Current advances and algorithmic solutions in speech generation

Dina Oralbekova[1], Orken Mamyrbayev[2], Dinara Kassymova[3], Mohamed Othman[4]
[1, 2]Institute of Information and Computational Technologies, Almaty, Kazakhstan
[1, 3]Information and Communication Technologies Department, Academy of Logistics and Transport, Almaty, Kazakhstan
[4]Malaysia Department of Communication Technology and Networks, University Putra Malaysia, Serdang, Malaysia
[4]Laboratory of Computational Science and Mathematical Physics, Institute for Mathematical Research, University Putra Malaysia, Serdang, Malaysia
[1]Corresponding author
E-mail: [1]*dinaoral@mail.ru*, [2]*morkenj@mail.ru*, [3]*d.kassymova@alt.edu.kz*, [4]*mothman@upm.edu.my*

Received 18 January 2024; accepted 27 February 2024; published online 4 April 2024
DOI https://doi.org/10.21595/vp.2024.23940

68th International Conference on Vibroengineering in Almaty, Kazakhstan, April 4-6, 2024

Abstract. Currently, Text-to-Speech (TTS) technology, aimed at reproducing a natural human voice from text, is gaining increasing demand in natural language processing. Key criteria for evaluating the quality of synthesized sound include its clarity and naturalness, which largely depend on the accurate modeling of intonations using the acoustic model in the speech generation system. This paper presents fundamental methods such as concatenative and parametric speech synthesis, speech synthesis based on hidden Markov models, and deep learning approaches like end-to-end models for building the acoustic model. The article discusses metrics for evaluating the quality of synthesized voice. Brief overviews of modern text-to-speech architectures, such as WaveNet, Tacotron, and Deep Voice, applying deep learning and demonstrating quality ratings close to professionally recorded speech, are also provided.

Keywords: speech synthesis, speech recognition, neural networks, natural language processing, acoustic modeling, voice generation.

1. Introduction

Effective communication with computers or other smart devices relies not only on the stage of human speech recognition but also on providing clear responses in the form of familiar human voices for natural human-machine interaction. The ability to convert text into audio enhances the naturalness of communication between humans and machines, making it more accessible for individuals with limited capabilities in their daily lives. Speech synthesis has become an integral part of various domains, including accessibility technologies, education, entertainment, and business.

In this article, we aim to contribute to the advancement of speech synthesis technologies by exploring contemporary methods and algorithms and analyzing their application in the context of current trends and challenges. While there are numerous reviews of speech synthesis methods [1]-[3], this research work stands out for covering more recent advancements and providing a comprehensive analysis of their effectiveness.

Our work goes beyond existing state-of-the-art approaches in several ways. Firstly, we provide a detailed examination of key methods in the field of speech synthesis, including concatenative and parametric synthesis, as well as approaches employing hidden Markov models and deep learning. Specifically, we delve into the advancements made in deep learning techniques, such as end-to-end models for acoustic modeling, which have shown promising results in enhancing the quality of synthesized speech.

Furthermore, we offer discussions on modern text-to-speech architectures, including innovative approaches such as WaveNet, Tacotron, and Deep Voice. These architectures, leveraging deep learning methods, exhibit a level of quality comparable to professionally recorded speech and play a significant role in the contemporary landscape of speech synthesis technologies.

Our original contribution lies in synthesizing recent advancements in speech synthesis methods and providing insights into their practical applications and potential limitations.

2. Standard speech synthesis system

There are two main approaches to speech synthesis: concatenative, based on using pre-recorded sound fragments, and parametric, which models speech using parameterized models.

– Concatenative speech synthesis. Concatenative speech synthesis is a method based on the use of pre-recorded sound fragments (or units) to form the target speech sequence. These units can be phrases, words, or even individual sounds. To create natural-sounding speech, various units are selected and concatenated according to the required text. There are two different approaches to concatenative speech synthesis: one uses Linear Predictive Coding (LPC) [4], and the other is based on Pitch-Synchronous Overlap and Add (PSOLA). The main drawbacks of these approaches are outlined in [5].

– Parametric speech synthesis. Parametric speech synthesis is a method based on modeling various aspects of speech sound using parameterized models. These models control key speech characteristics such as pitch height, speech rate, and timbre. Common methods include statistical parametric speech synthesis [6], speech synthesis based on Hidden Markov Models (HMM) [7], and speech synthesis using Deep Neural Networks (DNN) [8].

The main advantages and limitations of these approaches are presented in Table 1.

