Journal of King Saud University - Computer and Information Sciences 36 (2024) 102131

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University - Computer and Information Sciences

journal homepage: www.sciencedirect.com

Review Article

Planning the development of text-to-speech synthesis models and datasets with dynamic deep learning

Hawraz A. Ahmad [a], Tarik A. Rashid [b,*]

[a] Software and Informatics Department, College of Engineering, Salahaddin University-Erbil, Erbil, Iraq
[b] Computer Science and Engineering Department, University of Kurdistan Hewlêr, Erbil, Iraq

ARTICLE INFO

Keywords:
Deep learning
Text-to-speech (TTS)
Parametric synthesis
Concatenative synthesis
Text analysis

ABSTRACT

Synthesis of Text-to-speech (TTS) is a process that involves translating a natural language text into a speech. Speech synthesisers face a major challenge when recognizing the prosodic elements of written text, such as intonation (the rise and fall of the voice in speaking), and length. In contrast, continuous speech features are influenced by the personality and emotions of the artist. A database is maintained to store the synthesized speech pieces. Its output is determined by how similar the person utters the words and how capable they are of being implied. In the past few years, the field of text-to-speech synthesis has been heavily impacted by the emergence of deep learning, an AI technology that has gained widespread popularity. This review paper presents a taxonomy of models and architectures that are based on deep learning and discusses the various datasets that are utilised in the TTS process. It also covers the evaluation matrices that are commonly used. The paper ends with a look at the future directions of the system and reaches to some Deep learning models that give promising results in this field.

1. Introduction

A speech synthesizer is an electronic system used to process information. As its applications expand, the quality of this type of software rises. For instance, speech synthesis can be utilized to assist people with visual impairments in communicating effectively (Sak et al., 2006). Speech synthesis can also be utilized to teach various languages proper pronunciation and spelling. Nowadays, smartphones have the capability of listening to users' questions and providing answers using a wide range of intelligent assistants, such as Google Assistant, Siri and Microsoft's Cortana (Lopez et al., 2018). Artificial Intelligence has been heavily studied in the field of speech synthesis. The TTS system's goal is to automatically convert text into speech. It does so through two main steps. The first involves text analysis, which converts the input string into a phonetic or symbolic representation.

The second step involves creating speech waveforms. There are two main techniques used in speech synthesis: traditional and deep learning. In the former, machine learning is used for developing TTS systems, while deep learning is focused on developing systems that can perform more complex tasks (Khan and Chitode, 2016). Concatenative speech synthesis is a type of research that uses the concatenation and selection of segments of the human voice. It is commonly performed on a single voice, and it can be distinguished from other methods by the way it selects and stores the correct syllables and tone (Ning et al., 2019; Zena et al., 2009). various schemes are designed to perform this type of synthesis, such as the Epoch Synchronous (Kayte et al., 2015) Add and Spacing (ESNOLA) (Mandal and Datta, 2007), the Pitch Synchronous Add and Spacing (PSOLA) (Norbert et al., 2000), the Time Domain Time Synchronous Add and Spacing (TDPSOLA) (Toma et al., 2010; Matthe-yses et al., 2006), and the EMBROLA (Gopi et al., 2013).

This style of synthesis is often denoted as a concatenative synthesis. The main differences between this and other methods are the way they store the signals in their database and the restoration procedure. There are two main types of models used for developing TTS systems: the autoregressive and non-autoregressive models (Khanam et al., 2022). We will discuss over the past decade, several papers on speech synthesis using Deep Machine Learning have been published. Although they presented good literature reviews, they mainly focused on techniques for specific languages. Also, they didn't discuss the technologies or methods involved in the paper. The contribution of this study that makes it different from other TTS reviews is that we have done this review on Kurdish TTS. A thorough evaluation of the Kurdish Text TTS may help

* Corresponding author.
E-mail addresses: hawraz.dizayi@su.edu.krd (H.A. Ahmad), tarik.ahmed@ukh.edu.krd (T.A. Rashid).

https://doi.org/10.1016/j.jksuci.2024.102131
Received 11 May 2024; Received in revised form 17 June 2024; Accepted 10 July 2024
Available online 26 July 2024

advance the technology. This paper includes some contributions summarized as follows:

1) Structure of deep learning models in a taxonomy.
2) Analyse TTS and datasets and discover widely utilized assessment matrices.
3) It presents a historical context and provides future directions for researchers in Text-to-Speech synthesis.
4) Developers and researchers can influence the reviews to gain a deeper understanding of the Kurdish TTS's shortcomings and strengths. This feedback can aid in identifying areas of improvement, like pronunciation and naturalness.
5) It can provide developers and researchers with valuable insight into how to make their creations more user-friendly.
6) The reviews can also help developers and researchers identify the specific use cases that are most commonly used by the Kurdish TTS.
7) A quality assurance measure, such as a Kurdish TTS review, can assist a developer in identifying issues that need attention. Reviews are helpful in the continuous refinement and development of the Kurdish TTS. They can provide valuable insight that can help improve the system and make it more useful to its users.

The paper is moulded into sections. The second section covers an overview of the survey methodology. Section three talks about the Kurdish Language. Section four has descriptions of the TTS Challenges. Finally, the rest Sections highlight the basic speech synthesis and evaluation metrics that are commonly used.

2. TTS literature review

Text is a common form of computer interaction that most people expect to feel as natural as talking to other humans. With the help of speech synthesis, a computer can convert written texts into voice messages. The process of speech synthesis concerns the creation of speech using various other forms of media, such as text, gestures, and lip movements. Text-to-speech technology targets to help convert any language into speech.

There are many applications of text-to-speech technologies, particularly in the area of voice assistance.

1. Voice Assistants: Text-to-speech is a core component of several voice assistants, such as Apple Siri, Google Assistant, and Amazon Alexa. These assistants utilize TTS to accurately convert text-based instructions into speech, allowing users to converse with them using spoken language.
2. Accessibility: Text-to-speech technology is a vital component of making information more accessible to people with visual impairments. It can be used by screen readers to read aloud the content of documents, websites, and applications.
3. Navigation Systems: Text-to-speech is commonly utilized in GPS navigation devices. It converts text-based prompts into spoken instructions, which enables pedes trains and drivers to safely navigate roads.
4. E-learning and Education: Text-to-speech technology is commonly utilized in educational applications and e-learning platforms to provide audio-supported content.
   It helps individuals with different learning styles and abilities access books and articles.
5. Call Centers and Customer Service: In call centre applications, TTS can automate the process of answering frequently asked questions by providing answers without requiring human operators to answer them. It handles large numbers of calls and delivers consistent information to consumers.
6. Multilingual Applications: Text-to-speech technology can be utilized in various applications, such as language learning tools and translation services.
7. Audio Books and Podcasts: Text-to-speech technologies can be used to convert various types of written content, such as articles and books, into audio files that can be played on a device. They make it easier for people who prefer auditory entertainment and learning.
8. Assistive Devices for the Elderly: Text-to-speech technology can also be integrated into various smart home systems, which are designed to help the elderly use their devices more easily. It allows them to control their devices and receive information.

Various methods have been used for creating speech, such as parametric and concatenation synthesis. A concatenation synthesis is a process that involves the creation of pre-recorded sound segments, such as syllables, sentences, and individual phones. These segments can be stored in various forms, such as spectrograms and waveforms.

2.1. Statistical parametric synthesis

A parametric synthesis is a process that uses a set of parameters and a function to modify a human-recorded voice. Statistical parametric synthesis is a process that involves training and then synthesis. During the training phase, we analyze the various parameters of an audio sample to get a sense of its composition. We then use a statistical model to estimate these parameters. Hidden Markov Models (HMM)s are known to provide the most accurate results. They are used in the synthesis process to create the final speech output. Fig. 1. Shows tts using statistical parametric synthesis.

The advantages of using a statistical model for synthesis are its flexibility and the lack of need for storing the audio sample in a database. In most cases, the output of a synthesized speech is not good enough. This is the reason why Deep Learning techniques are used. In deep learning, the Target speech X can be created by taking into account the sequence Y of the text.

$$X = \text{argmax}P(X|Y, Parametersforthemodel) \tag{1}$$

Usually, the text is passed to a feature generator, which creates various acoustic features, such as the spectrogram or fundamental frequency. A Neural vocoder then converts the output into a speech segment.

2.2. WaveNet Model

In (Oord et al., 2016), a study proposed the WaveNet framework, which is a form of neural network that can create raw audio waveforms. It is a deep-generative model of raw audio waveforms. The WaveNet model was the first to create speech-like audio waveforms modelled after the raw audio signal. It can output 16,000 samples per second and is an autoregressive algorithm, which indicates that each sample is dependent on its previous iteration

$$P_\theta(x) = \prod_{t=1}^{T} P\left(x|x_1, \cdots, x_{t-1}\right) \tag{2}$$

The model was completely autoregressive and probabilistic. The researchers utilized a pair of English and Mandarin-Chinese datasets with varying levels of speech data. They evaluated the performance of WaveNets through a series of subjective paired comparisons. They also used the Tacotron framework, which was proposed by (Wang et al., 2017). The author utilized a fully convolutional neural network and dilated convolutions to create autoregressive models. The WaveNet framework was inspired by two of the most popular image generators, namely Pixel Recurrent Neural Networks (PixelRNN) and Pixel Convolutional Neural Networks (PixelCNN) as in Fig. 2.

The authors trained the models by using real waveforms created by human speakers. After the training, the network produces the final output. One of the main disadvantages of WaveNet is that it requires a lot of computation to perform properly. This can make inferences very

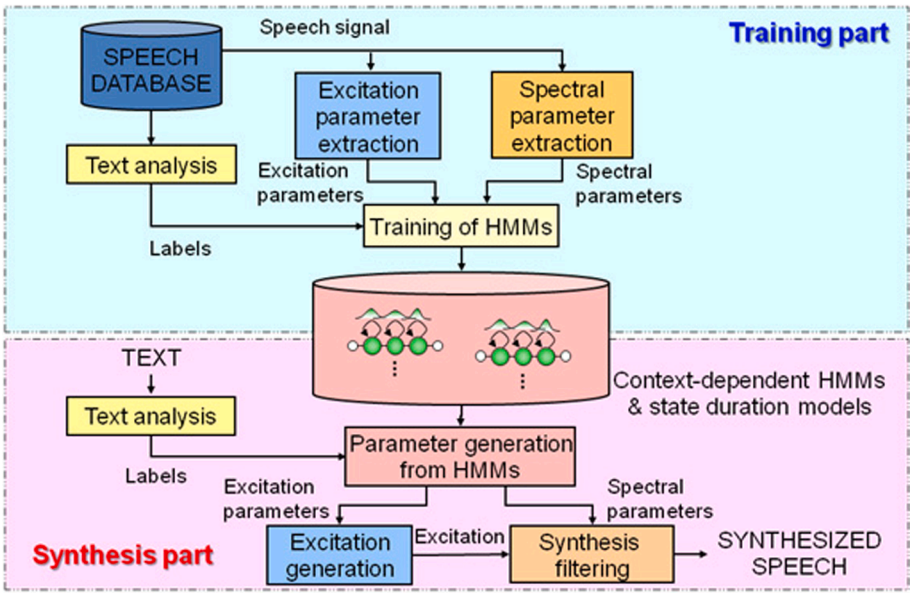

Fig. 1. Statistical Parametric Synthesis (Zena et al., 2009).

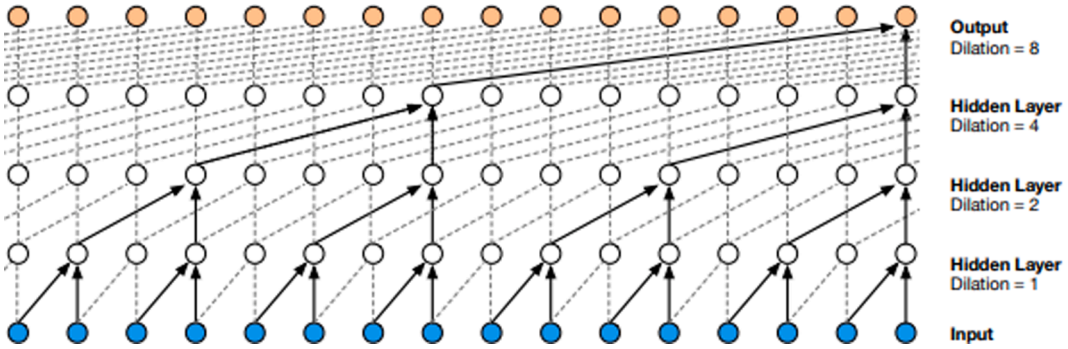

Fig. 2. A stack of causal convolution layers Visualization (Oord et al., 2016).

slow and costly. WaveNet's first version, which is in English, has a Mean Opinion Score (MOS) of 4.21. This is significantly higher than the previous models, which were typically around 3.67–3.86. The Fast WaveNet platform was able to reduce the complexity of WaveNet's original structure (See Fig. 3). The goal of the system was to reduce the number of layers to L, which is the number of connections in the network. By implementing a caching system, no redundant calculations were made.

2.3. Deep voice

The Deep Voice framework was developed by Baidu to help develop the next generation of speech synthesis (Arik et al., 2017) as shown in Fig. 4. It features four neural networks that work together to produce a pipeline. The framework features a segmentation model that can identify the boundaries between various phonemes. It uses a combination of RNN and CNN networks to predict the alignment of the target phonemes with the sounds. The framework is also used to convert graphemes into phonemes. It utilizes a multi-layer encoding and decoding model with a GRU cell. It can predict the duration and fundamental frequencies of phonemes. It learned these tasks by training two sets of interconnected layers with unidirectional GRU cells. The WaveNet framework was used by the authors to create a final audio file. It features a modified conditioning network that samples linguistic features at the desired frequency.

