ICARC 2023

Sabaragamuwa University of Sri Lanka

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

N.N. Perera[1], G. U. Ganegoda[2]

[1] Faculty of Information Technology, University of Moratuwa,
Moratuwa, Sri Lanka

[2] Faculty of Information Technology, University of Moratuwa,
Moratuwa, Sri Lanka

Presenting Author: N.N. Perera

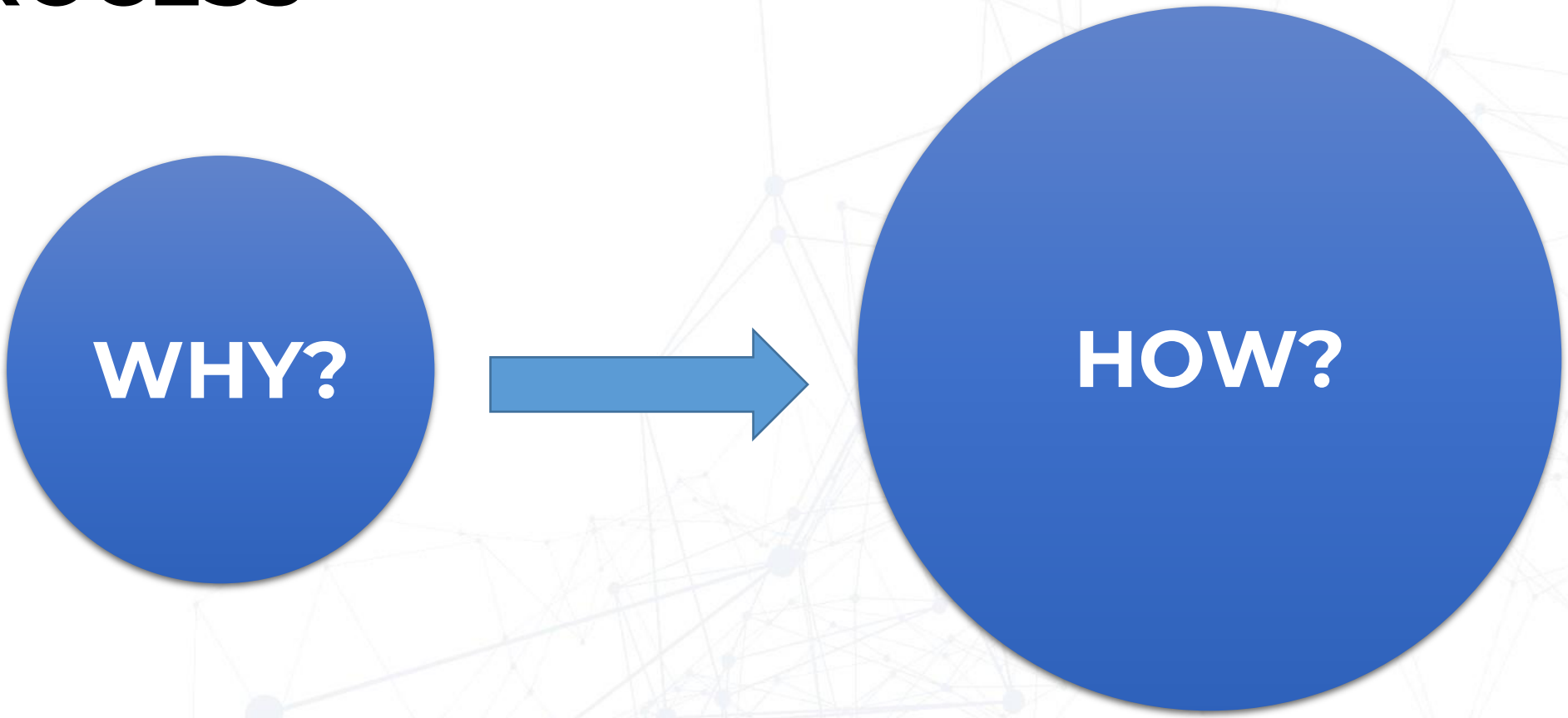
# OVERVIEW

- Introduction
- Literature Gap and Paper Selection Process
- Speech Generation Process
- Speech Generation Techniques
- Neural Network-Based Approaches in Speech Synthesis Techniques
- Neural Network-Based Approaches in Postfilters
- Neural Network-Based Approaches in Vocoding
- Proposed Approach
- Limitations

# 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.
- For the past two centuries, humans have tried to develop many devices and systems that are capable of generating speech
- With the advent of the digital and computational eras, speech synthesis began to go the extra mile, extending the technology's capabilities.
- In the last two decades Artificial Intelligence Scientists were tend to research more about Neural networks-based approaches in speech synthesis techniques to embed emotions and voice variations in synthesized speeches.

