DEEP LEARNING METHODS: A REVIEW

Jyoti Seth[1], Shobha Bhatt[2]
[1]Dept. of Computer Science Engineering, NSUT, Delhi-110078, India
[2]Dept. of Computer Science Engineering, NSUT, Delhi-110078, India
[1]jyoti.seth.pg21@nsut.ac.in [2]Shobha.bhatt@nsut.ac.in

Abstract—This review paper investigates advances in speech synthesis using deep learning approaches. For many years, speech synthesis has been an important topic of research, and with recent advances in deep learning, new ways to generating more natural-sounding speech have been proposed. The study presents an introduction of several deep learning approaches used for speech synthesis, such as Generative Adversarial Networks (GANs), WaveNet, and Tacotron,Deep Bidirectional Long-short term memory(DBLSTM). It also highlights the difficulties that researchers encounter, such as the availability of training data, model complexity, and evaluation criteria. Finally, the paper concludes with potential future avenues for deep learning-based speech synthesis research.

Keywords- Speech Synthesis, WaveNet, Tacotron, Generative Adversarial Networks (GAN),Deep Bidirectional Long-short term memory(DBLSTM).

I. INTRODUCTION

Text-to-speech (TTS) conversion, often known as speech synthesis, is the process of generating natural-sounding speech from text. With the development of speech synthesis technologies, from the previous formant based parametric synthesis [1,2], waveform concatenation based methods [3][4] to the current statistical parametric speech synthesis (SPSS) [6], the intelligibility and naturalness of the synthesized speech have been improved greatly. Deep learning has sparked a surge of interest in examining its potential for speech synthesis. Deep learning has proven its ability to learn complicated patterns and produce high-quality results in a variety of applications, including speech processing.

Deep learning algorithms for speech synthesis, such as Generative Adversarial Networks (GANs), WaveNet, and Tacotron, Deep bidirectional Long Short-term memory (DBLSTM) have been proposed by researchers in recent years. These approaches have yielded encouraging results in terms of producing natural-sounding speech that closely resembles human speech. Furthermore, deep learning techniques have enabled speech synthesis in different languages and styles, making it a helpful tool for a variety of applications such as text-to-speech systems, virtual assistants, and the entertainment industry.

Despite the positive results, there are still hurdles in deep learning-based speech synthesis research. One significant problem is the availability and quality of training data. Another problem is the complexity of deep learning models and the amount of compute required for training. Furthermore, measuring the quality of synthesised speech is subjective, thus objective assessment criteria are required to precisely quantify the quality of synthesised speech.This review paper seeks to provide an overview of current advances in speech synthesis using deep

learning techniques. It will go over the various deep learning approaches used for voice synthesis, the obstacles that researchers confront, and probable future avenues for study in this subject.

The remaining part of the paper is structured as follows.Section 2 describes the Literature Review . Section 3 describes an overview of speech synthesis in that we have discussed about the various advantages and challenges occurs during the research. Section 4 describes the various methods used for speech synthesis like WaveNet, Tacotron , Generative Adversarial Networks(GAN),Deep Bidirectional Long-short term memory(DBLSTM).Section 5 describes the Result and Discussion of different techniques followed by a comparative analysis of different techniques .Finally, conclusions and future work suggestions are presented in Section 6.

II. LITERATURE REVIEW

Several studies have explored the use of deep learning techniques for speech synthesis, and their results have been promising. In this literature review, we will discuss some of the most significant contributions to the field of speech synthesis using deep learning.

Van den Oord et al. (2016) proposed WaveNet, a deep learn- ing model for speech synthesis[6]. WaveNet is a generative model that can generate speech one sample at a time while taking into account the history of preceding samples. This method produces high-quality, natural-sounding speech.

Arik et al. (2017), for instance; offered a modified version of WaveNet that could generate speech in real-time, opening up new opportunities for interactive applications. [7]

Sotelo et al. (2017) suggested a multi-speaker WaveNet model that could create speech for numerous speakers, in- creasing the model's versatility and adaptability [8]. WaveNet has also been used in conjunction with other deep learning models to generate speech.

Shen et al (2018) "A Fully Convolutional Neural Network for Speech Synthesis" Based on the WaveNet architecture, this study developed a fully convolutional neural network for speech synthesis. In comparison to the original WaveNet model, the authors demonstrated that their model could gener- ate high-quality speech waveforms with fewer parameters and lower computing cost [8].

