Speech Synthesis Technology: Status and Challenges

Caiyue Chen*

Electrical and Computer Engineering, College of Engineering, University of Massachusetts Amherst, Amherst, Massachusetts 01003, United States of America

Abstract. In recent years, speech synthesis technology has been more widely used in the field of artificial intelligence and human-computer interaction due to the excess of machine learning models to deep learning models. With the rise and development of applications such as intelligent voice assistants, voice navigation systems, generative macro modelling, and virtual reality, Users' demand for voice systems is not limited to the generated sound as cold as robots and full of "inhuman" tones and rhythm, it is also desired to generate speech that is more natural, fluent, and free of mechanical sensations. This paper reviews the recent development of speech synthesis techniques and the interpretation of each step in the order of speech synthesis steps, two parts of acoustic modeling and wave pattern synthesis are described in detail. In addition, this article aims to introduce some typical Speech Synthesis technology and also summarizes the current applications and future prospects in the field of speech synthesis research.

1 Introduction

Speech synthesis technology is a branch of natural language processing, with its earliest history going back to the early 20th century. The "Voder" device, invented by humans in 1930 using mechanical and electronic devices, was a preliminary exploration of speech synthesis. By the end of the 20th century, with the progress of computing power, speech synthesis technology has also been developed accordingly. Since 1990, statistical models based on Hidden Markov Models (HMM) and Gaussian Hybrid Models (GMM) have become popular, representing that Text-to-Speech (TTS) technology has entered the stage of statistical parameter synthesis. Although the synthesis method based on statistical parameters improves the overall quality of synthesized speech compared to earlier methods, it is still insufficient in terms of tone naturalness. Since the 21st century, with the development of computer science, artificial intelligence, and other disciplines, data-driven machine learning methods have begun to occupy a dominant position. Especially with the introduction of deep learning, they have brought great progress to the field of TTS.

Due to the long history of speech synthesis and the emergence of various methods and models at every step of speech synthesis. In order to review the speech synthesis technology in recent years, this article provides a detailed introduction and classification of acoustic

* Corresponding author: caiyuechen@umass.edu

modeling and waveform synthesis in Chapter 2. The acoustic modeling part is mainly divided into three models based on parametric speech synthesis (SPSS), end-to-end, and generative adversarial networks (GANs), and each one introduces its typical models. For the waveform synthesis part, introduced by the autoregressive model WaveNet, two other optimization models, GANs-based vocoder and small size vocoder, are described to solve the problems of slow training and generation and unnatural synthesis of speech. More details about acoustic modeling and waveform synthesis will be introduced in Chapter 2. Moreover, because of the diversity of speech datasets and the complexity of speech synthesis problems, this article provides a detailed introduction in Chapter 3 to the mainstream speech datasets currently used for model training, as well as evaluation metrics used to assess the quality of synthesized speech. In addition, the current application of TTS is mainly summarized in Chapter 4, and future works and related works will be listed in Chapter 5.

2 Steps in speech synthesis

In this section, this paper reviews the stages of speech synthesis, but is limited to the length of the article, focusing on acoustic modeling and waveform synthesis are expanded upon below. Speech synthesis generally includes four parts, they are text analysis, acoustic modeling, waveform synthesis and audio output. Fig. 1 is a process of speech synthesis.

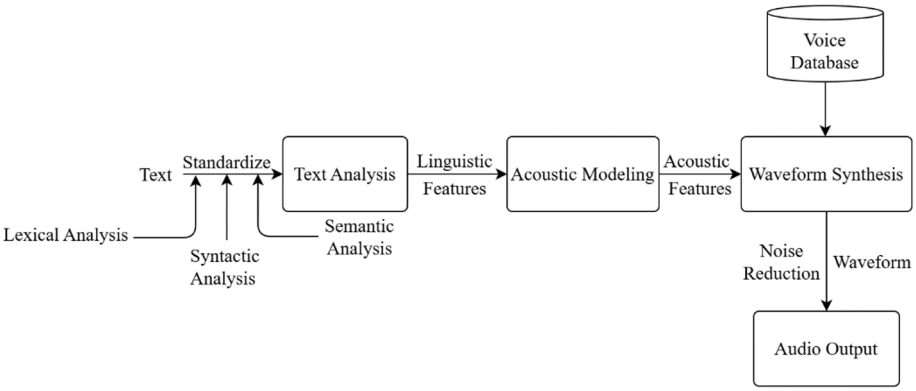

Fig. 1. A Process of Speech Synthesis (Photo/Picture credit: Original).