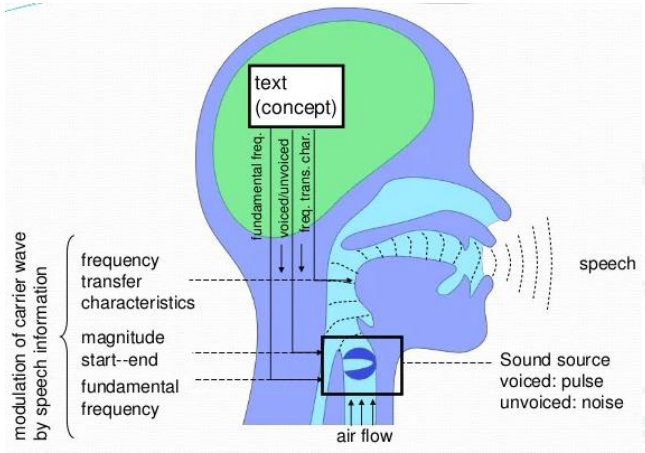
# LITERATURE GAP AND PAPER SELECTION PROCESS



# SPEECH GENERATION PROCESS

## Human Speech Generation Process

- Speech formulation
- Human Vocal Generation mechanism using vocal cords and muscles
- Acoustic wave generation




## Artificial Speech Generation Process

- Input the text.
- Text analysis (Sentence segmentation, word segmentation, text-normalization, part-of-speech tagging, non-standard words tagging, pronunciation analysis, and prosodic analysis will take place in this step.)
- Speech generation (Prosody prediction and waveform generation process take place in this step.)
- Synthesized output

# SPEECH SYNTHESIS TECHNIQUES

## Formant Synthesis

- The source-filter method is used.
- Likely to produce robotic and unnatural voices.
- Relying on linguistic rules to generate parameters needed for speech synthesis and coarticulation.

## Articulatory Synthesis

- human articulator behaviour is directly modeled here.
- This technique is not commonly used as it is difficult to implement.

# SPEECH SYNTHESIS TECHNIQUES (Contd...)

## Concatenative Synthesis

- The spoken sentence is broken down into sentences, words, syllables, demi-syllables etc.
- To produce new sentences, the above parts of recorded samples are concatenated and rearranged.
  - **Domain-specific synthesis**
    Used in domain-specific applications like speaking clocks, speaking calculators, medical call centers, weather reports, and train announcements. These systems are highly restricted by vocabulary.
  - **Diaphone synthesis**
    It is used in most software that allows free speech synthesis. It produces a robotic voice rather than a genuine-sounding voice using recorded diaphones. Tends to create glitches.
  - **Unit Selection synthesis**
    A large database of recorded speeches is used. It provides the maximum naturalness (very similar to human actual voice). This technique keeps whole words and phrases in the database. Therefore, it is much memory-consuming and needs a large storage for the database.

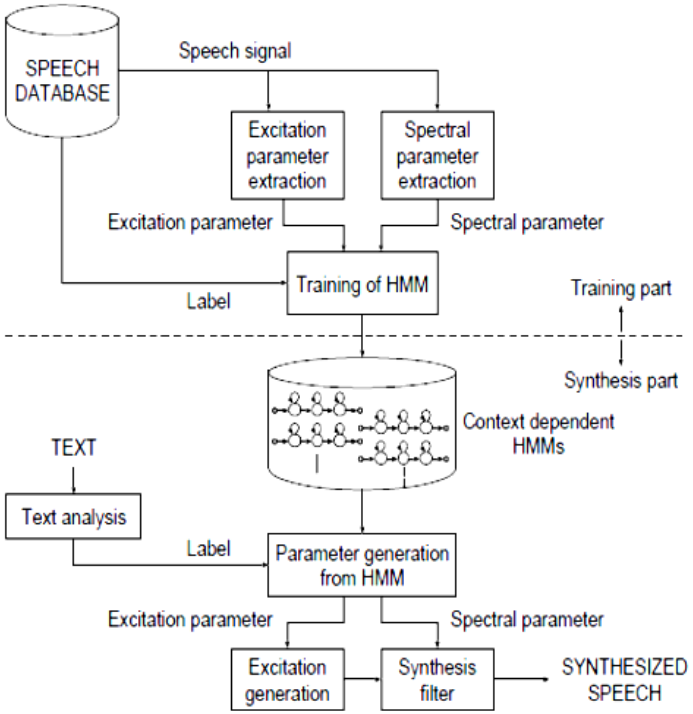
# SPEECH SYNTHESIS TECHNIQUES (Contd...)