Wang et al. (2018), for example, offered a hybrid technique that used WaveNet to generate the mel-spectrogram, which was then fed into a neural vocoder to synthesise speech. This method enabled the production of high-quality, natural- sounding speech [9].

Wang et al. (2017) developed a two-stage architecture consisting of a text-to-mel-spectrogram model and a WaveNet vocoder in their initial Tacotron work. The text-to-mel- spectrogram model generates mel-spectrograms from input text, which are subsequently utilised to synthesise voice by the WaveNet vocoder. The authors exhibited Tacotron's ability to generate high-quality, natural-sounding speech[10].

Tacotron 2, Shen et al. (2018) updated the WaveNet vocoder with a WaveRNN-based neural vocoder. This change made speech synthesis faster and increased the quality of the gen- erated speech[11].

Kim et al. (2018) proposed a Tacotron architecture update that introduced an extra encoder network to improve the alignment between input text and speech attributes [12].
Ren et al. (2019) suggested a novel technique for directly producing mel-spectrograms from input text using a feed- forward transformer network in FastSpeech. This change en- abled faster and more efficient speech synthesis than previous Tacotron versions [13].
Zhang et al. (2019) introduced a Tacotron architecture mod- ification that enabled for speech synthesis in various languages by employing a shared phoneme embedding space [14].
Goodfellow et al. (2014) presented GANs as a framework for training generative models in a seminal publication. GANs function by training two neural networks, a generator and a discriminator, in a two-player minimax game. The generator generates fictitious data, whereas the discriminator attempts to discern between genuine and fictitious data. The two networks are trained iteratively, with the generator attempting to gener- ate data that the discriminator cannot differentiate from actual data and the discriminator attempting to separate genuine data from generated data.[15]
Donahue et al. (2018) proposed WaveGAN, a GAN-based model for producing raw audio samples, in their study. Wave- GAN produced high-quality speech that was almost indistin- guishable from natural speech [16].
Karras et al. (2020) proposed Hi-Fi GAN, a GAN-based model for creating high-fidelity voice, in their work. Hi-Fi GAN was able to synthesise speech at a sampling rate of 24 kHz, which is greater than the sampling rate of most existing GAN-based speech synthesis models[17].
Zhang et al. (2019) proposed Spectrogram-based Adver- sarial Generative Network (SAGAN), a GAN-based model for generating mel-spectrograms from input text, in their study. SAGAN was capable to producing highquality, natural- sounding speech [18].
Arik et al. (2017) published ”Real-time Neural Text-to- Speech”: The Deep Voice model, a deep neural network architecture based on a sequence-to-sequence framework with attention mechanisms and deep Bi-LSTM layers for creating high-quality speech, was introduced in this study. After being trained on a vast corpus of speech data, the model displayed real-time synthesis skills with human-like speech quality [19].
Y. Wang et al. (2017) published ”Tacotron: Towards End-to- End Speech Synthesis”: This study offered an end-to-end neu- ral text-to-speech model with attention mechanisms and deep Bi-LSTM layers based on a sequence-to-sequence framework. The model was trained on a large dataset of speech recordings and produced cutting-edge results in terms of naturalness and likeness to the target speaker [20].
Deep Voice 2: Multi-Speaker Neural Text-to-Speech” by A. Arik et al. (2017): The Deep Voice model has been revised to accommodate multi-speaker synthesis in this article. The authors developed a speaker embedding layer that learns a low- dimensional representation of each speaker’s speech, which is subsequently used as input to the deep Bi-LSTM layers. The model was trained on a huge dataset of speech recordings from numerous speakers and displayed high-quality and diversified synthesis capabilities [7].
Y. Ren et al. (2019): ”FastSpeech: Fast, Robust, and Con- trollable Text to Speech”: A feed-forward transformer model with a duration predictor and a deep Bi-LSTM-based acoustic

predictor was proposed in this work as a novel approach to voice synthesis. The model was trained on a large dataset of speech recordings and was able to provide high-quality, controlled output in real time [13].