2.1 Text analysis

Before text analysis, texts have to be “standardized” to the original text. This step includes text normalization, punctuation processing, and the conversion of numbers, date, and time. After prior work, texts are transformed into linguistic features that contain rich phonetic and rhythmic information. Table 1 summarizes some general tasks in text analysis.

Table 1. Common Tasks in Text Analysis.

Task	Description
Text Normalization	Converts incoming non-standardized text (numbers, abbreviations, symbols, etc.) into a standard form for speech systems.
Word Segmentation	Divide a continuous stream of text into individual words or meaningful units.
Part-of-Speech (POS) Tagging	Label each word in the text with the corresponding lexical gender (noun, verb, etc.)
Prosody Prediction	Predict characteristics such as rhythm, stress, and intonation for each word or phoneme based on context, syntax, and semantics.
Grapheme to Phoneme (G2P)	Converts written words (graphemes) into corresponding sounds (phonemes).
Polyphone Disambiguation	Determine the correct pronunciation of words with multiple possible pronunciations based on context (more common in Chinese).

2.2 Acoustic modeling

Acoustic modeling is the process of transforming the text's linguistic features and prosodic information into acoustic features, which are then used to generate the speech waveform. In the following, this paper will introduce the acoustic model in statistical parametric SPSS, End-to-End TTS and other acoustic models respectively.

2.2.1 Acoustic model in SPSS

In SPSS, generating acoustic features using the HMM statistical model is the most common way [1]. The HMM model sees speech as a Markov chain of hidden states, each of which generates a specific sound feature. By training the HMM, the system can learn the probability distribution of each phoneme or sub-phoneme, and then generate speech waveforms or speech signals recognizing the input based on these distributions. Although HMM performs well in processing time series data (such as speech) and can cope with variation and noise in speech, HMM assumes that states are independent and uses simple Gaussian distribution to describe timbre characteristics, which cannot capture complex sound patterns [2]. Fig. 2 illustrates the workflow of the HMM.

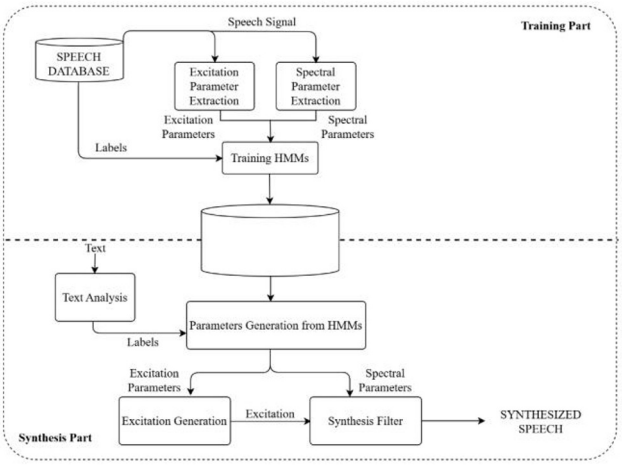

Fig. 2. HMM Workflow Diagram [3].

2.2.2 Acoustic models in end-to-end TTS

End-to-end models typically include encoders, decoders, and attention mechanisms. The encoder encodes input text or speech features into hidden representations, and the decoder generates speech features or directly generates speech waveforms based on these hidden representations. The attention mechanism is used to align the input text with the generated speech features. When dealing with long sentences or complex text, the End-to-End model is usually able to generate more natural speech, but because the training and reasoning process of the end-to-end model requires a lot of computational resources and a lot of labeled data to train, the performance will be significantly degraded when the data is insufficient. Table 2 introduces three common acoustic models in End-to-End TTS.

Table 2. Common Acoustic Models in End-to-End TTS.

End-to-End TTS model	Model Name	Description
RNN-based Models	Tacotron Series [4]	Tacotron utilizes an encoder-attention decoder framework that takes characters as input and outputs a linear spectrogram; Compared to the Tactron, Tacotron 2 offers significant improvements in connectivity TTS, parametric TTS, and neural TTS [5].
CNN-base Models	DeepVoice Series [6]	DeepVoice augmented with a convolutional neural network to obtain linguistic features; DeepVoice3 enhances DeepVoice with improved network architecture and multi-speaker modeling.
Transformer-based Models	FastSpeech Series [7]	FastSpeech utilizes an explicit duration predictor to extend the phoneme hidden sequence to match the length of the melting spectrogram; Compared to FastSpeech, FastSpeech2 alleviates the text-to-speech one-to-many mapping problem by providing more information about the differences, such as pitch, duration, and energy.