## Harmonic plus Noise Model (HNM)

- The speech signal is considered as a sum of harmonic and noise components.

## Hidden Markov Model (HMM)

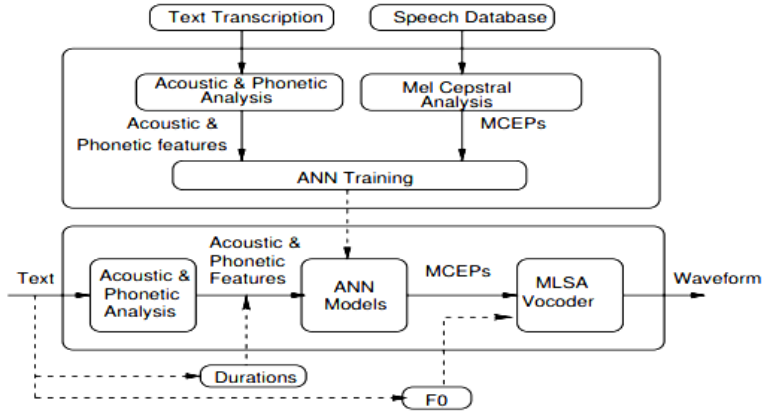
- Also named as statistical parametric synthesis. If the state sequence cannot be determined from the signal sequence then it is considered hidden. The statistical acoustic model of this technique is trained using context-dependent Hidden Markov Models.
- The training phase and the synthesis phase.
- It should be decided during the training phase which features the model should be trained for.
- Two steps in the synthesis process.
- The feature vectors for a given phone sequence must be estimated.
- To convert those feature vectors into audio signals, a filter is used.

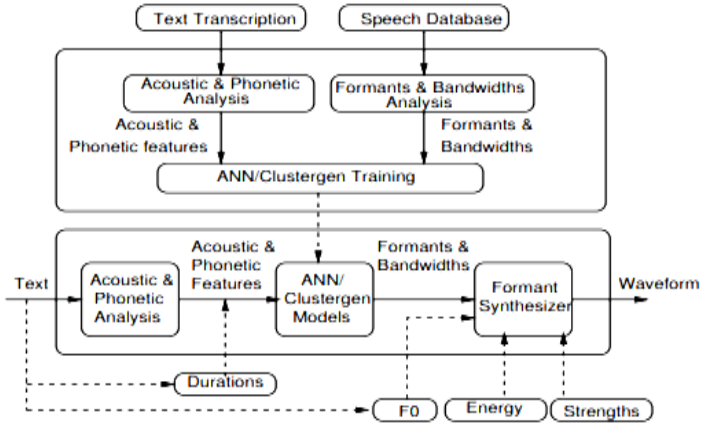


# NEURAL NETWORK BASED APPROACHES IN SPEECH SYNTHESIS TECHNIQUES

## 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).
- Presented two methods
  - predicting Mel-Cepstral Coefficients and synthesizing speech using the MLSA vocoder.
  - Using formant characteristics to create a statistical parametric synthesis.