W. Ping et al. (2018) published ”Neural Speech Synthesis with Transformer Network”: A transformer-based speech syn- thesis model that blends self-attention mechanisms with deep Bi-LSTM layers was proposed in this paper. The authors also incorporated a novel training aim called maximum likelihood with teacher forcing, which enhanced the quality of the synthesised speech greatly. The model was trained on a large dataset of speech recordings and produced cutting-edge results in terms of naturalness and likeness to the target speaker [21].

III. AN OVERVIEW OF SPEECH SYNTHESIS

A. Basic Overview of Speech Synthesis

Text-to-speech (TTS) conversion, often known as speech synthesis, is the process of generating natural-sounding speech from text. It has several uses, including virtual assistants, accessibility tools, the entertainment industry, and education. Individuals with speech problems or those who require speech synthesis to communicate effectively can benefit from speech synthesis. It can also help people with visual impairments by providing an alternative way to absorb text-based content.It is a cutting-edge technology in the field of information processing

[22] , especially for the current intelligent speech interac- tion systems. Several approaches for speech synthesis have been developed throughout the years, including rulebased, concatenative, and parametric synthesis. While these methods have been successful in generating speech, they have limits in producing natural-sounding speech that closely resembles human speech. Deep learning has made great progress in speech synthesis in recent years, enabling a more advanced and effective technique to generating natural-sounding speech. Figure 1 shows the basic diagram of speech synthesis, which consists of text as input, which is fed to text analysis, phonetic analysis, prosodic analysis, and speech synthesis, which will provide speech as output.

B. Advantages of Speech synthesis

1. Accessibility: Through TTS, users with visual impair- ments or reading challenges can read the written article. TTS

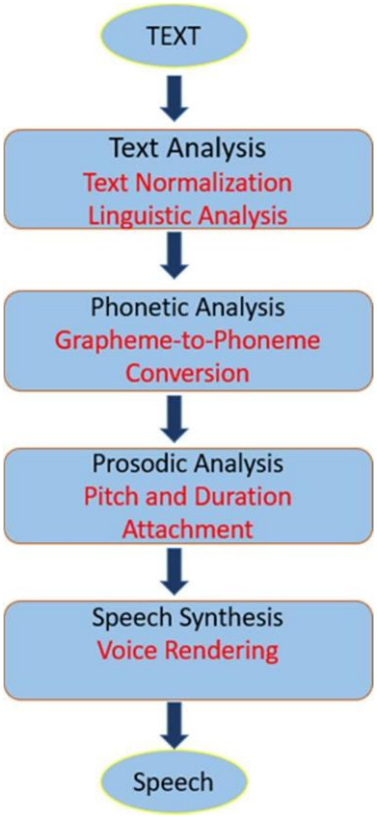

Fig. 1. :
Basic Diagram of Speech Synthesis

allows those individuals to access the content of the research paper without having to rely on others to read it aloud to them.

2. Proofreading: Hearing the material read aloud by a TTS system might aid in the identification of errors or unusual phrasings that may have been overlooked during the editing process. This is especially beneficial for spotting faults that the writer's eye may miss.

3. Multitasking: Reading while listening to the TTS system can help readers save time and increase productivity. A re- searcher, for example, could listen to the TTS system while travelling, exercising, or performing housework.

4. Language Learning: TTS systems can also assist language learners in improving their pronunciation and comprehension. Learners can improve their listening skills and learn accurate pronunciation by listening to the system read the text aloud.

5. Customization: Many TTS systems allow users to change the voice and tempo of the speech to make it simpler to comprehend and more enjoyable to listen to.

C. Challenges of Speech synthesis

Some of the issues in speech synthesis are as follows:

i. There are various problems in text analysis that involve text pre-processing, such as numerals, abbreviation.

ii. Today, correct prosody and pronunciation analysis from written text is also a major issue.

iii. Speech data recording circumstances.

iv. There are no explicit emotions in the written language, and the pronunciation of proper and foreign names is fre- quently exceedingly strange.

v. Language-specific and feature extraction issues exist. Furthermore, the number of prospective customers and mar- ketplaces varies greatly between countries and languages, affecting the number of resources available for creating voice synthesis.