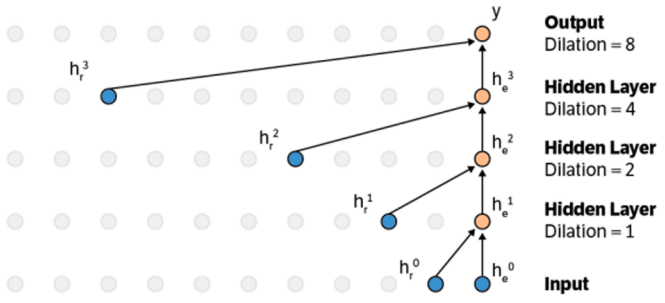

Fig. 3. Fast Wavenet where the computational complexity of a given output is expressed as O(L) and the number of layers is L (Le Paine et al., 2016).

In (Arik et al., 2017), Deep Voice 2, which is an updated version of the Deep Voice 1 framework as in Fig. 5. It supports multi-speaker setups. They then trained the model on two different datasets, one of which is a VCTK dataset, and the other is a single-speaker English dataset.

In the paper, the authors show that we can enhance Tacotron with similar techniques, and we can also use WaveNet to replace the instrument's vocoder. The results of their analysis were very promising, and Deep Voice 2 with the WaveNet-based model achieved a 3.53 MOS.

The researchers (Ping et al., 2018) then compared the Deep Voice-3 model which is shown in Fig. 6 model with the models that were built using RNNs. They also experimented with other methods for generating waveforms. The WaveNet consistently performed well. They achieved

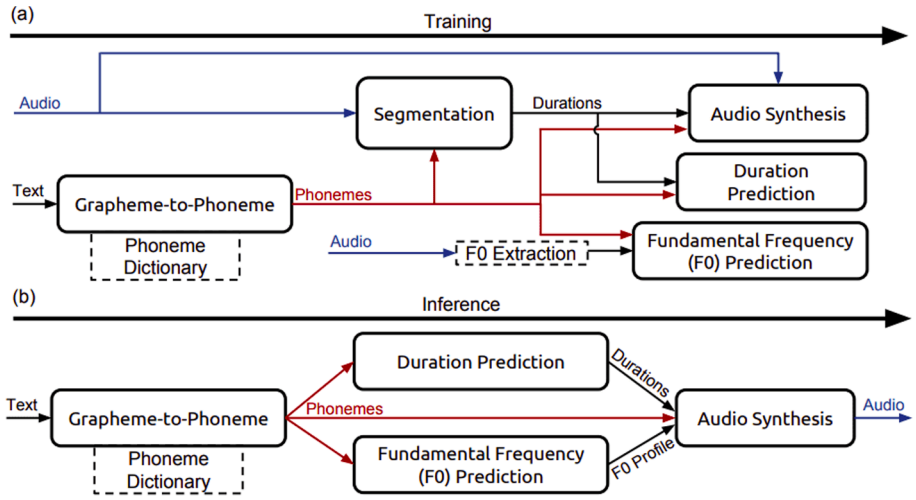

Fig. 4. Deep Voice System diagram representing (a) the training process and (b) the inference process (Arik et al., 2017).

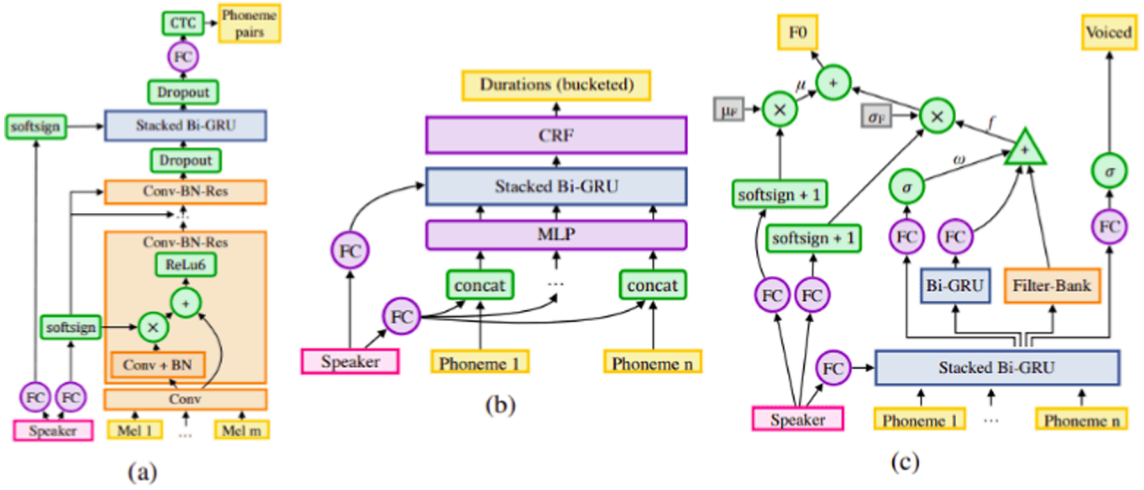

Fig. 5. Structural design for the multi-speaker (a) segmentation, (b) duration, and (c) frequency model (Arik et al., 2017).

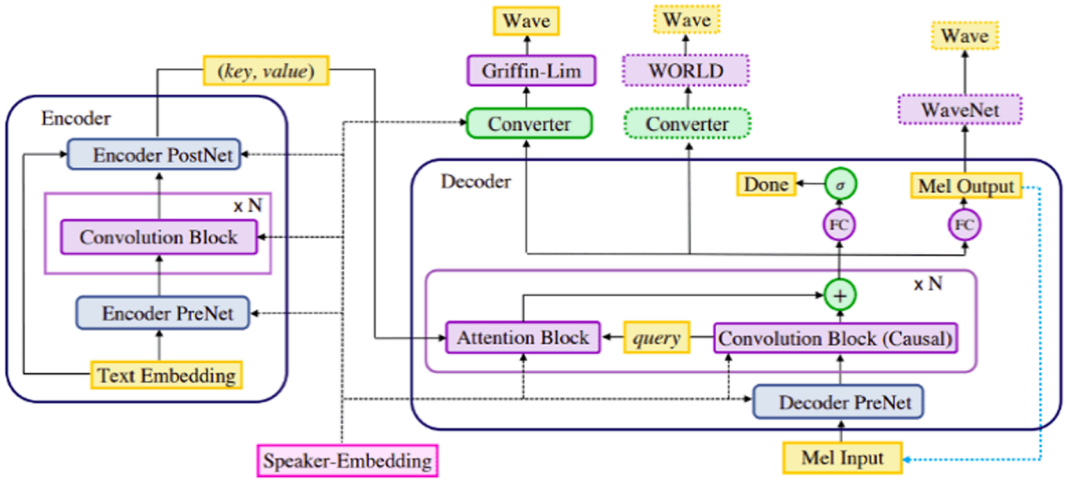

Fig. 6. Deep Voice 3 utilizes residual convolutional layers to encode input text into a form per-timestep key and value vectors for an attention-based decoder (Ping et al., 2018).

real-time inference by implementing optimized GPU and CPU kernels. The team was able to obtain a MOS of 2.67.

The Deep Voice 3 model is completely redesigned. It has a single model that's ideal for parallel computing, and the authors proposed a completely convolutional style of architecture. This is different from RNN models, and they were able to achieve the best results with WaveNet.

The researchers built the model using three components. One of these is the encoding component, which is a convolutional network. It transforms various features, such as phonemes and characters, into a condensed audio file. The second component, known as the decoder, is a low-dimensional audio file that forecasts the vocoder parameters' final shape. The third built the model by implementing a converter, which is an unsupervised post-processing network that considers the hidden states of the vocoder. This component can be used in the future to rely on context knowledge. For training, the researchers used over 20 h of English single-speech data, over 44 h of VCTK data, and over 800 h of LibriSpeech data.

2.4. Tacotron

In 2017, Google released the Tacotron system (Wang et al., 2017), the system was designed and built to provide end-to-end services. It featured a sequence model that followed the famous encoder-decoder framework. Fig. 7 shows the block diagram of the Tactron structure.

As input characters, the model takes the final speech and outputs its raw spectrogram, which is later converted to a waveform. The CBHG is a type of module that's used in neural machine translation. It's a representation pipeline that's designed to extract sequences from a set of data as in Fig. 8.

The researchers (Shen et al., 2018) trained Tacotron using a pair of English and Mandarin-Chinese speech data sets. They compared the performance of the framework with a parametric model and a concatenative method. The study's results indicated that the Tacotron framework outperformed the parametric system. The researchers utilized the Tacotron 2 framework with fewer complex building blocks as in Fig. 9. It shifted from CBHG stacks to LSTM and convolutional structures in the decoder and encoder, and it doesn't utilize a reduction factor. Instead of using additive attention, the researchers utilized location-sensitive attention. The system was trained on a 24.6-hour speech dataset. The model achieved a maximum output of 4.53 on a scale of 5.

The Tacotron 2 architecture is more streamlined and improves upon its predecessor. The new version of the Tacotron 2 architecture features a bidirectional LSTM and three convolutional layers. It replaces the CHBG and PreNets modules. The original mechanism for addressing additive attention was enhanced by location-sensitive attention. A new Autoregressive RNN is now implemented by utilizing a Pre-Net, two LSTMs, with a 5-layer Convolutional Post-N.

A modified WaveNet is then employed as the Vocoder in this architecture. It follows the parallel and pixel CNN++ methods. Instead of linear-scale spectrograms, the generated Mel spectrograms are sent to the Vocoder. The WaveNet algorithm was then used in the first version of Tacotron. It replaces the Griffin-Lin algorithm. The latest version of Tacotron 2 achieved a remarkable MOS of 4.53.

In (Zhang et al., 2019), proposed a method that combines prosodic annotation with the Tacotron model to produce natural Chinese and rhythm-based expressions. The training batch included over 30 h of

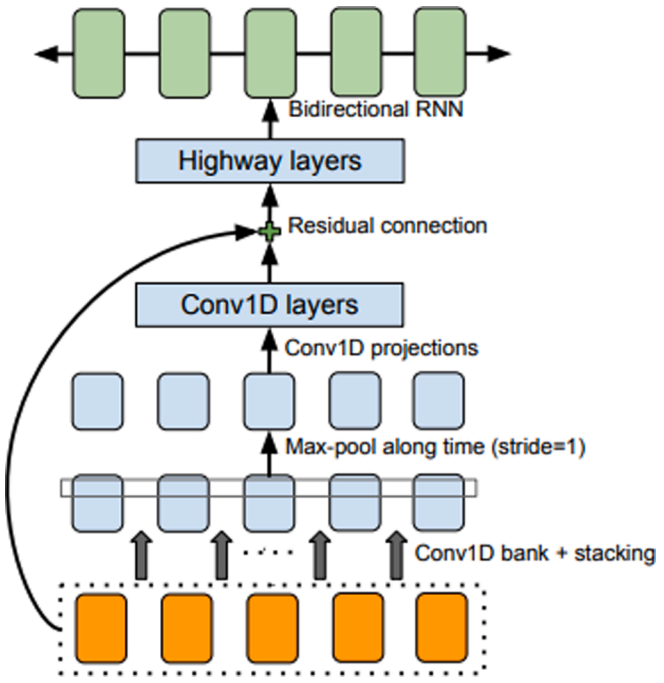

Fig. 8. The CBHG (1-D convolution bank + highway network + bidirectional GRU) module was adapted from (Lee et al., 2016).

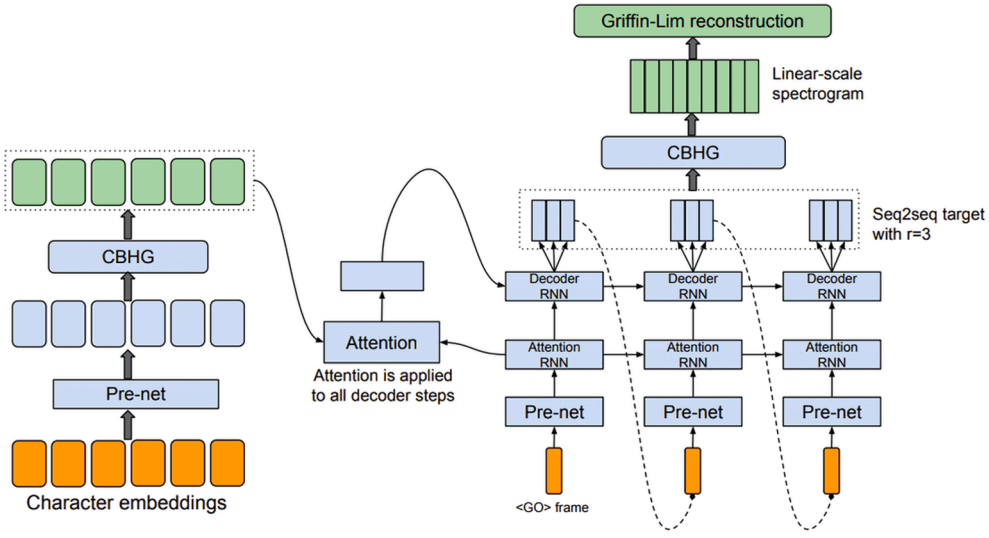

Fig. 7. The model architecture takes characters as input and outputs a spectrogram corresponding to the corresponding raw signal. This is then fed to a Griffin-Lim algorithm for synthesizing speech (Wang et al., 2017).

