A Comprehensive Review on Speech Synthesis Using Neural-Network Based Approaches

N. N. Perera
Faculty of Information Technology
University of Moratuwa
Moratuwa, Sri Lanka
nimnaperera98@gmail.com

G. U. Ganegoda
Faculty of Information Technology
University of Moratuwa
Moratuwa, Sri Lanka
upekshag@uom.lk

***Abstract*— Speech is a primary mode of communication between human beings. With that, the need of creating artificial speech became a dream of humankind from the beginning of the 1980s. As a result of those experiments, many Speech synthesis systems were developed throughout the past few decades and further many advancements were made to the existing systems as well. However, with the beginning of the Artificial Intelligence era, these advancements were rapidly improved with the aid of neural networks. These newly found advancement techniques and methods were led to reemerge the research area of "Speech synthesis using neural networks-based approaches". This paper reviews currently available speech synthesis techniques, techniques that use neural networks-based approaches, advantages, disadvantages, and limitations. Further this review paper intends to suggest a new hybrid approach and what are modifications can be done in the upcoming future as well.**

Keywords—Speech synthesis, Neural networks

I. INTRODUCTION

Speech Synthesis has many application areas in the present world such as educational applications, telecommunication, and media applications, and in the robot industry as well. But moreover, Speech Synthesis is commonly used in accessibility applications. This can be considered as the most important and most useful application in speech synthesis. From the beginning of the 1980s, the demand for artificial speech became a human fantasy thanks to these various applications.

For the past two centuries, humans have tried to develop many devices and systems that are capable of generating speech and those are used to investigate phonetic phenomena. And the development of these devices and systems changed how humans communicate with each other human-to-human and human-to-machine. The development of speech synthesis systems can be divided into three main eras [1].

1. The mechanical and electro-mechanical era - In this era, mechanical devices were used to generate human speech.
2. The electrical and electronic era - During this time, speech synthesis began to move away from mechanical and electromechanical systems. Instead of mechanical devices, using pure electrical and electronic devices started here.
3. The digital and computational era - With the development of computers and artificial intelligence speech, synthesis has become a very popular topic in the industry. Now it's more accurate when comes to speech synthesis, thus it's not speech synthesis anymore it's a kind of voice cloning, even we can't say the difference between original and synthesized speeches.

With the advent of the digital and computational eras, speech synthesis began to go the extra mile, extending the technology's capabilities. When it comes to modern approaches to voice synthesis, artificial intelligence and machine learning are more prevalent.

More advanced text-extracting techniques, and natural language processing techniques to embed emotions into simple text, emotions plus voice variations in synthesized speeches signify these advancements in modern approaches. This is the place where neural network-based speech synthesis comes into act.

There were many Artificial Intelligence and Machine Learning Scientists who worked hard to research and test out these neural network-based speech synthesis approaches to generate high-quality and accurate speeches since to accept a synthesized speech should lie between acceptable range with regard to quality and accuracy [2]. This paper will discuss those neural network-based approaches in various speech synthesis techniques.

Further, this paper addresses a gap in the literature in the field of speech synthesis, by providing a comprehensive review on speech synthesis techniques using neural network-based approaches. Despite the presence of some review papers addressing speech synthesis techniques using traditional methods[3], a specific literature review that exclusively focused on neural network-based approaches could not be found. Therefore, this paper aims to fill that void by providing a thorough examination of the current state of research in this field.

The remainder of the paper is laid out as follows. The first section discusses the process of speech generation, followed by a discussion of existing speech synthesis techniques in the second half. The third section covers the neural networks-based approaches in those techniques that are discussed in the second section. The fourth and fifth sections hold the other neural network-based approaches in post-filters and a comparison of the current approaches with the suggested improvements will be discussed in the latter part of the paper.

II. SPEECH GENERATION PROCESS

When considering the artificial speech generation process, it is essential to know about the human speech generation process. Because the artificial speech-generation process mimics the speech-generation process of humans[4].