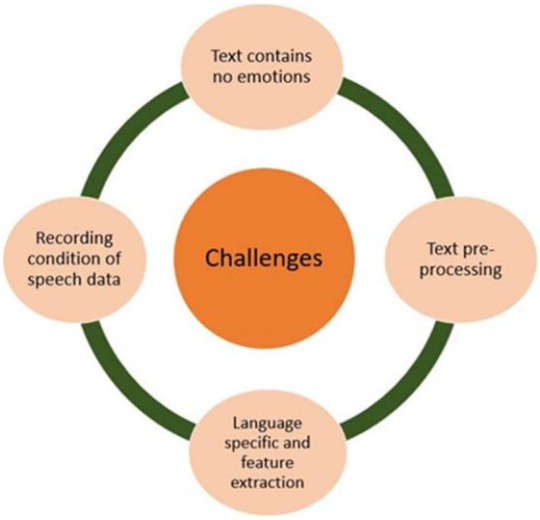

Fig. 2. Challenges of Speech Synthesis

D. History of Speech Synthesis

Text-to-speech (TTS), often known as speech synthesis, is a technology that allows robots to synthesise human-like speech using artificial intelligence algorithms. Speech synthesis dates back to the early 18th century, when French scientist Julien La Mettrie developed the concept of a ”talking machine.” Speech synthesis technology has advanced significantly since then, with new techniques being developed to improve the quality and naturalness of synthesised speech.

Early voice synthesis techniques focused on rule-based approaches, in which sets of rules were utilised to synthe- sise speech based on phonetic and grammatical principles. However, the capacity of these algorithms to generate natural- sounding speech remained limited, and their applicability was primarily limited to simple applications such as weather forecasts and stock quotes.

The advancement of digital signal processing (DSP) and speech analysis algorithms in the 1980s and 1990s resulted in the birth of more complex approaches for speech synthesis, such as formant synthesis and concatenative synthesis. To generate speech sounds, formant synthesis employs mathemat- ical models of the human vocal tract, whereas concatenative synthesis employs prerecorded speech segments.

To increase the naturalness and quality of synthesised speech, statistical parametric synthesis approaches such as hidden Markov models (HMM) [23] and deep neural networks (DNN) [6]have recently been created. These strategies enable

the model to understand the relationship between the input text and the related voice signal, resulting in very natural and expressive speech synthesis.

Speech synthesis has seen substantial growth in recent years, with applications ranging from virtual assistants and auto- mated customer service to accessibility solutions for those with speech impairments. However, the technique raises ethical difficulties, particularly in the context of voice cloning and the possibility of misapplication. Overall, the history of speech synthesis demonstrates the continual development and refining of systems for producing natural and expressive speech using artificial intelligence.

IV. SPEECH SYNTHESIS TECHNIQUES USING DEEP LEARNING

There are various techniques for speech synthesis using deep learning like - WaveNet, Tacotron, Generative Adver- sarial Networks(GAN),Deep Bidirectional Long Short term memory(DBLSTM) and many more. In this section we will discuss some of the techniques.

A. WaveNet

WaveNet is a deep generative audio synthesis model capable of producing high-quality audio waveforms. It is notable for its capacity to record complicated audio patterns and provide realistic audio output because it is built on a deep convolutional neural network architecture.

One of WaveNet's primary advantages is its capacity to create speech using a probabilistic model that captures the distribution of speech waveforms. This is accomplished by employing a deep neural network with dilated convolutions, which can capture long-range relationships in the speech stream. Several studies have investigated the use of WaveNet for speech synthesis, with encouraging results.

Despite the promising results, implementing WaveNet for speech synthesis presents certain obstacles. The computational cost of training and inference, which might be prohibitive for real-time applications, is one of the major problems. Furthermore, WaveNet necessitates a considerable amount of training data, which may be insufficient for some applications. Moreover, WaveNet is a powerful deep learning model for voice synthesis with promising outcomes. While there are still obstacles to overcome, such as processing cost and data availability, WaveNet has opened up new possibilities in speech synthesis and has the potential to improve the quality and naturalness of synthesised speech. Figure 3 shows the basic diagram of WaveNet model is represented by an input waveform that is fed into a stack of dilated causal convo- lutional layers. Each layer has several dilated convolutions with increasing dilation factors, allowing the receptive field to develop exponentially. Within each layer, gated activation units are used to combine the outputs of parallel convolutional pathways. Deep network residual connections serve to reduce training issues, while skip connections connect early layers directly to the output. The output layer converts the output of the final layer into the desired waveform representation.

WaveNet is taught to minimise the difference between the predicted and target waveforms during training. WaveNet can capture complicated audio patterns using this design.