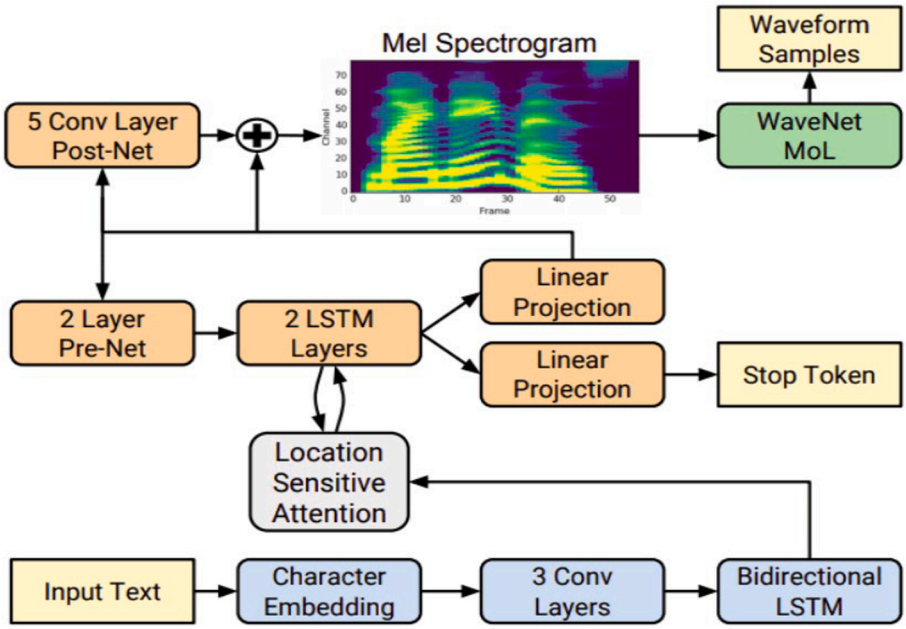

Fig. 9. The Tacotron-2 system's block diagram outlines the architecture (Shen et al., 2018).

Chinese-speaking female voices. The researchers also trained the Tacotron2 and Wavenet vocoders on up to 100 K measures to produce high-quality voices. They added prosodic phrasing by setting the parameters in three contexts: N-gram, P-Word, and Full-Sen.

To improve the model's prosodic phrasing, (Liu et al., 2020) suggested implementing a learning scheme of two tasks. They used the TH-CoSS and Mongolian speech data to train the model. They also utilized the algorithm of Griffin-Lim to fast-track the transformation. The same researchers performed listening exercises with twenty Mongolian and Chinese utterers. The MTL-Tacotron framework, which was designed for the training system, reliably outperformed other contrastive structures in out-of-the-domain inference. In addition, the researchers presented a unique training strategy that combines the student–teacher pipeline and the neural end-to-end framework. The researchers utilized the Griffin-Lim algorithm to create waveforms for their studies. To check the robustness and naturalness of their Chinese and English language training systems, they performed several tests on them. The proposed algorithm Tacotron-2-KD knowledge-distillation framework performed well against the baseline systems. For post-processing and encoder networks, the (He et al., 2020) suggested the use of the Tacotron module. It produced similar results to the CBHG model while using a less-parameterized approach. The researchers performed all of their experiments on the biaobei corpus, a woman who frequently utters Mandarin phrases.

The researchers evaluated the training system using 50 random sentences. After comparing the various training systems' performance, the researchers concluded that the DOP Tacotron framework performed better than the original Tacotron. In (Win and Masada, 2020) trained their model using Tacotron 2 on a Myanmar corpus for five hours. The result was a MOS = 3.89. Hayashi and colleagues also published an article about the ESPnet-TTS extension, which is an open-source speech processing framework. The authors' toolkit provides high repeatability and is compatible with various E2E-TTS models. The researchers (Fahmy et al., 2020) used Tacotron 2 to train their model on a 24-hour Arabic dataset from LJSpeech, and it achieved a MOS of 4.21. In (Weiss et al., 2021) trained it using the pre-built English model and the Wave-Tacotron neural network, which directly produces voice waveforms from the inputs. The researchers utilized the Autoregressive Decoder Loop's normalizing flow to expand the model's capabilities. Output waveforms are composed of fixed-length blocks with numerous samples in each. To test their model, they used two single-speaker datasets: one from LJSpeech and one from a private dataset, which has 39 h of speech. The researchers' model (Naderi et al., 2022) was able to achieve a MOS of 4.47. proposed a Persian TTS system that was based on Tacotron 2. To train the model, they created a Persian speech dataset for 21 h.

The framework (Vainer and Dušek, 2020) was developed through a convolutional system called SpeedySpeech, which is designed to generate spectrograms using phoneme-based methods. It was able to perform well in the evaluation of training and testing of the model on a large speech dataset. To analyze the model, the researchers surveyed 40 participants, including MUSHRA (Schoeffler et al., 2018). They found that the model performed significantly better than Tacotron 2 when it came to scoring. The researchers utilized a MelGan vocoder, which is used to create high-fidelity speech across different domains (Jang et al., 2021).

2.5. *Global style tokens (GST)*

GST (Wang et al., 2018) is a set of tools that can be used to enhance the capabilities of Tacotron-based architectures as in Fig. 10. They were developed by the authors using an unsupervised method. The goal of these tools is to model various speaking styles using their acoustic expressions.

A reference encoder is utilized during the training process which is to extract information about a short vector that encodes info about a speaking style, also known as prosody. It is then sent to a style token layer, which considers the contributions of each token to its embedding. A reference audio file can be used to encode the information, or a speech style can be manually controlled.

2.6. *TTS with Transformers*

Due to the dominance of the transformer in the natural language field, inevitably, they will eventually make their way into the TTS field. One of the main issues that they aim to address is the issue of training and inference and low efficiency.

One of the biggest issues that they face is the difficulty in modelling complex dependencies using RNNs. In 2018, they introduced the first generation of transformer-based architectures that allow multi-head attention systems to be trained in parallel.

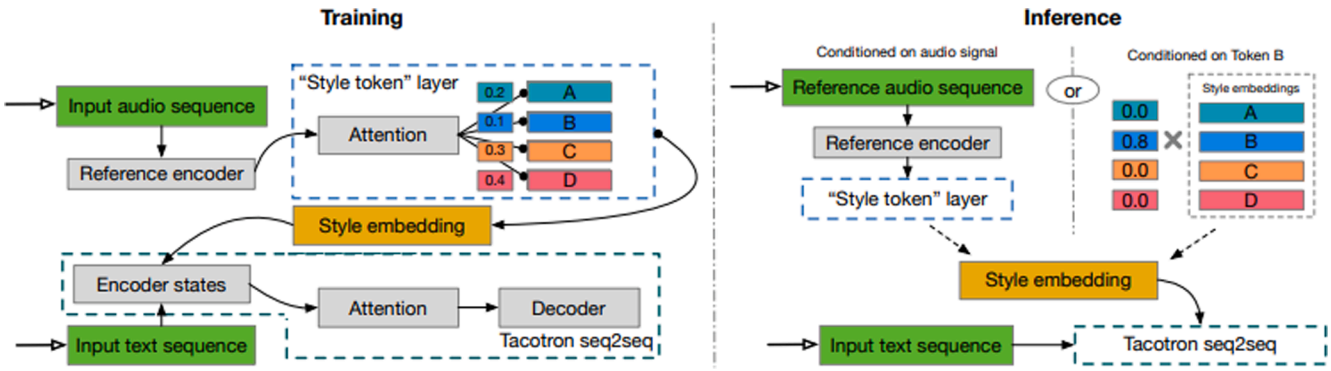

Fig. 10. The model illustrates the training target's log-myelogram, which is fed to a reference encoding by a style token layer (Wang et al., 2018).

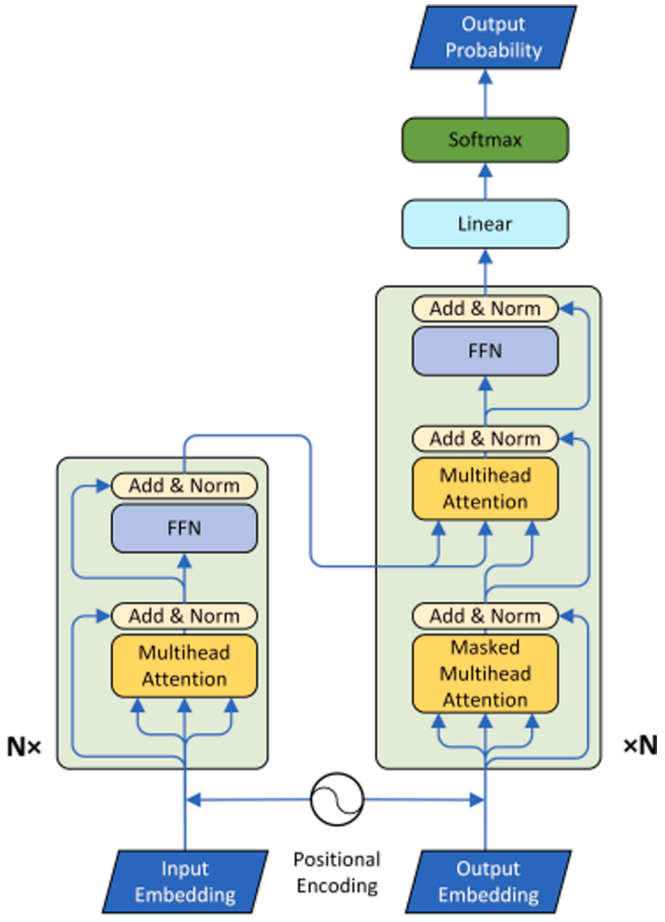

Fig. 11. System architecture of Transformer (Li et al., 2019).

The authors of the study (Li et al., 2019) presented a transformer network that can create neural speech synthesis as in Fig. 11. All texts were converted into phonemes to be used as input for the model. The researchers used 25 h of speech data from a US-English-English-speaking female dataset to train their model, which achieved a 4.39 evaluation on the MOS test. some characteristics that have:

1. A scaled positional encoding technique uses a sinusoidal form to record the position of a phoneme.
2. A 3-layer of CNN, like Tacotron 2, learns the phonemes' embeddings.
3. A decoder pre-net takes advantage of a Mel specification and projects it into a subspace that is like the phoneme embeddings.
4. A transformer-based Encoder replaces the bi-directional RNN.
5. A transformer-based decoder takes over the location-sensitive RNN. It has multi-head self-attention.

In 2017, Sotelo and colleagues presented the Char2Wav system, which is a voice synthesis framework that consists of a neural vocoder and readers. The two components were trained independently. They then utilized standard WORLD vocoder features as their targets and inputs (Sotelo et al., 2017). To train their model, the researchers used the DIMEX-100 and V-C-T-K datasets.

2.7. TTS with FastSpeech

The Fast Speech system was proposed by (Ren et al., 2019). It uses a similar strategy as a transformer. The framework is built on a feed-forward network, which is composed of various building blocks as in Fig. 12. The FastSpeech approach is like that used with the Transformers. It was able to achieve a 38x increase in performance. It is a parallel generation of mel-spectrograms that was furthermore performed. The difference between the previous model and the new one is that the former uses soft focus alignments for the phonemes while the latter uses hard alignments. A length regulator that can be used to change the voice speed by either shortening or lengthening the duration of the phoneme.

These include a span controller, the length predictors, and the feed-forward transformer. The time-consuming and difficult distillation of the student–teacher pipeline was among the issues encountered by FastSpeech. Also, the length of the extracted data from the teacher model was not precisely defined. In addition, the data loss caused by the simplification of the data structure affected the voice quality. To address these issues, (Ren et al., 2022) developed FastSpeech-2, a text-to-speech framework that is fast and high-quality. It was trained on a 24-hour speech dataset from LJSpeech. The model was able to achieve a score of 3.83. The DeepMind's EATS-end-to-end framework was developed to provide a text-to-speech (Donahue et al., 2021), generative model that is fast and accurate. It features an element known as the aligner, that converts the unaligned text to a representation that's aligned with the output. It also has a component known as the decoder, which increases the audio frequency of the aligner. They used a private dataset of 69 female and 69 male English speakers to evaluate the model.

2.8. TTS with flow-based TTS and WaveGlow (Prenger et al., 2018)

To understand flow-based TTS, let us first distinguish it from VAEs and GANs. Unlike GANs and VAEs, flow-based models do not rely on the probability density of our data, but rather on normalizing flows. Many TTS models have been proposed using the concept of flow-based modelling. Some of these include RealNVP, Glow, and NICE. Lillian Weng's excellent article can help you learn more about this idea. Flow-based models are commonly used in speech synthesis. Nvidia's Wave-Glow is a popular flow-based TTS model (Prenger et al., 2018). It

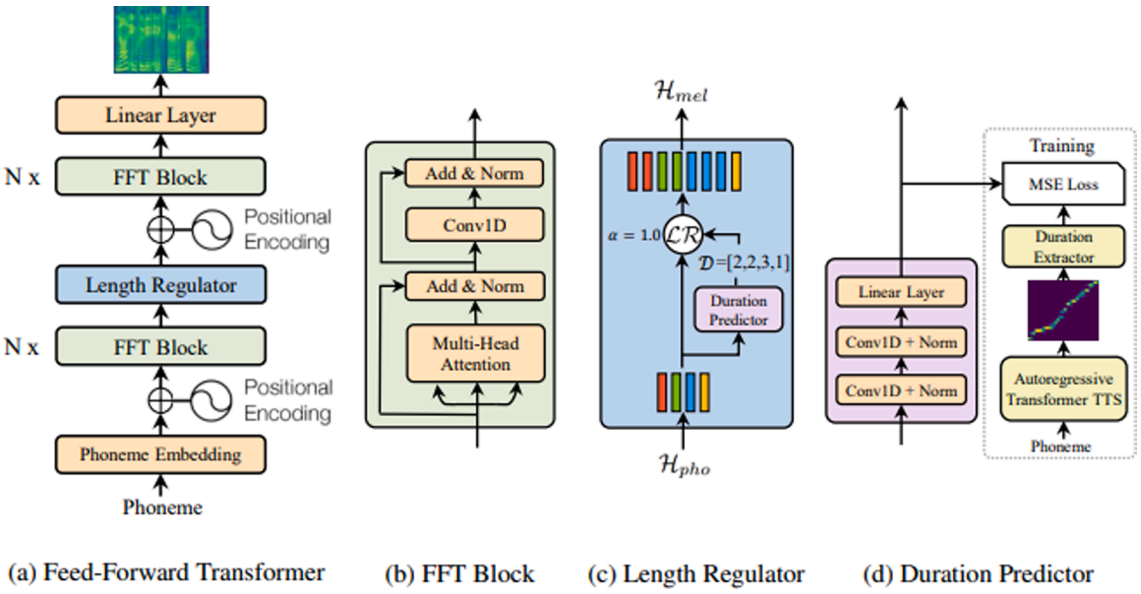

Fig. 12. The overall architecture for FastSpeech (Ren et al., 2019).

combines the insights of WaveNet and Glow, Auto-regression is not required for high-speed audio synthesis. WaveGlow is mainly utilized for speech generation. It does not have end-to-end capabilities. The goal of the training process is to minimize the log-likeliness of the data. This is done using Invertible neural networks. Fig. 13 shows the model architecture.