A. Speech Generation Process of Humans

There are three main steps in the human speech generation process as shown in Fig. 1 which are speech

979-8-3503-4737-1/23/$31.00 ©2023 IEEE

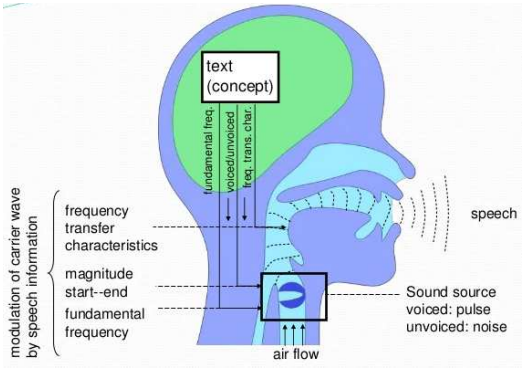

Fig. 1. *The speech production process in humans* [5]

formulation, Human Vocal Generation mechanism using vocal cords and muscles, and acoustic wave generation[4].

The process starts with speech formulation, which is the creation of a sequence of phonemes and prosodic features that represent the desired speech. The speech synthesizer then converts this sequence into a speech signal. The human vocal generation mechanism is the second step, where the vocal cords and muscles work together to modulate the airflow, resulting in the production of sound waves. Finally, the acoustic wave generation step forms the speech signal into sound waves that can be heard by the listener. This step is the result of the interaction between the speech signal and the vocal tract.

B. Speech Generation Process of Speech Synthesis Systems

In artificial speech synthesis systems, it mimics the human speech generation process. It follows the same flow as the human speech generation process. But this flow can be expressed more conveniently in the following manner[6].

1. Input the text.
2. Text analysis (Sentence segmentation, word segmentation, text normalization, part-of-speech tagging, non-standard word tagging, pronunciation analysis, and prosodic analysis will take place in this step. These tasks are carried out by processing the natural language it is a discrete-to-discrete transformation)[7][8].
3. Speech generation (Prosody prediction and waveform generation process take place in this step. And this is a discrete-to-continuous data transformation).
4. Synthesized output.

The text-to-speech transformation will be performed as described above. In this process, the second and third stages are the most critical. A simple change of a configuration in these stages can cause a lot of changes in the synthesized output. And it will lead to a major change in the accuracy of synthesized output as well.

In text analysis, there are three main steps which are text normalization, pronunciation, and prosodic analysis. In the first step – text normalization, breaking the input text into sentences and tokenization will take place. This step is essential to discover the sentences accordingly. The second step is identifying non-standard words. Words such as numbers, years abbreviations, and acronyms are considered non-standard words. Identifying those non-standard words will take place in this step. The next step of text analysis is pronunciation, which is finding the pronunciation for each word. The grapheme-to-phoneme algorithm generates a sequence of characters for this task which can be a rule-based or statistics-driven algorithm. Finally, the prosodic analysis. Prosody refers to the features that make sentences flow naturally. There are three main components of prosody which are phrasing, prominence, and intonation.

The next main stage of speech synthesis systems' speech generation process is "Speech generation". This stage is playing a significant role when considering the output quality and accuracy. Using neural networks-based approaches we can improve the quality of the generated speech output in this stage and this study is focused on these neural networks-based approaches and how to use them to improve the output quality.[8]

III. TRADITIONAL APPROACHES IN SPEECH SYNTHESIS TECHNIQUES

According to studies few main synthesis techniques are currently in use in the modern world.

A. Formant Synthesis

The source-filter method is used to synthesize the speech in format and is likely to produce robotic and unnatural voices[7]. The basic premise is that by replicating formant frequencies and amplitudes, the vocal tract transfer function can be accurately approximated. The speech is synthesized using these estimated frequencies and amplitudes. Formant synthesis does not utilize any human speech samples, instead relying on linguist-written rules to generate the parameters needed for speech synthesis and coarticulation (the transition from one phoneme to the next)[9].

B. Articulatory Synthesis

Articulatory synthesis turns out to be the technique to generate high-quality speech because human articulator behavior is directly modeled here. But this technique is not commonly used as it is difficult to implement[7][9].

C. Concatenative Synthesis

The spoken sentence is broken down into many tiny fragments via concatenative synthesis. sentence into words, words into syllables, them into demi-syllables, phonemes, diaphones, or triphones respectively. To produce new sentences, the above parts of recorded samples are concatenated and rearranged. With this technique, we can achieve maximum intelligibility and naturalness. Concatenative synthesis, on the other hand, has several drawbacks, including discontinuity distortions, a large memory need, and a long processing time[7]. And prosodic modification here will result, artifacts in the speech that makes speech unnatural[9]. Three main types can be found in this technique which are Domain-specific synthesis, Diaphone synthesis, and Unit selection synthesis[6][7].