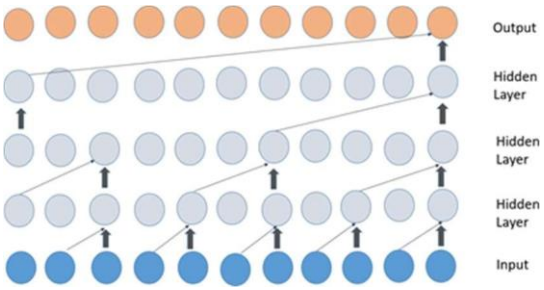

Fig. 3. Basic diagram of the WaveNet Model

B. Tacotron

Tacotron is a deep learning technique for speech synthesis that uses a sequence-to-sequence neural network to synthesise speech from text input.

Tacotron is an effective and extensively used deep learning technique for speech synthesis. To increase the quality and ef- ficiency of speech synthesis, researchers have proposed many tweaks and upgrades to the Tacotron design over time. The us- age of neural vocoders, extra encoder networks, feed-forward transformer networks, and shared phoneme embedding spaces are among the modifications. Figure 4 shows the basic diagram of Tacotron model in which there are three Layers that is Encoder,Decoder and post-processing which gives output in the form of waves and with the help of vocoders, it will generate sound waveforms.

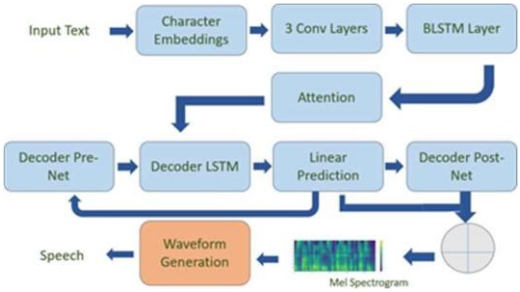

Fig. 4. Basic diagram of the Tacotron Model

C. Generative Adversarial Networks(GAN)

GANs are a form of deep learning model that has gained prominence in recent years due to their capacity to generate realistic data such as images, text, and speech.

While GANs have showed potential for speech synthesis, they do have certain drawbacks, such as difficulties in training, lack of interpretability, and limited control over the generated speech. Nonetheless, GAN-based speech synthesis is still an active area of research, with many researchers looking for

novel ways to increase the quality and efficiency of GAN- based speech synthesis.

Figure 5 show A GAN, or Generative Adversarial Network, is made up of two parts: a generator and a discriminator. The generator takes in random noise and produces synthetic samples such as pictures. The discriminator functions as a binary classifier, distinguishing between real and produced samples (from the target distribution). During training, the gen- erator's goal is to create samples that deceive the discriminator into thinking they are real, whereas the

discriminator's goal is to correctly categorise both actual and created data. The two components are trained adversarially, with the generator constantly improving its ability to generate realistic samples and the discriminator constantly improving its discrimina- tion skills.This iterative process drives the GAN to conver- gence, where the generator generates high-quality samples that closely resemble the target distribution and the discriminator is unable to distinguish between real and produced samples. The GAN framework allows for the creation of synthetic data that closely reflects the desired distribution of data.

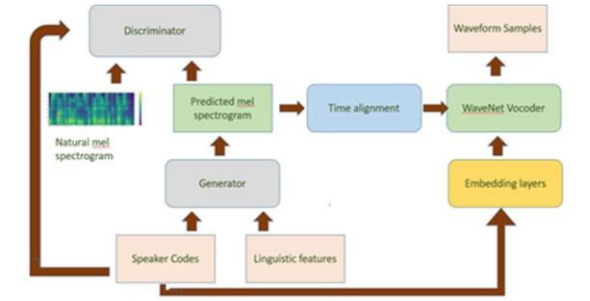

Fig. 5. Basic diagram of GAN Model

D. Deep Bidirectional LSTM (DBLSTM)

Deep bidirectional long short-term memory (DBLSTM) is a particular type of RNN. In a variety of applications, DBLSTM networks outperformed regular RNNs and other deep learning models. They are especially effective at tasks involving sequential data with long-term dependencies, such as speech synthesis, where the model must synthesise a natural- sounding speech waveform based on a sequence of input symbols.