The trained model will then run through the network with random sampling values. Flow-TTS and Glow-TTS are examples of similar systems. The Autoregressive network is used by Flowtron to generate speech. This shows that there are still research opportunities in this area of flow-based modeling.

In (Ping et al., 2019), presented ClariNet, a parallel wave-generating framework that is constructed by Gaussian inverse autoregressive flow (IAF). They were able to easily train it using the maximum likelihood method. They also introduced the first fully convolutional text-to-wave network for voice synthesis, which allowed them to quickly train their model. The researchers used an English speech dataset with over 20 h of audio to train their model. They were able to outperform the previous pipeline, which was only trained with WaveNet.

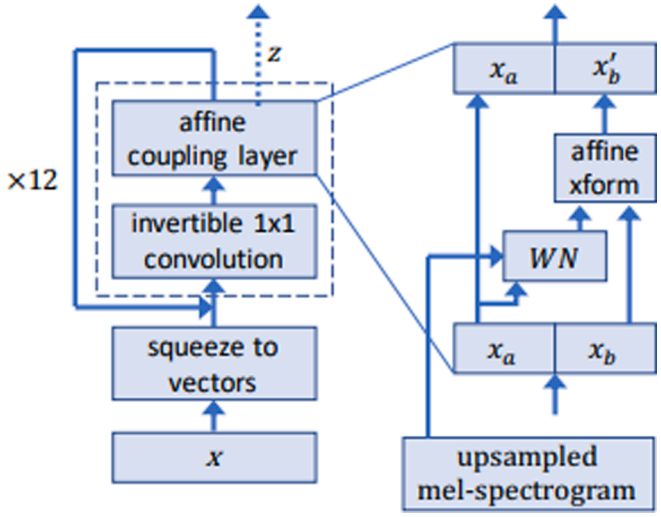

Fig. 13. WaveGlow Network Architecture (Prenger et al., 2018).

2.9. GAN-based TTS and EATS

One of the most significant works of recent years. DeepMind's EATS is an End-to-End Adversarial Text-to-Speech system that depends on Generative Adversarial Networks (GAN) (Donahue et al., 2021).

The EATS system employs the adversarial network training method. It can operate on raw phoneme sequences or pure text and produces output signals. The decoder and aligner are its basic submodules.

The aligner converts the raw input sequence into low-frequency features and writes them in an abstract space. The job of the aligner is to align the input sequence with the output. Meanwhile, the decoder takes advantage of its 1D convolutional capabilities to produce audio files. The system is trained as an adversarial entity. Fig. 14 shows the model architecture. A feed-forward neural network called EATS uses a variety of alignment schemes to generate audio. It can also capture temporal variations in the output.

Table 1 shows a list of summaries of some works in TTS, in addition to the methodologies used, the WaveNets performance for various studies was evaluated employing MOS and subjective paired comparisons. It performed well against the baselines of statistical-parametric and categorical voice synthesizers.

3. Kurdish language

This section covers the various dialects of the Kurdish language and provides an extensive explanation of the Central Kurdish writing system. In addition, we also talk about the pronunciation points and other aspects of the language. These explanations are significant since they help to develop a lexicon for any work. Most Kurds living in Iran and Iraq speak the Central Kurdish dialect. The written form of this language, which is known as Sorani, was developed during the 1920s (Nebez, 1993). It was then adopted as the official language of Iraq. The Northern Kurdish dialect is commonly used in areas such as Turkey, Northern Iraq, Syria, and Northwestern Iran. The Southern Kurdish is a language spoken in areas such as Kermanshah and Ilam in Iran and parts of Iraq. Other forms of the Kurdish language are also known as Hawrami or Gorani and Zazaki. The two dialect forms of Kurdish are the badeni or Kurmanji (Northern Kurdish) and the (Sorani) Central Kurdish. The former is written in Latin script, while the latter is usually written in Arabic script. Despite having a greater number of speakers, there are more written resources in Central Kurdish. The Central Kurdish writing system was first developed during the 1920s (Nebez, 1993). It has

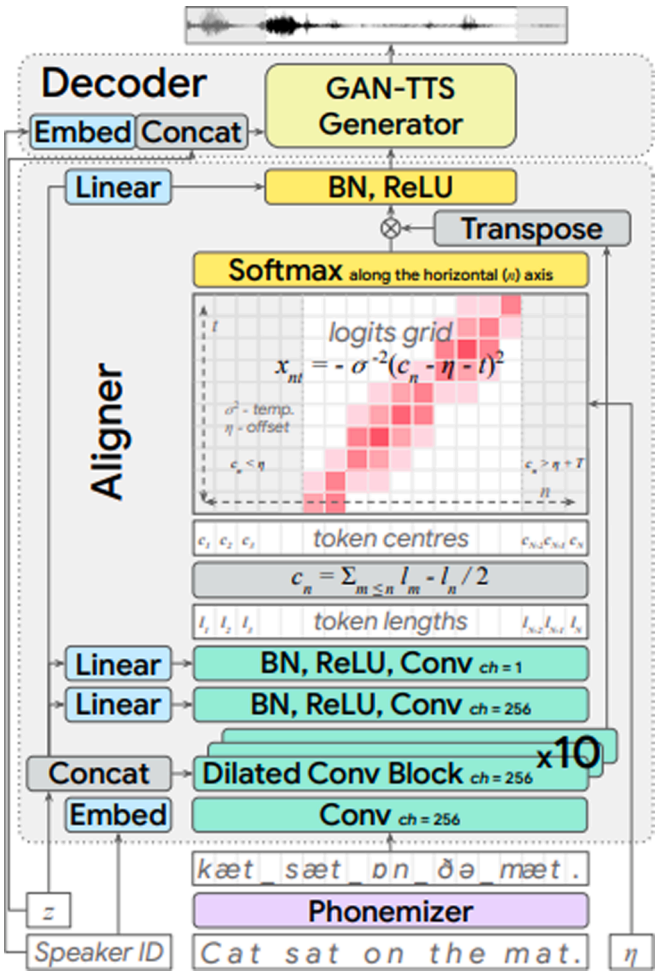

Fig. 14. A diagram of the generator, including the monotonic interpolation-based alignment (Donahue et al., 2021).

undergone a lot of changes since then. Table 2 shows the various features of the Central Kurdish's phonological system.

The central Kurdish writing system is like a phonemic system. Each letter of the language is assigned a phoneme, and some exceptions are made. For instance, the letters "ی" and "j" are both pronounced as "palatal," while the "i" is a "vocet." Similarly, the letters "و" and "w" are both pronounced as "bilabial," respectively. The character "و" is also repeated in a repeating characters form as "وو" and is pronounced as "u/" or a long vowel (Kurdistan Regional Government – Information Technology (2014)).

The Central Kurdish writing system is similar to that used in Arabic and Persian writing, but there are differences (Abdullah et al., 2024; Ahmad and Rashid, 2024). Table 3 shows the difference between the letters in Persian, Arabic, and Kurdish alphabets (Veisi et al., 2019). These unique letters can be used to differentiate languages. In Persian and Arabic, the Kasre and homograph problems are usually present (Bijankhan et al., 2011). On the other hand, in the Kurdish system, the corresponding letters for these problems are present in written texts. This means that these problems do not take place.

Table 4 shows some beginning works in Kurdish and some Kurdish TTS.

Table 5 shows some methods including the advantages and disadvantages which are commonly used in tts, however, some of these methods were introduced to overcome the limitations in the previous one.

Table 1
Summarizes the main points of the literature review.

No.	Reference	Year	Method	Dataset	Result
1	Oord et al., 2016	2016	WaveNet	North American English (24.6 hrs). Dataset of Mandarin-Chinese (34.8 hrs).	MOS 4.0
2	Wang et al., 2017	2017	Tacotron End-to-End	North American English dataset	MOS 3.82
3	Arik et al., 2017	2017	Deep Voice2 TTS (Multi-speaker)	English single-speaker (20 hrs) VCTK dataset multi-speaker (44 hrs)	MOS 3.53
4	Ping et al., 2018	2017	Deep Voice3 TTS (fully convolutional attention)	English single-speaker (20 hrs) VCTK dataset multi-speaker (44 hrs) Librispeech multi-speaker (820 hrs)	MOS 3.78
5	Sotelo et al. (2017)	2017	CHAR2WAV-End-to-End TTS	the VCTK and DIMEX-100 datasets	Their result was sufficient
6	Shen et al. (2018)	2018	Tacotron 2	Dataset of North American English (24.6 hrs)	MOS 4.526
7	Ping et al., 2019	2019	Clarinet: WaveNet (Parallel)	Dataset of internal English speech (20 hrs)	MOS 4.15
8	Zhang et al. (2019)	2019	End to End Tacotron TTS	BZSYP-Chinese database.	84 %
9	Li et al. (2019)	2019	Transformer-based TTS	US English dataset	MOS 4.39
10	Ren et al. (2019)	2019	FastSpeech- non-autoregressive End to End	LJSpeech dataset (24 hrs)	MOS 3.84
11	Ren et al., 2022	2020	FastSpeech 2- End to End TTS	LJSpeech dataset (24 hrs)	MOS 3.83
12	Vainer and Dušek, 2020	2020	SpeedySpeech	LJSpeech dataset	75.24
13	Liu et al. (2020)	2020	Taotron-2-KD	English and Chinese dataset	MOS-Engl (3.93) MOS-Chi (3.94)
14	He et al. (2020)	2020	DOP-Tacotron	Biaobei speech corpus-Mandarin (12 hrs)	MOS 3.683
15	Win and Masada (2020)	2020	Tacotron2	Myanmar corpus (5 hrs)	MOS 3.89
16	Fahmy et al. (2020)	2020	Tacotron 2-Transfer Learning	Nawar Halabi's Arabic Dataset (3 hrs)	MOS 4.21
17	Weiss et al. (2021)	2021	Wave-Tacotron	LJSpeech dataset private dataset (39 hrs)	MOS 4.47
18	Naderi et al. (2022)	2022	Tacotron 2	Dataset of Persi (21 hrs)	MOS 3.01–3.97

Table 2
Phonemes and Letters of Central Kurdish Language (Veisi et al., 2019).

No	Feature	IPA	Phoneme	Letter (isolated form)	Example
1	Voiced Stop	B	B	ب	باوان
2		D	D	د	کورد
3		d͡ʒ	Je	ج	گۆڤ
4		G	G	گ	درەنگ
5	Voiced Fricative	V	V	ڤ	بەیڤ
6		Z	Z	ز	ڕێزە
7		ʒ	Zh	ژ	کەژ
8		ɣ	Xe	غ	ساغ
9		ʕ	Ah	ع	بەتەعار
10		t	T	ت	دەرمانات
11	Unvoiced Stop	t͡ʃ	Ch	چ	ورچ
12		K	K	ک	بەووک
13	Unvoiced Fricative	H	H	ھ	ھەناسە
14		ʃ	Sh	ش	ڕەش
15		S	S	س	بەسک
16	Vibrant Flap	ɾ	R	ر	کار
17	Vibrant Trill	R	rr	ڕ	گۆڕین
18	Lateral	L	L	ل	لار
19		L	Ll	ڵ	ھەوڵ
20	Nasal	M	M	م	کەم
21		N	n	ن	کۆن
22	Approximant	J	Y	ی	کەی
23		W	W	و	چەوت
24	Vowels				
25	Front High	I	I	ی	نەوی
26	Central Low	Ä	Aa	ا	باران
27	Front Mid- low	ɛ	E	ێ	ێوێ
28	Back Mid	O	O	ۆ	زۆر
29	Central-back Mid-high	ʊ	U	و	کورد
30	Back High	U	Uu	وو	لووت
31	Front Low	A	A	ە	بەش
32	Central-front Mid-high	I			دڵ

Table 3
Letters of Kurdish in comparison with Persian and Arabic letters (Veisi et al., 2019).