2.2.3 Formatting the title acoustic models in GANs

In addition to the acoustic models above, GANs are also widely used in speech synthesis. GANs consist of a generator and a discriminator. The generator is responsible for generating realistic speech waveforms from noise, while the discriminator is responsible for determining whether the generated speech waveforms are similar to real speech. Through adversarial training between the generator and discriminator, the generator will gradually generate high-quality speech. However, the training process of GANs is usually unstable and prone to pattern collapse or training failure. Table 3 lists some typical acoustic models in GANs.

Table 3. A list of typical acoustic models in GANs.

Models	Year	Features
WaveGAN	2018	Suitable for processing 1D audio data.
Parallel WaveGAN	2019	Lightweight and efficient for real-time applications.
MelGAN	2019	Fast and capable of parallel processing.
HiFi-GAN	2020	High fidelity with multi-scale and multi-subband discriminators.
VocGAN	/	Adaptable to different acoustic tasks.

2.3 Waveform synthesis

In 2016, Oord et al. proposed the most widely used vocoder to date [8]. WaveNet is a neural network-based autoregressive model. Since it uses a multilayer extended causally gated convolutional network, its network structure is very complex, and its training and inference are slow due to the large amount of time required for autoregressive waveform generation. In order to refine and optimize the shortcomings of WaveNet, the mainstream approach is mainly to speed up training and inference. This section describes some waveform synthesis models that have been optimized on top of WaveNet. Table 4 shows these models and their details.

Non-Autoregressive Vocoder [9]. To compensate for the slow training and inference speed of WaveNet, it has been found that the generation speed can be significantly improved by parallelizing the processing of speech waveforms. Because there is no need to rely on the previously generated speech sample to generate subsequent samples, and the ability to generate multiple samples at the same time, more and more non-autoregressive models are starting to appear.

Small Size Vocoder [9]. In order to speed up the generation of waveforms, the simplest way is to use the Relu function and a low-dimensional convolution. For example, a multi-scale RNN structure is adopted, which is based on the principle of processing audio data at different times in each layer, and at the same time, weighted pruning or quantization techniques can be used to further narrow down the parameters of the model in order to increase the training speed.

GANs-based Vocoder [10]. Since sometimes the speech produced by WaveNet can be unnatural, in order to enhance the naturalness and smoothness of synthesized speech, GANs with generators and discriminators have been used for adversarial training of waveforms. And because the waveforms are generated by implicit generative models such as GANs, speech waveforms with different resolutions can be predicted at the same time, which ensures the waveform frequency and waveform details of the synthesized speech.

Table 4. Kinds of Vocoder.

Models	Types	Input	Parallel
WaveNet	/	Mel Spectrogram	N
WaveRNN	Small Size	Mel Spectrogram	N
FftNet	Small Size	Mel Spectrogram	N
GAN-TTS	Non-Autoregressive, GANs-based	Mel Spectrogram, Text	Y
MelGAN	Non-Autoregressive, GANs-based	Mel Spectrogram	Y
LPCNet	Small Size	Mel Spectrogram, Linear Prediction Coefficients (LPC)	N
Multi-Band WaveRNN	Small Size	Mel Spectrogram	N
DiffWave	Non-Autoregressive	Mel Spectrogram, Diffusion Noise	Y
WaveGrad	Non-Autoregressive	Mel Spectrogram, Diffusion Noise	Y
WaveFlow	Non-Autoregressive	Mel Spectrogram	Y
WaveVAE	Non-Autoregressive	Mel Spectrogram	Y
Parallel WaveGAN	Non-Autoregressive, GANs-based	Mel Spectrogram	Y
HiFi-GAN	Non-Autoregressive, GANs-based	Mel Spectrogram	Y
VocGAN	Non-Autoregressive, GANs-based	Mel Spectrogram	Y
Multi-Band MelGAN	Non-Autoregressive, GANs-based	Mel Spectrogram	Y

3 Datasets and evaluation metrics

Current speech synthesis research relies on a variety of open-source datasets that cover different languages, multiple speakers, and partial emotion labeling. Commonly used datasets include LJSpeech, VCTK, and LibriSpeech, which are suitable for multi-speaker speech synthesis but generally lack emotion labeling. Datasets such as IEMOCAP and EmoV-DB, while providing emotion labeling, but have limited sentiment categories and small data volumes. Chinese datasets such as CSMSC and AIShell-3 are widely used in Mandarin speech synthesis, but also lack emotion labeling. In addition, specialized datasets such as Korean Emotional Speech (KES) and Japanese Kamishibai and Audiobook Corpus (J-KAC) provide language-specific affective or storytelling speech data, but with limited diversity and scale. Although these datasets have provided important resources for speech synthesis research and facilitated the development of cross-linguistic and multi-speaker models, however, they suffer from several limitations such as scarce emotion annotation, insufficient data diversity, scale limitations, and uneven quality. These limitations restrict the research progress and model generalization capabilities of affective speech synthesis, especially in cross-cultural and multi-contextual applications. Therefore, future speech synthesis research still needs to make improvements in dataset diversity, labeling accuracy, and scale to support more complex and natural speech synthesis tasks. Table 5 lists the major open sources databases used in the papers covered in this review.