Figure 6 shows basic diagram of DBLSTM in which the input sequence is fed into two levels of LSTM units in a DBLSTM: one layer processes it forward in time, and the other layer processes it backward. Each LSTM unit contains a memory cell that may store data over time and several gates that govern the flow of data. The forward LSTM layer goes through the input sequence from start to finish, cap- turing dependencies in the forward direction. The backward LSTM layer reverse-processes the input sequence, capturing dependencies in the reverse direction. Both layers' outputs are concatenated to produce the final representation, which combines information from both directions. The architecture

of the DBLSTM allows it to capture bidirectional context and better model the relationships between past and present. Deep Bi-LSTM models have been used successfully in a variety of speech synthesis applications such as text-to-speech, multi- speaker synthesis, and controlled synthesis. These models have shown high-quality and real-time synthesis capabilities, making them a promising method for practical voice synthesis applications

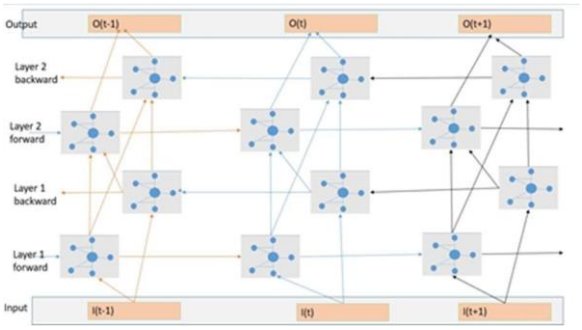

Fig. 6. Basic diagram of DBLSTM Model

V. RESULT AND DISCUSSION

In this section we will discuss the overall outcome of different techniques of Speech Synthesis using Deep learning based on the study conducted till now.

Deep learning-based voice synthesis has made great devel- opment in recent years, demonstrating remarkable potential for producing high-quality, natural-sounding speech. Deep neural networks such as WaveNet, Tacotron, and GANs have been critical in attaining this advancement.

WaveNet, a deep generative model based on autoregressive neural networks, has produced high quality speech waveforms directly from raw audio recordings.

Tacotron, on the other hand, is a sequence-to-sequence model that predicts the corresponding mel-spectrogram from input text and then converts it to audio using a vocoder. Tacotron and its variations have produced outstanding achieve- ments, particularly in terms of the naturalness and expres- siveness of synthesised speech. As per the study conducted Tacotron is best Speech Synthesis technique as it gave the best quality speech.

Another deep learning technology that has showed promise in speech synthesis is generative adversarial networks (GANs), which allow for the development of high-quality and realistic speech utilising a two-step procedure involving the generator and the discriminator.

In the field of speech synthesis, deep bidirectional long short-term memory (DBLSTM) networks have shown consid- erable potential. DBLSTM-based models have achieved state-of-the-art results in several speech synthesis tasks, including text-to-speech, multi-speaker synthesis, and controlled synthe- sis, because to their ability to capture long-term dependencies in sequential data and generate high-quality speech.

While deep learning-based speech synthesis has advanced significantly, there are still some issues to be solved. These include increasing model robustness and stability, lowering computing complexity, and creating more efficient and inter- pretable models. Overall, deep learning-based voice synthesis has immense potential to revolutionise the field of speech tech- nology, with several applications in disciplines such as virtual assistants, speech recognition, and speech rehabilitation.

Comparison table of different type of speech synthesis			
Technique Used	Key charac- teristic	Advantages	Limitations
1 Wavenet	Waveform- based model for directly generating speech at the waveform level. Long-term dependen- cies are modelled using dilated con- volutions.	capable of pro- ducing high- quality sound wave- forms	capable of pro- ducing high- quality sound wave- forms
2 Tacotron	A text- to-speech model that produces speech based on text input. The architecture is sequence- to-sequence, with attention mechanisms and deep Bi-LSTM layers.	End- to-end speech syn- thesis model that can generate high- quality speech wave- forms	Training the model is quite expen- sive.
3. GAN (Gen- erative Adver- sarial Net- works)	A dis- criminator network is used to discern between actual and fake speech samples, and a generator network is used to make speech that fools the discrimi- nator. Un- supervised learning of speech features is possible.	Produce high- quality and natural- sounding speech like real- world speech.	To obtain steady train- ing, it may be neces- sary to tune hyper- param- eters exten- sively.