Language	Letters
Kurdish only	ڤ /v/ ڕ /rr/ ڵ /ll/ ێ /e/ ۆ /o/ ە /a/
Kurdish and Persian	پ /p/ چ /ch/ گ /g/ ژ /zh/
Kurdish, Persian and Arabic	ء /eh/ ا /aa/ ب /b/ ت /t/ ج /je/ ح /he/ خ /kh/ د /d/ ر /r/ ز /z/ س /s/ ش /sh/ ع /ah/ غ /xe/ ف /f/ ق /q/ ک /k/ ل /l/ م /m/ ن /n/ و /w/ ھ /h/ ی /y/
Persian and Arabic	ث /th/ ص /ṣ/ ض /ḍ/ ذ /dh/ ظ /ẓ/ ط /ṭ/ ـ - /a/ ـ- /e/ ـ - /o/ ـ- (shaddah)

4. TTS challenges and limitations

The speech synthesis field is an interdisciplinary discipline that has a large scope of problems. Among the issues that are commonly encountered in its pre-processing is the management of non-standard words. In addition, it is also prone to issues related to the pronunciation and prosody of foreign and proper nouns. One of the most common issues that speech synthesizers encounter is the discontinuity and contextual influences in the wave concatenation method. This is because children's and women's voices have a higher pitch than that of men (Lemmetty, 1999). It is a challenging task to prepare text, as various language-related issues must be resolved. Every non-standard term has to have a phonetic equivalent, and full words should be formed from numbers and digits. According to (Macon, 1996), abbreviations should be extended into whole words so that they can be spoken as if they were written. In Kurdish, there are issues with Kasra, homographs, and automatic pronunciation. For instance, some words have the same letters but have different meanings like (چەک ،خاڵ، ، شاخ ، کار، شانە، دێ، جام، ڕووت، سەوزە، پەڵ، تەڕ).

Table 4
Summarizing the main points of the Kurdish literature review.

No.	Reference	Year	Method	Dataset	Result
1	Bahrampour et al. (2009)	2009	Concatenative (Allophone, Syllable, and Diphone)	Kurdish Language	Allophone MOS 2.45 Syllable MOS 3.02 Diphone MOS 3.51
2	Barkhoda et al. (2009)	2009	Concatenative (Allophone, Syllable, and Diphone)	Kurdish Language	Best quality score 3.5 Best Diagnostic-Rhyme-Test (DRT) 97 %
3	Daneshfar et al. (2009)	2009	Concatenative (Allophone)	Kurdish Language (2100 words)	Best quality score 2.4
4	Hassani and Kareem (2011)	2011	Concatenative (Diphone)	Kurdish Language (2100 words)	Best quality score 55 %
5	Fahmy et al. (2020)	2020	Tacotron 2-Transfer Learning	Nawar Halabi's Arabic Dataset	(3 hrs) MOS 4.21
6	Naderi et al. (2022)	2022	Tacotron 2	Persian dataset	(21 hrs) MOS 3.01 – 3.97

In general, the studies on the use of TTS need to be analyzed with the help of inconsistencies in their findings. This can help us understand the possible biases and limitations of the research. The following points are some of the inconsistencies:

1) The diversity of the samples in a study can also affect the generalizability of results. This issue can prevent the proper understanding of how TTS works with various demographic groups.
2) There's a lack of standardization when it comes to the evaluation of resulting metrics that are used to assess the effectiveness of TTS systems. These include the Word Error Rate (WER), the Mean Opinion Score (MOS), and the naturalness score. Because of this variability, it can be hard to compare the systems' performance.
3) Languages and speech corpus: The corpus utilized in training and assessing TTS systems can differ widely, and this can affect the findings. This can result in TTS systems' varying performance across different speech styles and languages.
4) Training and architecture of TTS systems are typically done using different approaches. This includes neural network frameworks, data augmentation methods, and optimization techniques. The training and system architectures' variations can affect their generalizability and performance.
5) A study on text-to-speech systems may look into different contexts and user scenarios, such as accessibility applications and in-car navigation. The variations in these settings can affect the relevance of the findings to specific applications.

The inconsistencies exhibited in the research methodologies utilized in studies on text-to-speech systems highlight the need for more transparency and standardization in the field. Future research on this technology should utilize evaluation metrics, sample sizes, user scenarios, and system architectures to maintain the highest levels of reliability.

5. Speech synthesis principles

One of the most common methods of human communication is through speech (Lemmetty, 1999). This process involves converting text into a voice, which is as close to human speech as possible while following certain pronunciation rules (Aida-Zade et al., 2010). Although

Table 5
The Limitations of various speech synthesis methods.

No.	Method	Limitations	Performance
1	Hidden Markov Model (HMM)	Due to the over-smoothing of the acoustic features, the created speech sounds are muffled.	The synthesized speech could sound robotic or lacking in natural prosody, which affected the overall quality and intelligibility of the output.
2	Deep Neural Network (DNN)	Noisy data and the system's performance can be affected by irregular information. The complexity of the data model can also make it hard to train and maintain.	By training on large amounts of high-quality data, DNNs capture complex patterns and dependencies in the data, leading to more accurate modeling of speech characteristics such as prosody, intonation, and timbre.
3	Deep Belief Network (DBN),	The training efficiency of DBNs is low. They tend to forget about logic and reasoning because they do not have the necessary representation.	DBNs typically require a large amount of labeled training data to learn meaningful representations. Acquiring and aligning such data for TTS can be challenging and resource-intensive.
4	Convolutional Neural Network (CNN).	Due to the overfitting issue, the computational cost can be high.	(CNNs) have shown promising performance in certain aspects of Text-to-Speech (TTS) systems, particularly in modeling the acoustic features of speech. While CNNs are commonly associated with image processing tasks, they can be adapted and applied to speech synthesis with notable results
5	Recurrent Neural Network (RNN)	The recurrent behavior of computation slows it down. It is also hard to train due to its exploding or vanishing gradient problems. It cannot complete the long sequence of steps.	RNNs are particularly well-suited for modeling sequential data, making them highly effective in capturing the temporal dependencies and linguistic context present in speech
5	WaveNet	The effects of slow processing and the errors made by the front-end components can affect the synthesis process.	Has demonstrated impressive performance in Text-to-Speech (TTS) systems. It has significantly advanced the quality and naturalness of synthesized speech by directly modeling raw audio waveforms
6	Deep Bidirectional Long Short-term Memory (DBLSTM)	A vocoder is needed to synthesize a waveform.	Has shown promising performance in Text-to-Speech (TTS) systems, particularly in modeling the temporal dependencies and context in speech. DBLSTMs combine the strengths of Bidirectional LSTMs (BiLSTMs) with the depth of deep neural networks, resulting in improved synthesis quality.
7	Long Short-term Memory (LSTM)	The recurrent behavior of computation slows it down. It is also hard to train due to its exploding or vanishing gradient problems. It cannot complete the long sequence of steps.	has been widely used in Text-to-Speech (TTS) systems and have demonstrated strong performance in capturing sequential dependencies and linguistic context
8	Restrictive Boltzmann Machine (RBM)	The training of data can also be affected by the fragmentation issue.	RBMs are generative models that can learn to capture hidden patterns in data, but they have limitations when it comes to speech synthesis
9	Tacotron	Too expensive to train the model	It has shown impressive performance in generating high-quality and natural-sounding speech.
10	GAN	Unsteady and some difficulty to train, and also the issue of Mode Collapse. Also the pattern-less GANs' learning process, the generator part will begin to degenerate, generating the same sample points repeatedly.	GANs have shown promise in improving certain aspects of TTS, but their performance can vary depending on the specific architecture and training setup

voice synthesis has been around for a long time, the current generation of models is more suitable for various applications, such as telecommunications and digital. The field of speech technology is a part of the NLP framework, which also includes other areas like speech recognition with dialog systems (Bakhsh and Alshomrani, 2014). Text-to-speech systems are commonly used in various applications. They are composed of a series of steps that are shown in Fig. 15.

The main function of text processing is to convert non-standard words into spoken ones. Text normalization is a process that involves changing non-standard terms, such as numbers, currencies, abbreviations, and numbers, into similar words that can be used for a certain phonetic conversion (Yang et al., 2017). The procedure of phonetic analysis is commonly used to convert lexical-orthographic symbols into words. It can be categorized into two main methods: rule-based and dictionary-based methods (Möbius et al., 1996). The latter allows for the preservation of a maximum amount of phonological information within a restricted vocabulary.

The rule-based method is commonly used to convert limitless words into spoken ones. It generates a group of letters to be used in the phonemes rule. Most text-to-speech programs use a combination of rules and vocabulary to manage their exceptions. The process of prosody generation is performed to extract supra-segmental attributes from written text, such as lengths, intonations, and stresses. It also tries to predict the pitch patterns of speech. Although pitch change is apparent in people's speech, it is not in written documents. According to Onaol, the attributes of speech signals are connected to variations in pitch, loudness, and syllables (Dutoit, 1996).

The following are some of the most popular synthesis techniques.

1) The most common form of synthesis is the formant. It's regarded as the earliest known technique for synthesizing speech. It was used for a long time, and it typically controlled the implementation of various procedures. The formant-synthesis method is based on a rule-based approach to describing vocal tracts' resonant frequencies. The formant-synthesis method combines various arguments, such as pitch, voicing, and noise levels, to create an artificial speech wave (Kayte et al., 2015). The articulatory synthesis method directly produces high-quality speech by showing the actions of humans. The various articulatory functions that are used in the synthesis process

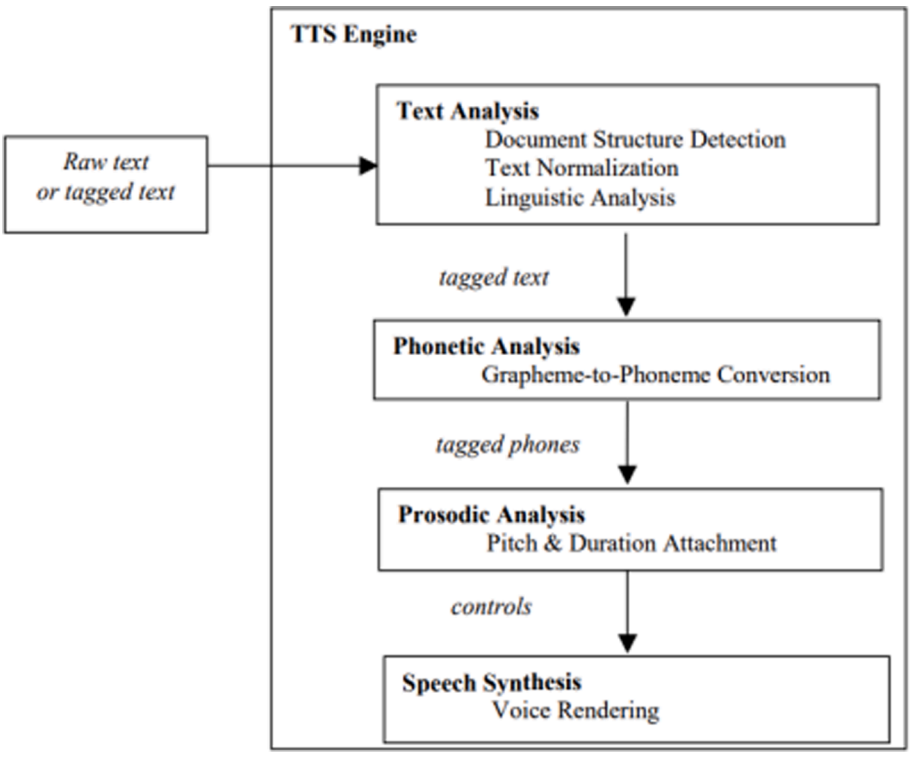

Fig. 15. TTS system architecture (Huang et al., 2001).

include tongue height, lip aperture, tongue tip protrusion, and tongue tip height. Two main challenges are related to the design and control of the system. The first is to gather a sufficient dataset for the model, while the other is to strike a balance between the easy-to-use and precise system (Klatt, 1987).

2) Concatenated synthesis follows the data-driven approach. It links pre-recorded and natural speech. Concatenative synthesis produces voice using various units, such as words, syllables, and demi syllables. The period of a device's operation affects the stability of synthesized speech. Long units with more genuineness require fewer concatenation points, and the number of them in the database grows rapidly. The space required for short units has decreased, but the methods for marking and sample collection have become more complex (Rashad et al., 2010).
3) One of the most common options is to use a statistical-parametric model for mapping parameter-to-parametric requirements. This method provides a more compact memory footprint than storing the entire dataset. Another common option is to use the HMM synthesis method, which is a statistical-parametric synthetic. It can be used for various applications, such as converting an individual's voice into a different one (Rashad et al., 2010).
4) The use of deep learning in speech synthesis is different from HMM-based methods in that it explicitly plots the linguistic features of the speech to the auditory ones. Deep neural networks can be used to study the inherent data features of complex models. They are very effective in identifying various data points. The main parts of a speech synthesis framework are the text analysis front end, a speech synthesizer, and an acoustic model. Although these mechanisms have been designed independently and are dependent on domain knowledge, errors can occur due to the training of each element. This is why end-to-end methods are becoming more popular. These methods allow the integration of various components into a single structure (Wang et al., 2016). The Wave-Net system is a generative framework that can be used for creating raw audio waveforms from the Pixel-RNN or CNN models. DeepMind first proposed this in 2016 as a feature for end-to-end voice synthesis. The Wave-Net framework can be used to create realistic sounding human-like speech sounds by the training process of a deep neural network model on a variety of speech samples (Goel et al., 2022). The Tacotron system is a complete speech synthesis framework that can be used to train models from audio sets and texts. This eliminates the need for feature engineering. Tacotron is a character-based model, which can be used in almost any language (Wang et al., 2016). It considers the text's context and maps it to a spectrogram for a strong voice approximation. Iteratively, it extracts the phase parameters from the data collected by the spectrogram (Griffin and Lim, 1984).

A comprehensive analysis of the outcomes of numerous experimental tests revealed that the synthesized speech of these models exhibited significantly better naturalness and speech quality. Speech synthesis can be subdivided into two main parts traditional speech synthesis techniques, and deep learning techniques.