Table 1. Main advantages and limitations of concatenative and parametric synthesis

Method	Advantages	Restrictions
Concatenative speech synthesis	– Concatenative synthesis is capable of providing more natural and realistic speech sounds as it utilizes recordings of natural human voices. – By using recorded fragments with different intonations and emotional nuances, concatenative synthesis can convey the richness of melodic features. – Recorded units allow for easy control over the pronunciation of specific words and phrases.	– The use of a large number of audio units requires significant resources for data storage and processing. – Complexity in adding new words or phrases to the system, as it necessitates recording new audio fragments
Parametric speech synthesis	– Parametric synthesis may consume fewer resources compared to concatenative synthesis as it relies on modeling speech aspects rather than storing audio fragments. – The parametric synthesis model can be easily adapted to include new words and phrases without the need to record new audio fragments.	– There may be limitations in achieving a high level of naturalness, especially in cases of complex intonations and emotional expression. – Parametric synthesis may be less suitable for precise control over the pronunciation of specific words and phrases.

Basic speech synthesis algorithm:

1. Text analysis and normalization. At this stage, input text to be synthesized is processed. The text can be a phrase, sentence, or even an entire passage. Data transformation involves converting various abbreviations and numbers into their textual representations.

2. Converting text into linguistic structure. In this stage, the text is analyzed, and for each phrase, the phonemes to be used are determined, as well as how each word should be pronounced in a given context. This involves working with linguistic rules and dictionaries.

3. Generation of speech units (phonemes, diphones, triphones, etc.). Based on the obtained

linguistic characteristics, the corresponding speech units are generated.

4. Speech synthesis. At this stage, the actual speech synthesis process occurs. In concatenative systems, speech units are concatenated to form complete speech.

5. Processing and quality improvement. The synthesized audio signal can undergo additional processing to improve its quality.

3. Speech synthesis methods

3.1. Statistical parametric speech synthesis

Speech synthesis is based on customizable parameters, defined by linguistic and prosodic rules. These parameters encompass intonation, stress, phoneme duration, speech rate, and can vary depending on the overall mood and emotional state of the individual at the current moment. The process also involves acoustic processing, where prosodic parameters are transformed into a speech signal using vocal cords and the vocal tract. All these stages form the foundation for building a speech synthesis system.

The key advantages of Statistical Parametric Speech Synthesis (SPSS) include high flexibility, quality, and relatively modest resource requirements for stable operation [9]. The flexibility of this method provides more efficient control over the intonation in synthesized speech, and the quality of the resulting speech using SPSS approaches natural sound.

Statistical Parametric Speech Synthesis includes the following components (Fig. 1): phonetic processing, extraction of linguistic and prosodic features, acoustic processing, and a vocoder.

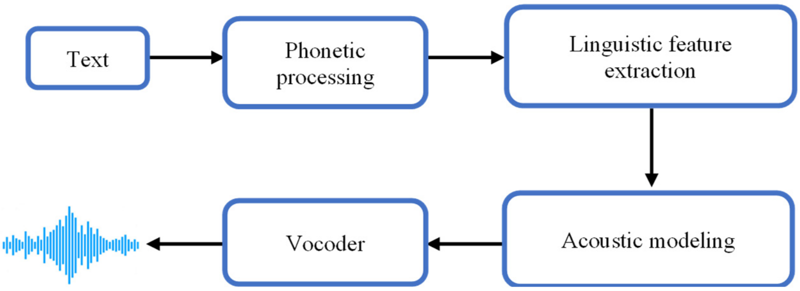

Fig. 1. Basic architecture of TTS technology

During phonetic processing, a phonetic vector is created for each word. In the acoustic processing stage, acoustic parameters are created for each phoneme fragment. After determining the duration of each phoneme, it is divided into windows, each of which can have a duration ranging from 15 ms to 40 ms. Then, acoustic feature vectors are generated for each window. In the vocoder stage, acoustic features are transformed into the audio signal of speech. The quality of modern speech synthesis systems depends on the completeness and correctness of selected prosodic and linguistic features influencing speech, for a specific language, and, of course, on the quality and accuracy of the annotation of the speech corpus.

3.2. Speech generation based on Hidden Markov Models

Speech synthesis based on Hidden Markov Models (HMM) is a method that involves modeling the temporal sequence of speech parameters using HMM. This process includes three main components: the state model, the observation model, and the alignment algorithm [10].

In the state model, the speech signal is described as a sequence of states, where each state represents a hidden state that changes over time. Transitions between states are determined by probabilistic transitions. Each hidden state is associated with an observable vector representing speech parameters, such as spectral coefficients. The observation model defines the probability distribution of these observations given a particular state.

The alignment algorithm determines the best match between observations and hidden states. The primary goal of speech synthesis based on HMM is to train the model parameters to maximize the probability of observed speech parameters given the hidden states. For each text fragment, HMM generates a sequence of hidden states, which is then used to synthesize the corresponding speech signal.