Comparison table of different type of speech synthesis			
Technique Used	Key charac- teristic	Advantages	Limitations
4 DBLSTM (Deep Bidirec- tional LSTM)	Long-term dependen- cies in sequential data are captured	Can fully benefit from con- textual informa- tion	To syn- thesise wave- forms, a vocoder is still re- quired.

	by a deep bidirectional long short-term memory network. Text-to- speech, multi- speaker synthesis, and controlled synthesis are all possible applications.		

VI. CONCLUSION

Speech synthesis using Deep learning have become one of the active area of research as it provide best quality of speech.Deep learning, which may use massive amounts of training data, has emerged as a significant tool for speech synthesis. Recently, an increasing number of studies using deep learning approaches or even end-to-end frameworks have been done and attained state-of-the-art performance. This study provides an overview of current advancements in speech synthesis, compares the benefits and drawbacks of various methods, and offers potential research avenues that can promote the development of speech synthesis in the future.

REFERENCES

[1] D. H. Klatt, "Review of text-to-speech conversion for English," The Journal of the Acoustical Society of America, vol. 82, no. 3, pp. 737–793, 09 1987. [Online]. Available: https://doi.org/10.1121/1.395275

[2] J. Baart and V. Van Heuven, "From text to speech; the mitalk system: Jonathan allen, m. sharon hunnicutt and dennis klatt (with robert c. armstrong and david pisoni): Cambridge university press, cambridge, 1987. xii+216 pp. £25.00," Lingua, vol. 81, p. 265–270, 07 1990.

[3] "Emotional stress in synthetic speech: Progress and future directions," Speech Communication, vol. 20, no. 1, pp. 85–91, 1996, speech under Stress. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639396000465

[4] M. Chu, H. Peng, Y. Zhao, Z. Niu, and E. Chang, "Microsoft mulan - a bilingual tts system," vol. 1, 05 2003, pp. I–264.

[5] S. N. Kayte, M. Mal, and J. Gujrathi, "Hidden markov model based speech synthesis: A review," International Journal of Computer Appli- cations, vol. 130, pp. 975–8887, 12 2015.

[6] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," CoRR, vol. abs/1709.08041, 2017. [Online]. Available: http://arxiv.org/abs/ 1709.08041

[7] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping,
J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to- speech," 05 2017.

[8] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan,
O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," CoRR, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[9] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," 04 2018, pp. 4804– 4808.

[10] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. C. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings. OpenReview.net, 2017. [Online]. Available: https://openreview.net/ forum?id=B1VWyySKx
[11] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang,
Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous,
Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," CoRR, vol. abs/1712.05884, 2017. [Online]. Available: http://arxiv.org/abs/1712. 05884
[12] S. Kim, S. Lee, J. Song, and S. Yoon, "Flowavenet : A generative flow for raw audio," CoRR, vol. abs/1811.02155, 2018. [Online]. Available: http://arxiv.org/abs/1811.02155
[13] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, FastSpeech: Fast, Robust and Controllable Text to Speech. Red Hook, NY, USA: Curran Associates Inc., 2019.
[14] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. J. Skerry-Ryan,
Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," CoRR, vol. abs/1907.04448, 2019. [Online]. Available: http://arxiv.org/abs/1907.04448
[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley,
S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.
[16] C. Donahue, J. J. McAuley, and M. S. Puckette, "Synthesizing audio with generative adversarial networks," CoRR, vol. abs/1802.04208, 2018. [Online]. Available: http://arxiv.org/abs/1802.04208
[17] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," CoRR, vol. abs/1912.04958, 2019. [Online].
Available: http://arxiv.org/abs/1912. 04958
[18] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2019.
[19] S. O¨ . Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky,
Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep voice: Real-time neural text-to-speech," CoRR, vol. abs/1702.07825, 2017. [Online]. Available: http://arxiv.org/abs/1702.07825
[20] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly,
Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis,
R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to- speech synthesis model," CoRR, vol. abs/1703.10135, 2017. [Online].
Available: http://arxiv.org/abs/1703.10135
[21] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality TTS with transformer," CoRR, vol. abs/1809.08895, 2018. [Online]. Available: http://arxiv.org/abs/1809.08895
[22] Y. Qian, F. Soong, Y. Chen, and M. Chu, "An hmm-based mandarin chinese text-to-speech system," 01 2006, pp. 223–232.

[23] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis,” vol. J83-D-II, 09 1999.