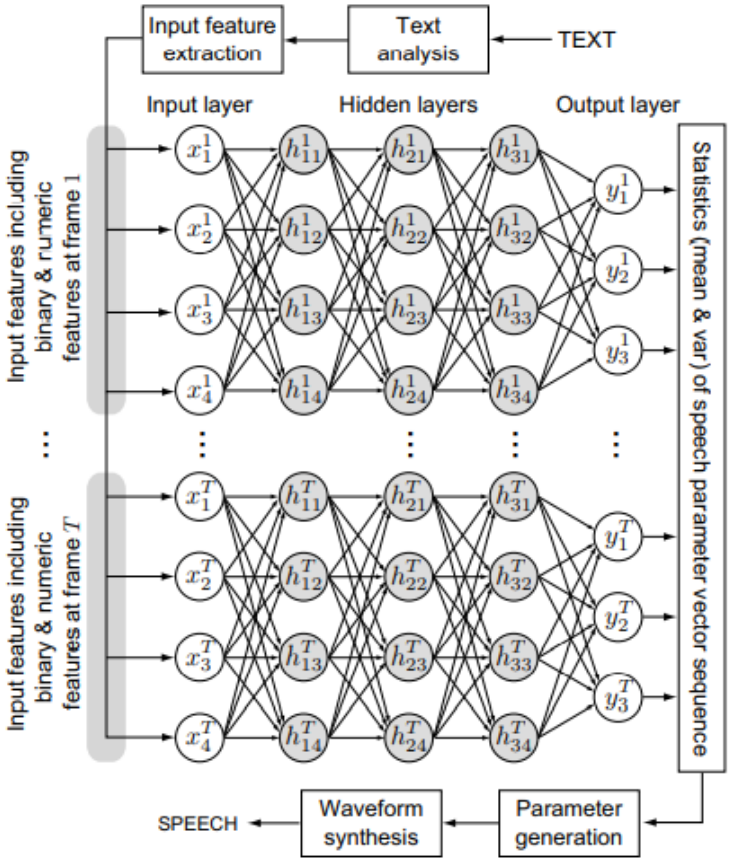


# NEURAL NETWORK BASED APPROACHES IN SPEECH SYNTHESIS TECHNIQUES (Contd...)

## Deep Neural Network based Approach

- 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.



# NEURAL NETWORK BASED APPROACHES IN SPEECH SYNTHESIS TECHNIQUES (Contd...)

## Recurrent Neural Network based Approach

- 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.
- With the use of LPCNet, this output in WaveRNN can be improved further

# NEURAL NETWORK BASED APPROACHES IN POST-FILTERS

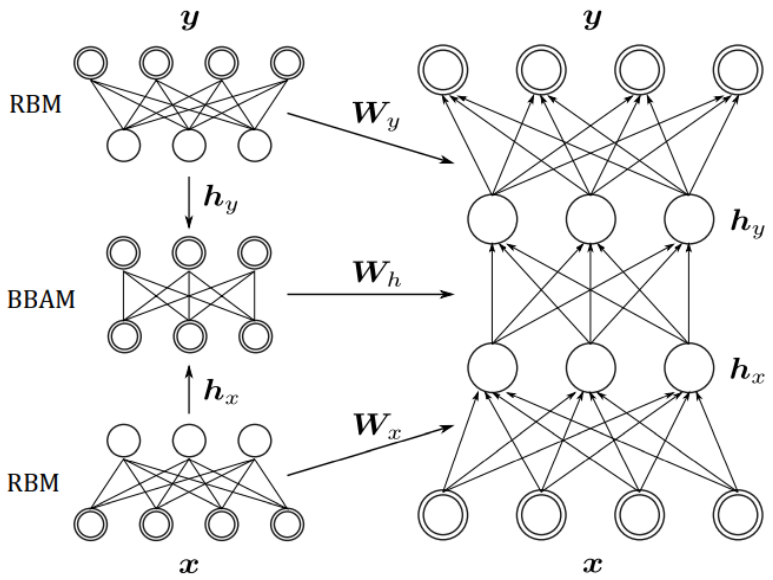
## Generative Adversarial Network based Postfilter

- The quality of synthetic speech is restricted because of three primary factors: vocoding, acoustic model accuracy, and over-smoothing. In this postfilter is mainly focused on the over-smoothing factor.
- A Generative Adversarial network is a neural network that uses an adversarial process to estimate a generative model and it contains with two main components.
  - Generative network
  - Discriminator network
- Researchers have proposed a learning-based postfilter that learns the acoustic differences directly from the data with three changes to naïve GAN architecture.
  - **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** – 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.

# NEURAL NETWORK BASED APPROACHES IN POST FILTERS (Contd...)

## Deep Neural Network based Stochastic Postfilter

- Feedforward Deep Neural Network with four layers.
- x is the input synthesized spectral envelop. y signifies the corresponding natural spectral envelop.
- The selected architecture is layer-by-layer trained using a cascade of a Bernoulli bidirectional associative memory (BBAM) and two restricted Boltzmann machines (RBMs
- 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.



# LIMITATIONS

## Emotions

The impact of emotions in human verbal communication is limited in Speech synthesis systems.

## Prosody

Prosody is extremely important in conveying a whole communication experience between the speaker and the listener. The prosody feature allows the synthesizer to change the pitch of the voice to produce output that sounds like it was spoken by individuals in conversation. Quality of prosody is one of the main problems which face modern speech synthesis engines.

## Preprocessing – Text analysis

Depending on the language pre-processing text analysis can be complex because of the high number of combinations in vowels and consonants in the relevant language

## Ambiguities

The problem of ambiguity exists in many forms such as ambiguity of homographs, Syntactic ambiguity. These ambiguities wreak havoc on the ability to create high-quality speeches.

## Naturalness

The naturalness of the generated speech. Still most speech synthesis systems are generating robotic voices which make synthesized speech outputs unnatural.

# CONCLUSION

- Synthesizing speech artificially, for better communication between humans and machines, became an essential thing in the recent past.
- 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 the better output concerning accuracy and quality of synthesized speech.

# THANK YOU