3.3. Speech generation using deep learning and end-to-end models

Speech generation using deep learning represents an advanced method in the field of speech synthesis, relying on the effective application of deep neural networks (DNN) to model and reproduce speech characteristics. Various deep learning architectures, such as recurrent neural networks (RNN), long short-term memory (LSTM), convolutional neural networks (CNN), and transformers, are employed within the context of speech generation [11]-[12]. These architectures are capable of extracting complex temporal and spatial dependencies in speech data. Deep models require the representation of text in a numerical form. Commonly used methods include word embeddings, providing contextual understanding of words in a sentence.

For model training, annotated data representing text-speech pairs are essential. The model is trained based on this data with the goal of minimizing the difference between synthesized and real audio:

$$min_\theta \sum_i MSE(s_i, \hat{s}_i), \tag{1}$$

where s_i – real audio signal, $\hat{s}_i$ – synthesized audio signal, θ – model parameters, MSE is typically performed using Mean Squared Error.

End-to-end models have the capability to directly synthesize speech signals from text after being trained on extensive datasets. This approach streamlines the process of statistical parametric synthesis or deep learning-based synthesis, replacing a complex pipeline of multiple modules with a single component in the form of a neural network.

3.4. Architectures of modern text-to-speech systems

WaveNet represents an innovative audio generation model utilizing a fully convolutional and autoregressive approach. This model can create compelling audio samples, taking into account various linguistic characteristics and synchronizing them with the original audio. WaveNet is capable of producing high-quality audio independent of input conditions.

The model itself is built on sequences of dilated convolutions with a progressively expanding receptive field for better data processing. It operates within a fully probabilistic and autoregressive approach, predicting the distribution for each audio signal based on previous data. It's essential to note that due to the structure of the sequential autoregressive process, WaveNet can be somewhat slow, limiting its efficient use in various sound synthesis applications.

Unlike some traditional systems, Tacotron does not require pre-recorded voice segments from a speaker. It is trained on text-audio pairs, enabling it to generate speech for arbitrary text. The main component of Tacotron is a recurrent neural network that learns the correspondence between textual input and audio output. The model takes a text sequence as input and generates a mel-spectrogram representing audio parameters. These parameters are then transformed into the time domain to obtain the final audio signal. Tacotron can be complemented with other architectures, such as WaveNet, to enhance speech generation quality.

Tacotron 2 has a more complex architecture than its predecessor. It consists of three main components: an encoder, a decoder, and a PostNet. During the encoding stage, text is transformed into a sequence of numerical symbols, which are then input into an Embedding layer. This layer creates high-dimensional vectors (512-dimensional) representing each symbol. The vectors then pass through a block of three one-dimensional convolutional layers. Each layer uses 512 filters of length 5, allowing for the consideration of the current symbol and its neighbors. The results of the convolution undergo normalization and ReLU activation. Then, the data passes through bidirectional LSTM layers with 256 neurons each, where the results of forward and backward passes are combined. After this, the decoder operates with a recurrent architecture, generating one mel-spectrogram frame at each step. A key element here is the mechanism of soft attention, providing flexibility and control over the generation process. The PostNet serves to enhance the generated spectrogram, crucial for improving audio output quality. The combination of these elements in Tacotron 2 allows for the creation of natural and high-quality speech, emphasizing the significance of soft attention in the synthesis process.

Similar to Tacotron, Deep Voice also consists of an encoder, a decoder, and an attention mechanism. In the encoder, text is transformed into input representations for the neural network, involving the conversion of characters into numerical vectors and subsequent transformation into mel-spectrograms representing sound information. The decoder is responsible for generating the audio signal based on input representations obtained from the encoder. Typically, the decoder operates in a recurrent architecture, where each time step produces a new audio fragment. The attention mechanism is used to better align the text and sound patterns, allowing the model to focus on different parts of the input text during audio fragment generation.

DeepVoice 2 utilizes a more complex architecture, including the use of Long Short-Term Memory (LSTM) and multi-layer convolutional networks for text processing. DeepVoice 2 has been significantly improved by incorporating a built-in multi-speaker function into the speech synthesizer, enabling the model to mimic hundreds of different voices.

DeepVoice 3 is a multi-system where segmentation, duration, and frequency blocks have been removed. Instead, a single model generates the mel-spectrogram or other audio features, which can then be decoded into an audio signal using WaveNet or another vocoder. A unique feature of DeepVoice 3 is its fully convolutional architecture, meaning the use of convolutional layers without recurrent neural networks.

4. Metrics for evaluating synthesized speech

The most common method for assessing the quality of generated audio is through the Mean Opinion Score (MOS). MOS is employed to measure human perception regarding the quality of an audio signal, specifically synthetic speech in this context. MOS involves surveying experts or native speakers, asking them to rate the quality on a scale, usually ranging from 1 to 5, where higher scores indicate better quality.

During the MOS evaluation, experts or users listen to audio segments and assign scores based on their perception of quality. The arithmetic mean of these scores forms the MOS.