5.1. Traditional speech synthesis

There are two types of methods for converting TTS files: parametric and concatenative. A concatenation method is used to directly create a speech stream by merging the various waveforms in the database. There are two types of concatenation schemes: the PSOLA method and the Linear Prediction Coefficients (LPCs) (Atal and Hanauer, 1971) method. The second method is mainly used to reduce the speech signal's storage capacity by implementing the LPC coding. It also produces a synthesized speech that is very natural. Since the speech codec is very important in preserving the details of the speech, it is commonly used to create a concatenation method that is simple and natural. However, since the concatenation points are not always the same, the overall effect of the method will be affected. To address this issue, a new type of concatenation algorithm known as PSOLA has been proposed. The PSOLA algorithm considers the target context when it comes to adjusting the concatenation points. This ensures that the final synthesized output maintains the original pronunciation while also conforming to the specified prosody.

A parametric speech synthesis is a process that involves using signal processing techniques to make speech from text. It takes a simulation of the human vocal process and uses a time-varying electronic filter to excite it. A source is defined as a sequence of periodic pulse sounds that represent the vocal cord's vibration, or a white noise that indicates the unvoiced speech. With the filter's parameters adjusted, the process can produce various types of speeches. Some of the common methods used for this process include the synthesis of synthesized vocal organ sounds, HMM, and DNN. The SPSS system is composed of various modules. One of these is a text analysis module, which can analyze the various acoustic features, such as frequency, duration, and spectral parameters. Another module is a parameter prediction tool, which can predict the acoustic features of a speech. Text analysis is mainly used to preprocess the input text before it is transformed into linguistic features that are then utilized by a speech synthesizer. These include normalization, word segmentation, and grapheme to phone.

Parametric speech synthesis is a process that involves using signal-processing techniques to generate speech from text. It takes a simulation of the human vocal process and uses a time-varying electronic filter to excite it A source can be a sequence of periodic pulse sounds that correspond to the vocal cord's vibration, or a white noise that indicates the unvoiced speech. With the filter's parameters adjusted (Xu, 2007), the process can produce various types of speeches. Some of the common methods used for this process include the synthesis of synthesized vocal organ sounds, HMM, and DNN (Meng, 2013; Zhuang et al., 2009). A complete SPSS system is composed of various modules. One of these is a text analysis module, which can analyze the various acoustic features, such as frequency, duration, and spectral parameters. Another module is a parameter prediction tool, which can predict the acoustic features of a speech. Text analysis is mainly used to preprocess the input text before it is transformed into linguistic features that are then utilized by a speech synthesizer. These include normalization, word segmentation (Zen et al., 2013), and grapheme-to-phoneme conversions (Fan et al., 2014).

It is widely known that the context information that a phoneme provides affects its acoustic features. This means that the acoustic features can be predicted using the context information. According to researchers, the process used in speech generation converts the context into a speech waveform by using a hierarchical structure. Through the use of deep structure models, researchers can now forecast the acoustic feature parameters of a speech synthesizer. This process, which is commonly referred to as speech synthesizer synthesis, is performed by synthesizing a speech waveform based on these parameters. Traditionally, the HTS_engine synthesizer (HTS, 2015) is used for synthesizing speech. Unfortunately, this method usually produces dull sounds. To enhance the quality of synthesized speech, a new algorithm known as STRAIGHT (Banno et al., 2007) is proposed.

5.2. Deep learning techniques

In the past few years, restricted Boltzmann machines (RBMs) (Ling et al., 2013) have been extensively employed in various applications, such as speech analysis and spectrogram coding. They are also commonly used in the training process of deep neural networks (DNNs) (Deng et al., 2010; Gehring et al., 2013). In speech synthesis, RBMs are commonly utilized to generate the acoustic parameters' spectral envelopes. They are then used to address the over-smooth issue by illustrating the distribution of high-dimensional envelopes. Despite the high naturalness of the model of speech synthesis built on the Deep Neural Network (DNN), it still has limitations when it comes to modeling the acoustic feature parameters. To solve these issues, the authors (Zen and Senior, 2014) proposed a method that considers the mixture density network. This method can predict the distribution of the output features under various input features. The authors noted that the use of mixed-density networks (MDNs) (Bishop, 1994) can allow them to map the input features of the speech synthesis model to the GMM parameters like mixture variance, weights, and mean.

Despite the success of the deep synthesis model in improving the naturalness process of synthesized speech, there still are some issues that need to be resolved. One of these is its limitation in the use of contextual information. For instance, it can only model the input features of fixed periods. Another limitation of the deep model is its capability to map each frame independently. This issue was addressed by the authors of a study (Graves and Schmidhuber, 2005), who proposed a method that uses neural networks. The main advantage point of this approach is its ability to map inputs and outputs using context information. Traditional RNNs are limited by their limitations when it comes to accessing context information. For instance, they can only map the effects of an input to a hidden layer and its output to an output layer, which can explode as it moves through the network. To solve this issue, the authors proposed a memory cell that can be used to store long-term dependencies. To fully utilize contextual information, the authors of the study (Graves et al., 2006) proposed a method called the bidirectional LSTM model. This allows the system to map the input features of speech to acoustic ones. In general, we can summarize the research works done in speech synthesis in Fig. 16.

A summary of the studies held in TTS can be seen in three stages:

a) In the first stage, studies conducted analyzed the performance of various text-to-speech synthesis techniques. They found that the statistical parametric method performed better than concatenative and rule-based methods in terms of flexibility and naturalness when controlling various speech attributes. The studies highlighted the weaknesses and strengths of different synthesis techniques. This helps in comprehending the various aspects of speech synthesis.
b) In the second stage, the studies conducted in 2018 revealed that the employ of deep neural networks could enhance the intelligibility and naturalness process of synthesized speech. Their findings support the current understanding of the various methods used in speech synthesis (Mehrish et al., 2023).
c) The third stage, conducted by studies investigated the use of wavelet models for enhancing text-to-speech capabilities. In their studies, they noted that the models can produce expressive and natural speech, making them more effective than traditional methods. The studies highlight the potential of neural networks in the area of speech synthesis, which would lead to advancements in the field. The studies presented in this section provide valuable insight into the various aspects of speech synthesis. They also highlight the advancements that can be achieved through deep learning and neural networks. These findings are valuable for practitioners and researchers in the field as they help in advancing the techniques.

6. Common datasets

A good TTS dataset involves several steps. The first step is gathering a large speech data amount, which can be collected from different sources, such as public domain recordings and audiobooks. The resulting dataset ensures that the synthesized speech can cater to a wide variety of users. After the raw speech data has been collected, it undergoes a thorough cleaning process to remove any traces of background noise. The data is then annotated to ensure alignment of the speech segments with the text. This step is very important to train the TTS models on how to understand the link between speech and text. One of the most important factors that a good TTS system needs to consider when it comes to developing its capabilities is the ability to support multiple languages. Having a well-defined multilingual dataset can help train the models to accurately perform various tasks in different languages.

1) American English Speech Synthesis (19.46 Hours – Corpus-Female) (Verma and Chafe, 2021):

This is a female audio file of American English, recorded by a native speaker. It features a sweet sound and an authentic accent. The coverage

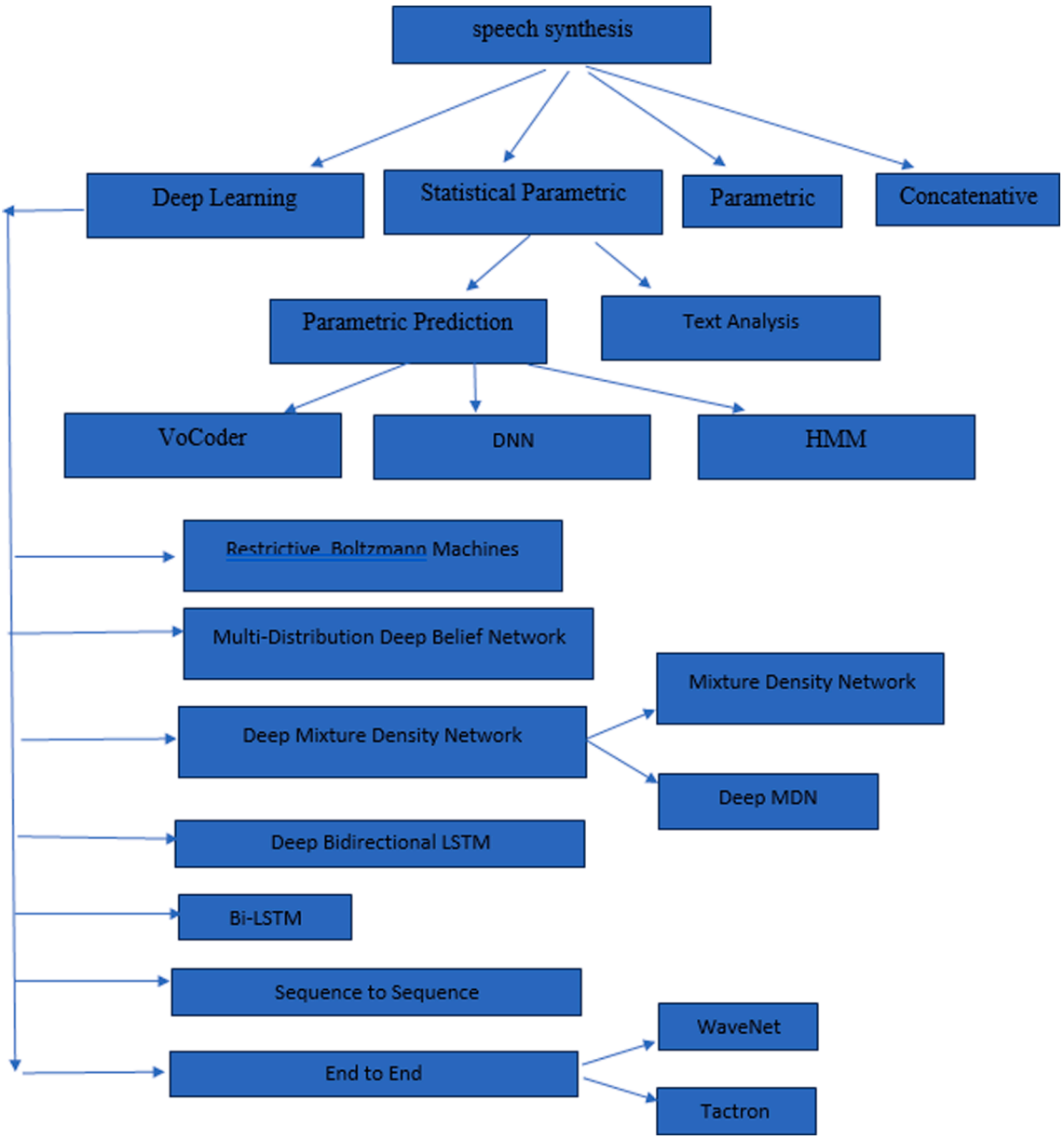

Fig. 16. Methodologies used for Speech Synthesis.

of the phoneme is balanced. A professional phonetician helps in the annotation. This works well with the speech synthesis needs of researchers.

2) American English Speech Synthesis (20 Hours – Corpus-Male) (Verma and Chafe, 2021):

The audio file of American English is composed of male voice recordings by native speakers with an authentic accent. The coverage of the phoneme is balanced and professionally annotated. This works well with the speech synthesis needs of researchers.

3) Japanese Synthesis (10.4 Hours – Corpus-Female) (Kaneko et al., 2017; Saito et al., 2018):

The audio file of Japanese is recorded by a native speaker with an authentic accent. Its coverage is professionally annotated and balanced. It perfectly fits with the needs of speech synthesis researchers.

4) Chinese Mandarin Multi-Emotional Synthesis Corpus (22 People) (Oord et al., 2016):

The Chinese Mandarin multi-emotional synthesis corpus is composed of 22 individuals, each with their unique emotional content. Its recording features balanced tones, emotional text, syllables, and phonemes. The annotation by a professional phonetician ensures that the corpus accurately fits with the speech synthesis researchers' needs.

7. Evaluation metrics

One of the most common ways to determine whether a system is meeting the needs of its users is by asking them to evaluate it subjectively. Another method is to analyze the system using an objective criterion that can be automated. Finally, behavioral evaluation is a less common approach. This is a test that aims to determine if the system can improve the performance of its users. A brief introduction about the current state of the art in the evaluation of systems using subjective, behavioral, and objective criteria is presented. Also despite the lack of

comprehensive quality data, a wide variety of metrics can be utilized to evaluate a system. Although some of these are independent, comprehensiveness is typically regarded as a requirement for most speech-based systems when it comes to task success. Despite the limitations of speech-based systems, they can still support various tasks.

7.1. Objective assessment for TTS

The goal of objective assessment is to classify a system. Automated scoring methods are attractive because they can reduce the need for subjective evaluations, but they do not align well with how people perceive information. Although automated scoring methods are useful for system tuning, they do not replace the need for subjective assessments. Also, not every trait can be evaluated objectively. For instance, some objective measures need to be capable of comparing and learn about the noise signal in real life to be effective. In most cases, objective measures are focused on intelligibility, prosodic correlates, and segmental quality. When trying to assess "naturalness," the goal is usually to capture the features of the speech that are most commonly used in a noisy environment. However, this approach tends to ignore prosody and considers it as a secondary issue. Speech quality assessment aims to use the Perceptual Evaluation of Speech Quality (PESQ) family standards and the melceptstral distortion (MCD) (Onaolapo et al., 2014) to evaluate the original speech and then compare and score the synthetic utterances. This method is performed by warping the two signals to align them. In addition to this, the distance between the two signals is also averaged over time.

Numerous efforts have been made to develop advanced scoring techniques for synthetic speech that are based on machine learning. Although the correlation between stimulus-level and system-level results is generally good, it is not as good as with AutoMOS (Patton et al., 2016). More impressive results came from AutoMOS, but this system was only trained on one speaker and is not widely available. Due to the emergence of high-quality synthesis models, such as WaveNet (Oord et al., 2016), we have finally gotten a synthesizer that can produce speech waveforms that are both human-like and natural (Malisz et al., 2019). These models store a lot of information about the actual sound and appearance of the speech. It is possible to say that a trained synthesizer can identify a speech waveform that is similar to a human-like speech without having access to a comparable recording. This could be useful in assessing the accuracy of the system's performance. However, it is not yet clear if this method can be applied to other speakers.