D. Harmonic plus Noice model

In this model, the speech signal is considered a sum of harmonic and noise components. Prosodic modifications of speech are needed for high-quality speech synthesis. This

model is parametric and so prosodic features can be modified easily with good quality[9].

E. Hidden Markov Model

This is also named statistical parametric synthesis. If the state sequence cannot be determined from the signal sequence, then it is considered hidden. The statistical acoustic model of this technique is trained using context-dependent Hidden Markov Models (HMM)[7].

The training phase and the synthesis phase are the two key phases as shown in Fig. 2. It should be decided during the training phase which features the model should be trained for. There are two steps in the synthesis process. First and foremost, the feature vectors for a given phone sequence must be estimated. Then, to convert those feature vectors into audio signals, a filter is used[9][3].

Further in this paper, the review will be focused on this technique, since most of the neural network-based approaches are focusing on this HMM-based speech synthesis.

IV. Neural Networks-based Approaches in Speech Synthesis Techniques

A. Artificial Neural Network based approach

Artificial Neural Network (ANN) models are recognized for their ability to capture complicated and nonlinear mapping, as well as their ability to generalize. A mapping from the text (linguistic space) to speech is necessary for the context of speech synthesis (acoustic space). Here, the paper presents two methods that are predicting Mel-Cepstral Coefficients and synthesizing speech using the MLSA vocoder (Fig. 3) and using formant characteristics to create a statistical parametric synthesis [10].

In [10] researchers have experimented with the Mel-cepstral-based ANNs synthesis basically in 5 ways which are One network for all the phones, a Separate network for vowels and consonants, a Separate network for each state, One network for all the states, Experiments with Different architectures. To evaluate these experiments, Mel cepstral Distortion (MCD) has been used. The lesser the MCD value better it is.

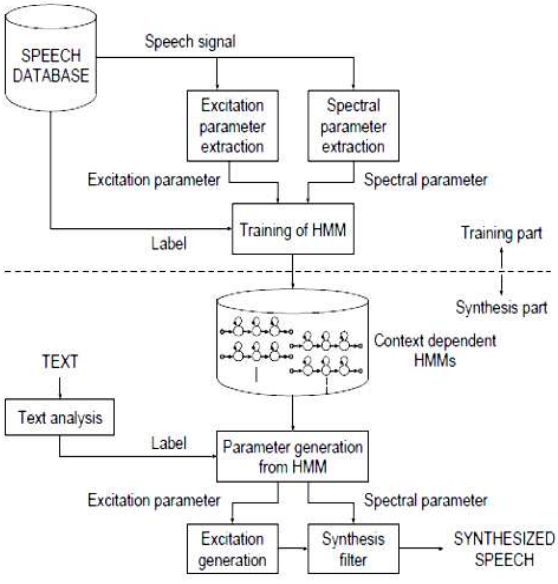

Fig. 2. *Typical architecture of HMM-based speech synthesis system* [3]

When considering the Formant prediction, Researchers in [10] use two methods to compare which are prediction using CLUSTERGEN and prediction using ANN. While both ANN models and CLUSTERGEN can create necessary trajectories in general. But because of ANN's capacity to generalize, ANN models appear to yield smoother trajectories than CLUSTERGEN.

B. Deep Neural Network-based approach

When considering the decision trees and Deep Neural networks (DNN), there are a few factors that we should consider.

- Complex functions of input characteristics, such as XOR, d-bit parity function, or multiplex issues, are difficult to represent using decision trees. Decision trees will be too big to reflect such instances. DNNs, on the other hand, may compactly represent them.
- A partition of the input space is used in decision trees, and each area linked with a terminal node has its own set of parameters. As a result, the amount of data per region is reduced, and generalization is.
- The amount of computing required to train a DNN via back-propagation is often significantly more than that
- required to create decision trees. DNNs need matrix multiplication at each layer during the prediction step, whereas decision trees just require traversing trees from their root to terminal nodes using a subset of input characteristics.

Fig. 4 shows a sample speech synthesis system based on a deep neural network. The given text is analyzed first and then converted into a series of features till the n-th frame.

Then using the deep neural network which includes hidden layers the extracted features will be forward propagated and will give output features. These features include the following.