It's important to note that MOS is a subjective assessment, and different experts may give different ratings based on their perception and preferences. Nevertheless, using the average MOS helps establish an overall assessment of synthesized speech quality.

Another method is the objective speech quality assessment algorithm, based on automatic metrics comparing the synthesized audio signal with the original. One widely used algorithm in this field is PESQ (Perceptual Evaluation of Speech Quality). This algorithm also seeks expert ratings on a scale from 1 to 5.

5. Key challenges requiring improvements in speech generation

Despite ongoing efforts to enhance models and introduce new effective algorithms, there are challenges and issues in speech synthesis that remain unresolved or require further refinement.

1) Intonation and Emotional Expression.

Even modern speech synthesis systems may struggle to achieve complete naturalness and convey emotional nuances accurately. Enhancing models' ability to not only reproduce accurate sounds but also convey appropriate intonation, emotional color, and expressiveness in speech.

2) Speed and Efficiency Enhancement.

Speech synthesis should be fast and efficient to ensure smooth and instantaneous speech generation in various applications like voice assistants. Developing algorithms that operate at high speeds, making them applicable for real-time synthesis in online applications.

3) Handling Different Languages and Accents.

Issues persist in synthesizing speech for languages with relatively limited resources, including Turkic language families.

Enhancements in these areas would contribute to higher quality and universality in speech synthesis systems, crucial for diverse applications.

6. Conclusions

This article has explored various speech synthesis methods, including concatenative, parametric, and hidden Markov model-based approaches, as well as deep learning-based approaches. Deep learning models such as WaveNet, Tacotron, and Deep Voice present promising architectures for text-to-speech conversion, delivering quality ratings close to professionally recorded speech.

The assessment of synthesized voice quality has emerged as a critical aspect of Text-to-Speech technology advancement. Metrics such as clarity, naturalness, and precise intonation modeling serve as benchmarks for refining speech generation systems.

As part of our future work, we aim to focus on enhancing the quality of synthesis for agglutinative languages, such as Kazakh, because of agglutinative languages pose unique challenges due to their complex morphological structure, requiring specialized techniques for accurate synthesis.

Acknowledgements

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19174298).

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Z. Mu, X. Yang, and Y. Dong, “Review of end-to-end speech synthesis technology based on deep learning,” *arXiv:abs/2104.09995*, Apr. 2021.

[2] O. Nazir and A. Malik, “Deep learning end to end speech synthesis: a review,” 2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC), IEEE, 2021.

[3] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, “A review of deep learning based speech synthesis,” *Applied Sciences*, Vol. 9, No. 19, p. 4050, Sep. 2019, https://doi.org/10.3390/app9194050

[4] N. Kaur and P. Singh, “Conventional and contemporary approaches used in text to speech synthesis: a review,” *Artificial Intelligence Review*, Vol. 56, No. 7, pp. 5837–5880, Nov. 2022, https://doi.org/10.1007/s10462-022-10315-0

[5] B. S. Atal and S. L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” *The Journal of the Acoustical Society of America*, Vol. 50, No. 2B, pp. 637–655, Aug. 1971, https://doi.org/10.1121/1.1912679

[6] E. Cataldo, F. R. Leta, J. Lucero, and L. Nicolato, “Synthesis of voiced sounds using low-dimensional models of the vocal cords and time-varying subglottal pressure,” *Mechanics Research Communications*, Vol. 33, No. 2, pp. 250–260, Mar. 2006, https://doi.org/10.1016/j.mechrescom.2005.05.007

[7] H. Zen et al., “The HMM-based speech synthesis system (HTS) version 2.0,” in *ISCA Workshop on Speech Synthesis*, 2007.

[8] F. B. Meng, “Analysis and generation of focus in continuous speech,” Tsinghua University, Beijing, China, 2013.

[9] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, “A style control technique for HMM-based expressive speech synthesis,” *IEICE Transactions on Information and Systems*, Vol. E90-D, No. 9, pp. 1406–1413, Sep. 2007, https://doi.org/10.1093/ietisy/e90-d.9.1406

[10] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, “Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005,” *IEICE Transactions on Information and Systems*, Vol. E90-D, No. 1, pp. 325–333, Jan. 2007, https://doi.org/10.1093/ietisy/e90-1.1.325

[11] H. Zen, K. Tokuda, and T. Kitamura, “An introduction of trajectory model into HMM-based speech synthesis,” in *5th ISCA Speech Synthesis Workshop*, Jan. 2004.

[12] O. Mamyrbayev, A. Kydyrbekova, K. Alimhan, D. Oralbekova, B. Zhumazhanov, and B. Nuranbayeva, “Development of security systems using DNN and i and x-vector classifiers,” *Eastern-European Journal of Enterprise Technologies*, Vol. 4, No. 9(112), pp. 32–45, Aug. 2021, https://doi.org/10.15587/1729-4061.2021.239186