7.2. Subjective assessment for TTS

This is a widely used technique to measure the quality of interaction by asking users to provide their opinions on various quality dimensions, such as perceived intelligence, likability, and intelligibility. However, this method is risky since it does not take into account the users' expectations and needs. This drawback is usually solved by conducting extensive surveys, which attempt to address every aspect of a user's interaction profile. Unfortunately, this method can lead to infirm responses due to boredom or fatigue (Lavrakas, 2008). Also, the questionnaires don't usually take into account the user's expectations, which can affect the quality of interaction. Nonetheless, it is still a useful tool for assessing the interaction quality.

Most of the time, questionnaires are used to capture a global view of signal quality by asking users to provide their opinions on various quality dimensions, such as perceived intelligence, likability, and intelligibility. They can also target more complex systems by asking users to provide their opinions on multiple stimuli with a hidden reference and anchor (MUSHRA) (International Telecommunication Union, 2015). A different approach has been developed to address the issue of TTS-related issues in an in-progress interaction (Edlund et al., 2015). This method involves having third parties evaluate the interaction and provide a binary response in the moments when the issues arise. The advantage of this method is that it can provide a subjective assessment of the interaction quality, while physiological and behavioral metrics are usually hard to interpret. However, certain factors such as eye/mouse tracking and EEG can help identify mismatches between the actual realization and the user's expectations.

In speech synthesis various evaluation metrics are commonly used, such as the Mean Opinion Score (MOS), the Voiced/Unvoiced (V/UV), the F0 Root-Mean-Square Error (RMSE), and the Mel cepstral (MCEP) (Zen and Senior, 2014). Among these, the MOS has been used in most papers for objective evaluation. Speech quality is the most frequently used metric in speech synthesis to estimate the performance of various models, algorithms, and architectures. MOS is commonly used to calculate the average quality score, which makes it the most popular figure. Some of the works have used these evaluations MOS (Zen and Senior, 2014; Ali et al., 2021; Takamichi, 2017; Li et al., 2021; Valentini-Botinhao et al., 2016; Batista et al., 2019; Wang et al., 2017; Sun et al., 2015; Möbius et al., 1996). MCEP (Zen and Senior, 2014; Rebai and BenAyed, 2013; Verma and Chafe, 2021; Luong et al., 2017; Choi et al., 2019; Oyucu, 2023). Voiced/unvoiced (V/UV) (Zen and Senior, 2014; Shechtman and Mordechay, 2018; Chen et al., 2018; Choi et al., 2019; Lee et al., 2018; Wang et al., 2017). A lower RMSE is desirable than a greater one (Zen and Senior, 2014; Shechtman and Mordechay, 2018; Luong et al., 2017; Chen et al., 2018; Choi et al., 2019; Lee et al., 2018; Wang et al., 2017; Sun et al., 2015).

7.3. Behavioral assessment for TTS

If the established measures of intelligibility are not sufficiently accurate after a few years, then they should be replaced by measures that are more likely to be used in experimental systems (Benoît et al., 1993). These include the process of word error rate estimations, rhyme tests, and word edit distance (Voiers, et al., 1975; Jekosch, 1992). Due to the advent of more complex systems, the need for precise measures of intelligibility has become less of a concern. In some experimental setups, they play a vital role. In addition to intelligibility, comprehensiveness is another measure that is not as well studied. This is the degree to which a given message's pragmatics and semantics have been understood. Some researchers claim that comprehensiveness can be evaluated indirectly by asking questions that can help a listener determine how well they have been able to grasp the message's content (Fellbaum, 2014; Duffy and Pisoni, 1992).In terms of assessing the effectiveness of TTS in interactive systems, behavioral performance is most commonly used. This is done by determining the amount of retrieved data that a listener has after interacting with a dialogue system (Betz et al., 2018). Other metrics such as efficiency and effectiveness are also taken into account to evaluate dialogue systems (International Telecommunication Union, 2003).

Although long interactions are typically regarded as an indication of poor system performance, they can also be used to assess the quality of a system that is designed to entertain. For instance, if a game software has a long interaction time, it might be considered a good system. However, quality metrics are not always independent of the testing application they are evaluating. There was another kind of behavioral analysis utilized in (Gustafson et al., 2005), where participants' verbal adaptation to a dialogue task was evaluated. The results indicated that a higher degree of adaptation was associated with a better user experience. Although performance metrics are usually used to evaluate a system's comprehensiveness, they can also be misleading when it comes to identifying the factors that caused a particular interaction. For instance, if a user has a high level of listening effort, then a system might not be able to explain how it has been able to achieve this.

One of the most important factors that can be considered Regarding evaluating the efficacy of TTS is the continuous monitoring of the interaction. This was done through a combination of physiological and behavioral metrics. Through a visual world paradigm, the researchers were able to explore the effects of a TTS on listener comprehension. A

simple GUI-based game was analyzed by the researchers to see how it performed in terms of performance and response times. They also experimented with different physiological metrics such as EEG and pupil dilation (Antons et al., 2012; Govender and King, 2018).

8. Conclusion

The field of speech synthesis has started to focus on the naturalness and expressiveness of speech. However, the use of concatenative and parametric methods has hindered the synthesis' naturalness. Deep learning-based methods are increasingly being used in the field of speech synthesis to emphasize the natural attributes of speech. This paper presents a review of some works related to the Kurdish language taxonomy of synthesized speech, which includes various models and architectures commonly used in the process. Despite the technological advancements that have occurred in the field of speech synthesis, there is still a huge opportunity for developing high-quality speech. With the emergence of lightweight deep neural network architectures, it can be possible to produce Kurdish speech from text shortly.

Funding

This research received no external funding.

CRediT authorship contribution statement

Hawraz A. Ahmad: Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Tarik A. Rashid:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is available based on requirements.

Acknowledgments

The authors would like to thank Salahaddin University-Erbil and the University of Kurdistan Hewlêr for supporting and facilitating this research work. The authors would also like to thank Giga Ant Technology (Gigant) for funding this project.

References

Aida-Zade, K.R., Ardil, C., Sharifova, A.M., 2010. The main principles of text to speech synthesis system. World Acad. Sci. Eng. Technol. 37 (1), 13–19. https://doi.org/10.5281/zenodo.1070639.

Abdullah, A.A., Veisi H., Rashid, T.A., 2024. Breaking walls: pioneering automatic speech recognition for Central Kurdish: end-to-end transformer paradigm. arXiv: 2406.02561v1. doi: 10.48550/arXiv.2406.02561.

Ahmad, H. A., & Rashid, T., 2024. Gigant-Ktts Dataset: Towards Building an Extensive Gigant Dataset for Kurdish Text-to-Speech Systems. Available at SSRN 4826641.

Ali, A. H., Magdy, M., Alfawzy, M., Ghaly, M., Abbas, H., 2021 Arabic speech synthesis using deep neural networks. In: Proc. International Conference on Communications, Signal Processing, and their Applications. pp. 1–6.

Antons, J.-N., Schleicher, R., Arndt, S., Möller, S., Porbadnigk, A.K., Curio, G., 2012. Analyzing speech quality perception using electro-encephalography. IEEE J. Sel. Top. Signal Process. 6, 721–731.

Atal, B.S., Hanauer, S.L., 1971. Speech analysis and synthesis by linear prediction of the speech wave. J. Acoust. Soc. Am. 50 (3), 637–655. https://doi.org/10.1121/1.1912375.

Arik, S., Diamos, G., Gibiansky, A., Miller, J., Peng, K., Ping, W., Raiman, J., Zhou, Y., 2017. Deep voice 2: multi-speaker neural text-to-speech. arXiv preprint arXiv: 1705.08947.

Bahrampour, A., Barkhoda, W., Azami, B.Z., 2009. Implementation of three text to speech systems for Kurdish language. In: Iberoamerican Congress on Pattern Recognition. Springer, Berlin, Heidelberg, November, pp. 321–328.

Bakhsh, N.K., Alshomrani, S., 2014. A comparative study of Arabic text-to-speech synthesis systems. Int. J. Inf. Eng. Electron. Bus. 6 (4), 27–31. https://doi.org/10.5815/ijieeb.2014.04.04.

Banno, H., Hata, H., Morise, M., Takahashi, T., Irino, T., Kawahara, H., 2007. Implementation of realtime STRAIGHT speech manipulation system: report on its first implementation. Acoust. Sci. Technol. 28, 140–146. https://doi.org/10.1250/ast.28.140.

Barkhoda, W., ZahirAzami, B., Bahrampour, A., Shahryari, O.K., 2009. A comparison between allophone, syllable, and diphone based TTS systems for Kurdish language. In: 2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), December, pp. 557–562.

Batista, C., Cunha, R., Batista, P., Klautau, A., Neto, N. Utterance copy in formant-based speech synthesizers using LSTM neural networks. In: Proc. 8th Brazilian Conference on Intelligent Systems, 2019. pp. 90–95.

Benoît, C., Grice, M., Hazan, V., 1993. The SUS test: a method for the assessment of text-to-speech synthesis intelligibility. Speech Comm. 18, 381–392.

Betz, S., Carlmeyer, B., Wagner, P., Wrede, B., 2018. Interactive hesitation synthesis: modelling and evaluation. Technol. Interact. 2, 1.

Bijankhan, M., Sheykhzadegan, J., Bahrani, M., Ghayoomi, M., 2011. Lessons from building a Persian Written Corpus: Peykare. Lang. Resour. Eval. 45 (2), 143–164.

Bishop, C.M., 1994. Mixture Density Networks. Neural Computing Research Group, Aston University, Birmingham, UK. Tech. Rep. NCRG/94/004;.

Chen, L., Yang, H., Wang, H., 2018. Research on Dungan speech synthesis based on deep neural network. In: Proc. 11th International Symposium on Chinese Spoken Language Processing. pp. 46–50.

Choi, H., Park, S., Park, J., Hahn, M. 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.

Daneshfar, F., Barkhoda, W., Azami, B.Z., 2009. Implementation of a text-to-speech system for Kurdish language. In: 2009 Fourth International Conference on Digital Telecommunications. IEEE, July. Pp. 117–120.

Deng, L., Seltzer, M.L., Yu, D., Acero, A., Mohamed, A.R., Hinton, G., 2010. Binary coding of speech spectrograms using a deep auto-encoder. In: Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, Makuhari, Japan, 26–30 September. pp. 1692–1695.

Duffy, S.A., Pisoni, D., 1992. Comprehension of synthetic speech produced by rule: a review and theoretical interpretation. Lang. Speech 35, 351–389.

Donahue, J., Dieleman, S., Bińkowski, M., Elsen, E., Simonyan, K., 2021. End-to-end adversarial text-to-speech. arXiv preprint arXiv:2021.03575.

Dutoit, T.A., 1996. Short Introduction to Text-to-Speech Synthesis. TTS Research Team, TCTS Lab., Mons, Belgium. Available at: https://www.academia.edu/416871/A_Short_Introduction_to_Text_to_Speech_Synthesis?fbclid=IwAR1i0qc1SQjy1vNuOe9bIk2kPVp5TDuK2-uxdYKukv1-jV0_Cx7uVPVahR0 (accessed March 3, 2024).

Edlund, J., Tånnander, C., Gustafson, J., 2015. Audience response system-based assessment for analysis-by-synthesis. In: Proceedings of the 18th International Congress of the Phonetic Sciences (ICPhS 2015), Glasgow, UK.

Fahmy, F.K., Khalil, M.I., Abbas, H.M., 2020. A transfer learning end-to-end arabic text-to-speech (TTS) deep architecture. In: IAPR Workshop on Artificial Neural Networks in Pattern Recognition, September. Springer, Cham, pp. 266–277.

Fan, Y., Qian, Y., Xie, F.L., Soong, F.K., 2014. TTS Synthesis with bidirectional LSTM based recurrent neural networks. In: Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September.

Fellbaum, K., 2014. Anmerkungen zu den begriffen "Verständlichkeit" und "Verstehbarkeit" bei der Sprachqualitätsmessung. In: Elektronische Sprachsignalverarbeitung (ESSV), Tagungsband der 25. Konferenz, Dresden. pp. 240–247.

Gehring, J., Miao, Y., Metze, F., Waibel, A., 2013. Extracting deep bottleneck features using stacked auto-encoders. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May. pp. 3377–3381.

Goel, K., Gu, A., Donahue, C., R'e, C., 2022. It's Raw! Audio generation with state-space models. In: International Conference on Machine Learning.

Gopi, A., Sajini, T., Bhadran, V.K., 2013. Implementation of Malayalam text to speech using concatenative based TTS for android platform. In: Proc. Int. Conf. Control Commun. Comput., December.

Govender, A., King, S., 2018. Using pupillometry to measure the cognitive load of synthetic speech. In: Proc. Interspeech 2018, pp. 2838–2842.

Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM networks. In: Proceedings of the IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July–4 August. pp. 2047–2052.

Graves, A., Fernandez, S., Gomez, F., Schmidhuber, J., 2006. Connectionist Temporal classification: labeling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June. pp. 369–376.

Griffin, D.W., Lim, J.S., 1984. Signal estimation from modified short-time Fourier transform. IEEE Trans. Acoust. 32 (2), 236–243. https://doi.org/10.1109/TASSP.1984.1164317.