- Input features – binary answers to questions about linguistic contexts/ Numeric values
- Output features – spectral and excitation parameters/ time derivatives

The Fig. 5 graph signifies the trajectories of 5-mel-ceptral coefficients of natural speech and those predicted by the HMM which the scaling factor(alpha) is equal to one and deep neural network. When comparing two predicted graphs it is visible that DNN based approach is better than the conventional approach. There are some objective and subjective evaluations for these two approaches in [11] as

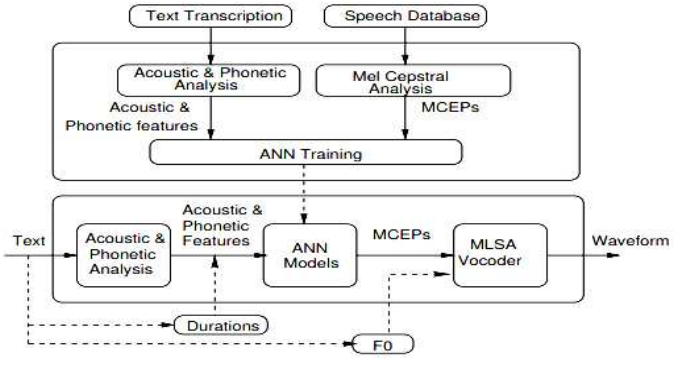

Fig. 3. *Mel cepstral based ANNs synthesis architecture:*[10]

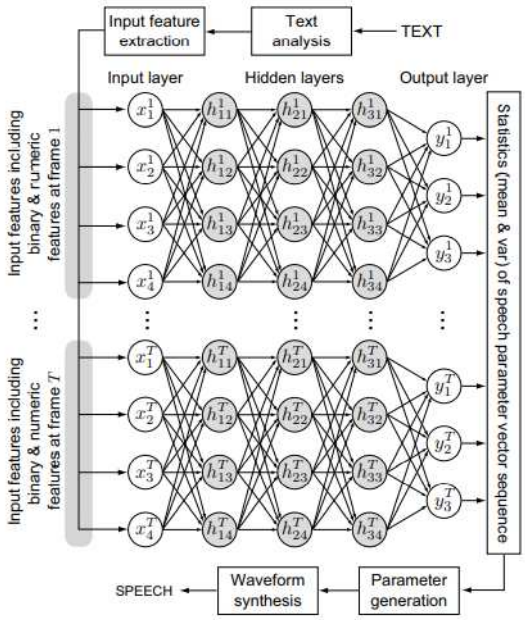

Fig. 4. *A sample speech synthesis system based on a deep neural network* [11]

well, which concludes the DNN-based approach is better than the conventional approach.

C. *Recurrent Neural Network-based approach*

In the recurrent neural network-based approach the most used algorithm is WaveRNN. WaveRNN is a single-layer recurrent neural network for audio synthesis that predicts 16-bit raw audio samples with high accuracy. A gated recurrent unit is the fundamental component of the WaveRNN paradigm. After that, there are two completely linked layers, followed by a softmax activation. Over the corresponding 8 bits, each portion feeds into this softmax layer, and the prediction of the 8 fine bits is conditioned on the 8 coarse bits. Instead of employing a single big output space, the Dual Softmax layer enables efficient prediction of 16-bit samples using two tiny output spaces, 2^8 values each, with 2^16 values. With the use of LPCNet, this output in WaveRNN can be improved further[2].

V. NEURAL NETWORK-BASED APPROACHES IN POST FILTERS

As well as the generation process after generating the speech, post filters are used to improve the quality of the synthesized output. Same as the generation process, these post-filters also follow two approaches which are traditional approaches and neural network-based approaches. Those neural network-based approaches in post-filters will be studied in this section.

A. *Generative Adversarial Network based Postfilter*

The quality of synthetic speech is restricted because of three primary factors: vocoding, acoustic model accuracy, and over-smoothing[12]. This postfilter is mainly focused on the over-smoothing factor. Over-smoothing happens in both the temporal and frequency directions in synthesized speech. Several attempts have been made in the respective directions to address the situation. In the frequency direction, for example, after producing parameters, a postfilter, which is frequent in speech coding, is used to highlight formants. A postfilter is used to improve spectral peaks in the time direction, global variance (GV) and variance scaling (VS) is used to increase the variation of a spectral feature trajectory, and a postfilter is used to enhance the modulation spectrum in the time direction (MS).

Considering this background, in the paper [13] itself, researchers have proposed a learning-based postfilter that learns the acoustic differences directly from the data.