Gustafson, J., Boye, J., Fredriksson, M., Johanneson, L., Königsmann, J., 2005. Providing computer game character with conversational abilities. In: Panayiotopoulos, T.,

Gratch, J., Aylett, R., Ballin, D., Olivier, P., Rist, T. (Eds.), Intelligent Virtual Agents, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 37–51.
Hassani, H., Kareem, R., 2011. Kurdish Text to Speech (KTTS). In: Tenth International Workshop on Internationalisation of Products and Systems. Kuching Malaysia. pp. 79–89.
He, T., Zhao, W., Xu, L., 2020. DOP-Tacotron: a fast Chinese TTS system with local-based attention. In: 2020 Chinese Control Decis. Conf. (CCDC), August, pp. 4345–4350.
HMM/DNN-Based Speech Synthesis System (HTS). 2015. Available online: https://hts-engine.sourceforge.net/ (accessed on 13 Apr 2023).
Huang, X., Acero, A., Hon, H.-W., 2001. Spoken Language Processing: A Guide to Theory, Algorithm & System Development.
International Telecommunication Union, 2003. Subjective quality evaluation of telephone services based on spoken dialogue systems. ITU-R Recommendation ITU-P.851.
International Telecommunication Union, 2015. Method for the subjective assessment of intermediate quality level of audio systems. ITU-R Recommendation ITU-R.BS.1534-3.
Jang, W., Lim, D., Yoon, J., 2021. Universal melgan: a robust neural vocoder for high-fidelity waveform generation in multiple domains. arXiv preprint arXiv:2011.09631.
Jekosch, U. The cluster identification test (CLID). In: Proceedings of the International Conference on Spoken Language Processing (ICSLP '92), Banff, Alberta, Canada, 1992, pp. 205–208.
Kaneko, T., Kameoka, H., Hojo, N., Ijima, Y., Hiramatsu, K., Kashino, K., 2017. Generative adversarial network-based post-filter for statistical parametric speech synthesis. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, June. pp. 4910–4914.
Kayte, S., Mundada, M., Gujrathi, J., 2015. Hidden markov model based speech synthesis: a review. Int. J. Comput. Appl. 130 (3), 35–39.
Kayte, S., Waghmare, K., Gawali, B., 2015. Marathi speech synthesis: a review. Int. J. Recent Innov. Trends Comput. Commun. 3 (6), 3708–3711.
Khan, R.A., Chitode, J.S., 2016. Concatenative speech synthesis: a review. Int. J. Comput. Appl. 136 (3), 0975–8887.
Khanam, F., Munmun, F.A., Ritu, N.A., Saha, A.K., Mridha, M.F., 2022. Text to speech synthesis: a systematic review, deep learning based architecture and future research direction. J. Adv. Inf. Technol. 13 (5).
Klatt, D.H., 1987. Review of text-to-speech conversion for English. J. Acoust. Soc. Am. 82 (3), 737–793. https://doi.org/10.1121/1.395275.
Le Paine, T., Khorrami, P., Chang, S., Zhang, Y., Ramachandran, P., Hasegawa-Johnson, M.A., Huang, T.S., 2016. Fast Wavenet Generation Algorithm. doi: 10.48550/arXiv.1611.09482.
Lee, J., Cho, K., Hofmann, T., 2016. Fully character-level neural machine translation without explicit segmentation. arXiv preprint arXiv:1610.03017.
Lee, J.Y., Cheon, S.J., Choi, B.J., Kim, N.S., Song, E., 2018. Acoustic modeling using adversarially trained variational recurrent neural network for speech synthesis. In: Proc. INTERSPEECH. pp. 917–921.
Lavrakas, P., 2008. Encyclopedia of Survey Research Methods. Sage Publications, Inc, Thousand Oaks, CA, Chapter, p. 743.
Lemmetty, S., 1999. Review of Speech Synthesis Technology. Helsinki University of Technology. Master's Thesis.
Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M., 2019. Neural speech synthesis with transformer network. In: Proc. AAAI Conf. Artif. Intell., July, pp. 6706–6713.
Li, Y., Qin, D., Zhang, J., 2021. Speech synthesis method based on Tacotron2. In: Proc. 13th International Conference on Advanced Computational Intelligence. pp. 94–99.
Ling, Z.H., Deng, L., Yu, D., 2013. Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis. In: Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May. pp. 7825–7829.
Liu, R., Sisman, B., Bao, F., Gao, G., Li, H., 2020. Modeling prosodic phrasing with multi-task learning in tacotron-based TTS. IEEE Signal Process Lett. 27, 1470–1474.
Lopez, G., Quesada, L., Guerrero, L.A., 2018. Alexa vs. siri vs. Cortana vs. Google assistant: a comparison of speech-based natural user interfaces. Proc. Int. Conf. Appl. Hum. Factors Ergon. 241–250.
Luong, H.T., Takaki, S., Henter, G.E., Yamagishi, J., 2017. Adapting and controlling DNN-based speech synthesis using input codes. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, March. pp. 4905–4909.
Macon, M.W., 1996. Speech Synthesis Based on Sinusoidal Modeling. Master's Thesis. Georgia Institute of Technology.
Malisz, G., Henter, Z., Valentini-Botinhao, C., Watts, O., Beskow, J., Gustafson, J., 2019. Modern speech synthesis for phonetic sciences: a discussion and an evaluation. In: Proceedings of ICPhS 2019, Melbourne, Australia.
Mandal, S.K.D., Datta, A.K., 2007. Epoch synchronous nonoverlapadd (ESNOLA) method-based concatenative speech synthesis system for Bangla. SSW 351–355.
Mattheyses, W., Verhelst, W., Verhoeve, P., 2006. Robust pitch marking for prosodic modification of speech using TD-PSOLA. In: Proc. IEEE Benelux/DSP Valley Signal Process. Symp. SPS-DARTS, June, pp. 43-46.
Mehrish, A., Majumder, N., Bhardwaj, R., Mihalcea, R., Poria, S., 2023. A review of deep learning techniques for speech processing. arXiv:2305.00359[eess.AS]. doi: 10.48550/arXiv.2305.00359.
Meng, F.B., 2013. Analysis and Generation of Focus in Continuous Speech. Tsinghua University, Beijing, China. Ph.D. Thesis.
Möbius, B., Schroeter, J., Van Santen, J., Sproat, R., Olive, J., 1996. Recent advances in multilingual text-to-speech synthesis. Fort. Der Akustik 22, 82–85.
Naderi, N., Nasersharif, B., Nikoofard, A., 2022. Persian speech synthesis using enhanced Tacotron based on multi-resolution convolution layers and a convex optimization method. Multimed. Tools Appl. 81 (3), 3629–3645.
Nebez, J., 1993. The Kurdish language from oral tradition to written language. In: Conference of "The Kurdish Language Toward the Year 2000". Sorbonne University and the Kurdish Institute, Paris; and Western Kurdistan Association Publications, London.
Ning, Y., He, S., Wu, Z., Xing, C., Zhang, L.J., 2019. Review of deep learning based speech synthesis. Appl. Sci. (Switzerland) 9 (19), 4050. https://doi.org/10.3390/app9194050.
Norbert, S., Peeters, G., Lemouton, S., Manoury, P., Rodet, X. Synthesizing a choir in real-time using pitch synchronous overlap add (PSOLA). In: ICMC, 2000.
Onaolapo, J.O., Idachaba, F.E., Badejo, J., Odu, T., Adu, O.I., 2014. A simplified overview of text-to-speech synthesis. Proc. World Congr. Eng. 1, 5–7.
Oyucu, S., 2023. A novel end-to-end Turkish text-to-speech (TTS) system via deep learning. Electronics 12 (8), 1900. https://doi.org/10.3390/electronics12081900.
Oord, A.V.D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K, 2016. Wavenet: a generative model for raw audio. arXiv preprint arXiv:1609.03499.
Patton, B., Agiomyrgiannakis, Y., Terry, M., Wilson, K.W., Saurous, R.A., Sculley, D., 2016. Automos: learning a non-intrusive assessor of naturalness-of-speech. CoRR, abs/1611.09207. Available online: http://arxiv.org/abs/1611.09207 (accessed on March 3, 2024).
Ping, W., Peng, K., Chen, J., 2019. Clarinet: parallel wave generation in end-to-end text-to-speech. arXiv preprint arXiv:1807.07281.
Ping, W., Peng, K., Gibiansky, A., Arik, S.O., Kannan, A., Narang, S., Raiman, J., Miller, J., 2018. Deep voice 3: scaling text-to-speech with convolutional sequence learning. arXiv preprint arXiv:1710.07654.
Prenger, R., Valle, R., Catanzaro, B., 2018. WaveGlow: A Flow-based Generative Network for Speech Synthesis, 2018. doi: 10.48550/arXiv.1811.00002.
Rashad, M.Z., El-Bakry, H.M., Isma'il, I.R., Mastorakis, N., 2010. An overview of text-to-speech synthesis techniques. In: International Conference on Communications and Information Technology - Proceedings, pp. 84–89.
Rebai, I., BenAyed, Y., 2013. Arabic text to speech synthesis based on neural networks for MFCC estimation. In: Proc. World Congress on Computer and Information Technology, June. pp. 1–5.
Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y., 2019. Fastspeech: fast, robust and controllable text to speech. Adv. Neural Inf. Process. Syst. 32.
Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y., 2022. Fastspeech 2: fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558.
Saito, Y., Takamichi, S., Saruwatari, H., 2018. 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. pp. 5299–5303.
Sak, H., Gung, T., Safkan, Y., 2006. A corpus-based concatenative speech synthesis system for Turkish. Turk. J. Electr. Eng. Comput. Sci. 14 (2), 209–223.
Schoeffler, M., Bartoschek, S., Stöter, F.R., Roess, M., Westphal, S., Edler, B., Herre, J., 2018. webMUSHRA—a comprehensive framework for web-based listening tests. J. Open Res. Softw. 6 (1).
Shechtman, S., Mordechay, M., 2018. Emphatic speech prosody prediction with deep LSTM networks. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 5119–5123.
Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R.A., 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), April, pp. 4779–4783.
Sotelo, J., Mehri, S., Kumar, K., Santos, J.F., Kastner, K., Courville, A., Bengio, Y., 2017. Char2wav: end-to-end speech synthesis. In: ICLR2017 workshop submission.
Sun, L., Kang, S., Li, K., Meng, H., 2015. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. IEEE International Conference on Acoustics.
Takamichi, S., 2017. 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. pp. 1308–1311.
Toma, S.A., Tarsa, G.I., Oancea, E., Munteanu, D.P., Totir, F., Anton, L.A., 2010. TD-PSOLA based method for speech synthesis and compression. In: Proc. 8th Int. Conf. Commun., June, pp. 241–250.
Valentini-Botinhao, C., Wang, X., Takaki, S., Yamagishi, J., 2016. Investigating RNN-Based Speech Enhancement Methods for Noise-Robust Text-to-Speech. SSW.
Veisi, H., Mohammad Amini, M., Hosseini, H., 2019. Toward Kurdish language processing: experiments in collecting and processing the AsoSoft Text corpus. Dig. Schol. Hum. https://doi.org/10.1093/llc/fqy074.
Vainer, J., Dušek, O., 2020. Speedy speech: Efficient neural speech synthesis. arXiv preprint arXiv:2008.03802.
Verma, P., Chafe, C., 2021. A generative model for raw audio using transformer architectures. In: International Conference on Digital Audio Effects, 2021.
Voiers, A.S., et al., 1975. Research on diagnostic evaluation of speech intelligibility. Air Force Cambridge Research Laboratories, Bedford, Massachusetts, Tech. Rep. AFCRL-72-0694.
Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q, 2017. Tacotron: towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135.
Wang, W., Xu, S., Xu, B., 2016. First step towards end-to-end parametric TTS synthesis: generating spectral parameters with neural attention. In: Proceedings of the Seventeenth Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September. pp. 2243–2247.
Wang, X., Takaki, S., Yamagishi, J., 2017. An RNN-based quantized F0 model with multi-tier feedback links for text-to-speech synthesis. In: Proc. INTERSPEECH. pp. 1059–1063.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R.J., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., Saurous, R.A., 2018. Style tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis. J. Chem. Inf. Model. 58, 123–135. https://doi.org/10.48550/arXiv.1803.09017.

Weiss, R.J., Skerry-Ryan, R.J., Battenberg, E., Mariooryad, S., Kingma, D. P., 2021. Wave-Tacotron: spectrogram-free end-to-end text-to-speech synthesis. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, June. pp. 5679–5683.

Win, Y., Masada, T., 2020. Myanmar Text-to-Speech System based on Tacotron-2. In: 2020 Int. Conf. Inf. Commun. Technol. Converg. (ICTC), October, pp. 578–583.

Xu, S.H., 2007. Study on HMM-based Chinese speech synthesis. Beijing University of Posts and Telecommunications. Ph.D. Thesis.

Yang, S., Xie, L., Chen, X., Lou, X., Zhu, X., Huang, D., Li, H., 2017. Statistical parametric speech synthesis using generative adversarial networks under a multi-task learning framework. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop, December. pp. 685–691.

Zen, H., Senior, A., 2014. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, May. pp. 3844–3848.

Zen, H., Senior, A., 2014. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In: Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 4–9 May 2014. pp. 3844–4848.

Zen, H., Senior, A., 2014. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, May. pp. 3844–3848.

Zen, H., Senior, A., Schuster, M., 2013. Statistical parametric speech synthesis using deep neural networks. In: Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing, Vancouver, BC, Canada, 26–31 May. pp. 7962–7966.

Zena, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Comm. 51 (11), 1039–1064.

Zhang, C., Zhang, S., Zhong, H., 2019. A prosodic Mandarin text-to-speech system based on Tacotron. In: 2019 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC), November, pp. 165–169.

Zhuang, X., Huang, J., Potamianos, G., Hasegawa-Johnson, M., 2009. Acoustic fall detection using Gaussian mixture models and GMM supervectors. In: Proceedings of the 34th IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April. pp 69–72.