A Generative Adversarial network is a neural network that uses an adversarial process to estimate a generative model and it contains two main components.

- Generative network.
- Discriminator network

To utilize a GAN for post-filtering they have made three changes to naïve GAN architectures.

- Conditional generative adversarial network – this model enables the generative network to generate a spectral texture that is realistically conditioned on the given synthesized speech.
- Residual representation – In here instead of the raw spectral texture, they have designed the postfilter to use the residual spectral texture generator. This helps the generator to get familiar with the subtle variations between natural and synthetic spectral textures.
- Convolutional architecture – To allow contraction in the spectral structure and temporal, the spectral structure must be flexible.

B. *Deep Neural Network based Stochastic Postfilter*

Statistical parameter-based speech synthesis is currently considered the most popular speech synthesis method because of its special features. But even if it is popular, it still tends to generate muffled outputs. Therefore, to improve this output quality many methods were introduced, including post-filters. In the recent past, a fascinating method was

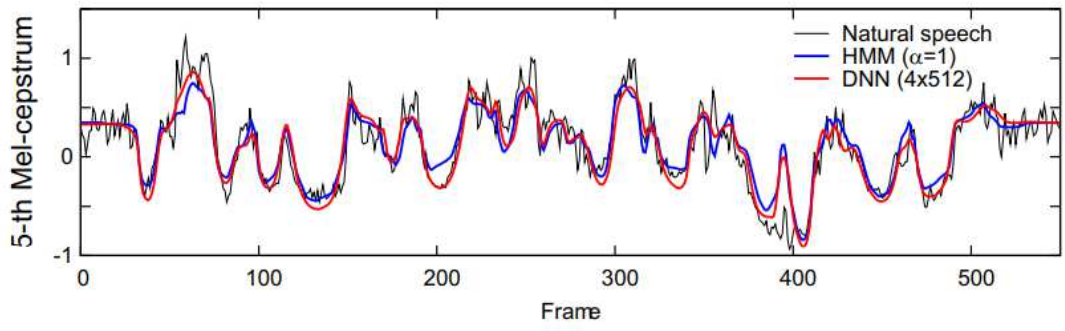

Fig. 5. *Trajectories of 5-th Mel-cepstral coefficients of a given phrase's speech predicted by the conventional HMM, DNN-based HMM systems, and natural speech of that given phrase* [11]

introduced to improve the quality of synthesized speech segments, which is based on modulation spectrum enhancement[14]. Using the same method this Postfilter is generated to enhance spectral parameter trajectories' natural frequency modulation. In [15] to model the conditional probability of the acoustic differences, researchers have introduced a deep neural network[16]. For this task, a Gaussian mixture model is commonly used in voice conversion[17]. But in this approach, a deep network is used instead of a gaussian mixture model because of its capabilities, such as allowing to conduct spectral shaping directly in the spectral domain and modeling highly dimensional and highly correlated data. This approach is very similar to the approach used in [16] which a Deep Neural Network is used for stochastic modeling of the difference between the spectra of natural and synthesized speech.

Fig. 6 shows the structure of a feedforward Deep Neural Network with four layers. x is the input synthesized spectral envelope. y signifies the corresponding natural spectral envelope. As shown in Fig. 6, the selected architecture is layer-by-layer trained using a cascade of a Bernoulli bidirectional associative memory (BBAM) and two restricted Boltzmann machines (RBMs)[18]. Further, as the Deep Neural Network's input and output, they have employed three consecutive frames of spectral envelopes. Therefore, to produce increased spectral envelopes, the parameter generation procedure of the HMM-based parametric speech synthesis approach is used.

VI. Discussion

The paper selection process for this review article involved conducting a comprehensive literature search using relevant keywords such as "speech synthesis" and "neural network". Papers were screened for inclusion based on their relevance to the topic and were limited to those published after 2010. The remaining papers were evaluated for both relevance and quality, with only those deemed high in both areas being included in the review. This process ensured that the review is up-to-date and presents a thorough analysis of the current state of research in the field of speech synthesis using neural-network-based approaches.

According to the studies it is visible that among other techniques unit selection technique and statistical parametric synthesis generate quality and accurate speech. But based on that it's not fair to put away all the synthesis techniques at once. Hence it is indeed required to critically evaluate the pros and cons of every speech synthesis technique.

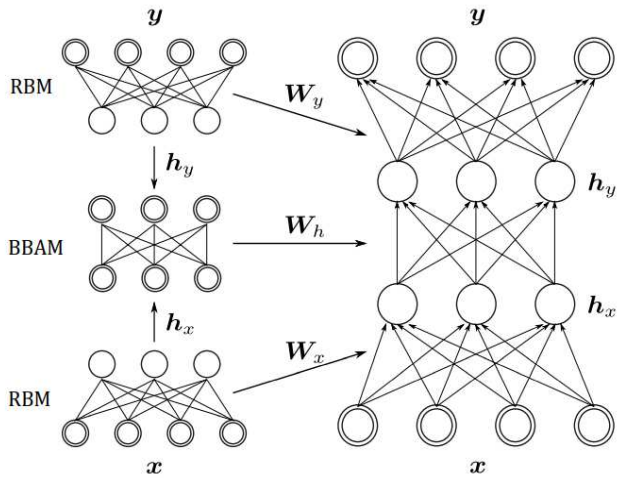

Fig. 6. *Structure and the training procedure of the proposed approach*[15]

Formant synthesis is an older speech synthesis technique that generates an infinite number of sounds and uses 3-5 formants to generate intelligible speech. It is rule-based, making it more intelligible than concatenative techniques, even at high speeds. Formant synthesis programs are also smaller in size, making them suitable for embedded systems with limited memory and processing power.

Articulatory synthesis mimics the human speech generation process, resulting in highly intelligible speech but it is not used in commercial applications. Instead, it is mostly used for studying purposes such as modeling speech synthesis and speech recognition. Concatenative synthesis is a corpus-based technique that uses pre-recorded samples, which can cause discontinuities and can require larger databases. However, it is commonly used in systems where discontinuity does not matter or which have limited vocabulary such as speaking clocks, train stations, etc.

Finally, it comes to statistical parametric synthesis which uses neural network-based approaches to improve the quality and accuracy. Because of parametric usage and neural network-based approaches, the major benefit of statistical parametric synthesis is that it can synthesize speech with a variety of voice characteristics, such as speaker individuality, speaking styles, and emotions, among others. According to [19] there are some other advantages also, which are,

- Wide coverage in acoustic space.
- Multilingual support
- Small footprint
- Robustness
- Unifying frontend (text analysis) and backend (waveform generation)
- Separate control in the spectrum, excitation, duration, etc.

When considering the statistical parametric speech synthesis technique over the unit selection concatenative method, which is considered the best concatenative technique, the statistical parametric method has significant advantages[3].

But on the other hand, it has a few drawbacks over unit selection. The main disadvantage is statistical parametric synthesis generates a buzzy voice which makes speech unnatural. According to [19], the suggested method is to use a hybrid approach that combines unit selection and Statistical parametric synthesis along with the neural network approach. Here it states target prediction of unit selection can be done using the statistical parameters and then by mixing natural units in the unit selection and generated speech in Hidden Markov Model the combined model can generate quality synthesized speech output.

When considering limitations on speech synthesis systems, some major and minor limitations can be seen. In speech synthesis systems the major limitations are in the quality of the synthesized output[20]. In [19], it highlights three main factors that affect these synthesized speeches' output quality. Those are over-smoothing, the accuracy of acoustic models, and vocoding. Nevertheless, rather than the

quality, there are a few limitations[1] such as emotions[21][22], prosody, preprocessing (text analysis), ambiguities, and naturalness which can be overcome using neural network-based approaches.

Despite the fact that this review addresses the gap in the neural network-based approaches for speech synthesis, it does not include details about the newest addition to neural networks, transfer learning. Transfer learning has become an increasingly popular approach in various fields and its potential for speech synthesis is yet to be fully explored. Therefore, it can be focused on conducting a comprehensive review and expanding methodologies of transfer learning-based approaches in the field of speech synthesis. This would provide a deeper understanding of the potential and limitations of transfer learning in speech synthesis and would help to guide future research in this area.

VII. Conclusion

Synthesizing speech artificially, for better communication between humans and machines, became an essential thing in the recent past. With that, humans started researching speech synthesizing for decades. Then the digital and computational era began; Machine learning and Artificial Intelligence came into the act. With the beginning of the Artificial Intelligence era, the advancements in speech synthesis began to improve rapidly while opening new pathways. These new pathways enabled researchers to explore the neural networks-based approaches in speech synthesis to get better output concerning the accuracy and quality of synthesized speech.

Acknowledgment

This review paper is supported by the Faculty of Information Technology, University of Moratuwa under the supervision of the Department of Information Technology and the Department of Interdisciplinary Studies.

References

[1] B. H. Story, "History of speech synthesis," *Routledge Handb. Phonetics*, pp. 9–33, 2019, doi: 10.4324/9780429056253-2.

[2] J. M. Valin and J. Skoglund, "LPCNET: Improving Neural Speech Synthesis through Linear Prediction," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May, pp. 5891–5895, 2019, doi: 10.1109/ICASSP.2019.8682804.

[3] S. Kayte, M. Mundada, and J. Gujrathi, "Hidden Markov Model based Speech Synthesis: A Review," *Int. J. Comput. Appl.*, vol. 130, no. 3, pp. 35–39, 2015, doi: 10.5120/ijca2015906965.

[4] J. V Vyas, "Study of Speech Recognition Technology and its Significance in Human-Machine Interface," vol. 3, no. 10, pp. 416–422, 2017.

[5] H. Zen, "Acoustic modeling in statistical parametric speech synthesis-from HMM to LSTM-RNN," 2015.

[6] K. Kuligowska, P. Kisielewicz, and A. Włodarz, "Speech synthesis systems: Disadvantages and limitations," *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 234–239, 2018, doi: 10.14419/ijet.v7i2.28.12933.

[7] S. Lukose and S. S. Upadhya, "Text to speech synthesizer-formant synthesis," *2017 Int. Conf. Nascent Technol. Eng. ICNTE 2017 - Proc.*, pp. 1–4, 2017, doi: 10.1109/ICNTE.2017.7947945.

[8] M. Z. Rashad, H. M. El-Bakry, I. R. Isma'il, and N. Mastorakis, "An overview of text-to-speech synthesis techniques," *Int. Conf. Commun. Inf. Technol. - Proc.*, pp. 84–89, 2010.

[9] Y. Tabet and M. Boughazi, "Speech synthesis techniques. A survey," *7th Int. Work. Syst. Signal Process. their Appl. WoSSPA 2011*, pp. 67–70, 2011, doi: 10.1109/WOSSPA.2011.5931414.

[10] E. V. Raghavendray, P. Vijayadityay, and K. Prahalladyz, "Speech synthesis using artificial neural networks," *Proc. 16th Natl. Conf. Commun. NCC 2010*, pp. 5–9, 2010, doi: 10.1109/NCC.2010.5430190.

[11] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 7962–7966, 2013, doi: 10.1109/ICASSP.2013.6639215.

[12] Y. Cui, X. Wang, L. He, and F. K. Soong, "A New Glottal Neural Vocoder for Speech Synthesis.," in *Interspeech*, 2018, pp. 2017–2021.

[13] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4910–4914.

[14] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in HMM-based speech synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 290–294.

[15] L. H. Chen, T. Raitio, C. Valentini-Botinhao, J. Yamagishi, and Z. H. Ling, "DNN-based stochastic postfilter for HMM-based speech synthesis," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. September, pp. 1954–1958, 2014, doi: 10.21437/interspeech.2014-441.

[16] L. H. Chen, Z. H. Ling, and L. R. Dai, "Voice conversion using deep neural networks with multiple frame spectral envelopes," *Submitt. to Interspeech*, 2014.

[17] D. Saito, H. Doi, N. Minematsu, and K. Hirose, "Application of matrix variate Gaussian mixture model to statistical voice conversion," 2014.

[18] L.-J. Liu, L.-H. Chen, Z.-H. Ling, and L.-R. Dai, "Using bidirectional associative memories for joint spectral envelope modeling in voice conversion," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7884–7888.

[19] H. Zen, K. Tokuda, and A. W. Black, "Statistical Parametric Speech Synthesis," vol. 004, 2009, doi: 10.1016/j.specom.2009.04.004).

[20] H. Zen, M. J. F. Gales, and Y. Nankaku, "Product of Experts for Statistical Parametric Speech Synthesis," vol. 20, no. 3, pp. 794–805, 2012.

[21] S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs," *Proc. - 9th Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC 2017*, vol. 2018-Febru, no. December, pp. 1613–1616, 2018, doi: 10.1109/APSIPA.2017.8282282.

[22] J. Lorenzo-Trueba, G. Eje Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis," *Speech Commun.*, vol. 99, pp. 135–143, 2018, doi: 10.1016/j.specom.2018.03.002.