Subjective evaluation metrics are metrics that assess the performance of a speech synthesis system and the quality of the generated speech through direct feedback from human listeners. These metrics rely on human perception and therefore capture subtle differences that are difficult for machines to quantify, such as speech naturalness, emotional expression, and overall listening experience. Table 6 lists the common subjective metrics applied for evaluating TTS models' performance.

Objective evaluation metrics are metrics that quantify and assess the quality of speech generated by speech synthesis systems through automated methods. These metrics do not rely on the subjective perception of the human listener, but rather use specific algorithms and formulas to analyze and compare the characteristics of speech signals to provide consistent and repeatable evaluation results. Table 7 lists the mainly metrics for evaluating TTS models' performance.

Table 5. List of main open source expressive.

Database	Year	Language	Data Size (h)	Multi Speaker
Blizzard Challenge	2018~2023	English	/	
IEMOCAP	2008	English	12	√
LibriSpeech	2015	English	1000	√
VCTK	2016	English	44	√
LJSpeech	2017	English	24	
CHEAVD2.0	2018	Mandarin Chinese	8	√
MELD	2018	English	13	√
CMU-MOSEI	2018	English	66	√
SEWA	2019	Multilingual	44	√
Chinese Standard Mandarin Speech Copus (CSMSC)	2019	Mandarin Chinese	12	
LibriTTS	2019	English	585	√
CH-SIMS	2020	Mandarin Chinese	2.5	√
MSP-Conversation	2020	English	15.15	√
Multilingual LibriSpeech	2020	Multilingual	50000	√
Aishell-3	2021	Mandarin Chinese	85	√
Emotional Speech Dataset (ESD)	2021	English/Mandarin Chinese	175	√
M3ED	2022	Mandarin Chinese	50	
English conversation corpus (ECC)	2022	English	93	√

Table 6. Subjective evaluation metrics for expressive speech synthesis models.

Metric	Description	Advantages	Disadvantages
MCD	Measures the squared differences between MFCCs of ground truth and synthesized samples.	Simple to compute, quantifies spectral differences.	Lacks perceptual relevance.
GPE	Percentage of frames where pitch deviates from reference by more than 20%.	Clearly identifies major pitch errors.	Ignores minor pitch variations.
VDE	Percentage of frames with incorrect voiced/unvoiced decision compared to reference.	Simple metric for voicing accuracy.	Does not account for context or perceptual impact.
FFE	Percentage of frames where pitch or voicing differs from the reference by more than 20%.	Captures both pitch and voicing errors.	Does not differentiate error types or perceptual significance.
WER	Percentage of substitutions, deletions, and insertions between reference and hypothesis transcripts.	Clear metric for speech recognition accuracy.	Ignores small errors that may not affect understanding.
BAPD	Measures differences in band aperiodicity between synthesized and ground truth speech.	Quantifies accuracy of noise and unvoiced sound modeling.	Ignores temporal dynamics and speech context.
RMSE	Square root of average squared differences between predicted and observed values.	Measures prediction accuracy.	Does not indicate whether errors are overestimations or underestimations.

Table 7. Objective evaluation metrics for expressive speech synthesis models.

Metric	Description	Advantages	Disadvantages
MOS	Five-point subjective ratings of naturalness and fluency of synthesized speech by human listeners.	Reflects human perception of audio quality.	Costly and subjective.
CMOS	Measures perceived quality difference between two samples on a scale from -3 to +3.	Direct, sensitive assessment of quality differences.	Influenced by individual biases.
DMOS	Ratings on a five-point scale for perceived degradation compared to a reference sample.	Provides average subjective degradation scores.	Inconsistent across different listener groups.
AB Preference Test	Listeners choose between two samples, A and B, or indicate no preference.	Useful for comparing two systems or algorithms.	Requires many participants for statistical significance.
ABX Preference Test	Listeners identify whether sample X matches A or B, with X being the target.	Double-blind, good for evaluating subtle differences.	Focuses on discernment, not subjective preference.
MUSHRA	Listeners rate multiple samples, including a reference and an anchor, on a percentage scale.	Fine-grained quality assessments, captures subtle differences.	Test complexity can lead to listener fatigue.

4 Future works

Future speech synthesis technology will be optimized in several directions to further enhance its naturalness, emotional expression, diversity and adaptability. For example, enhanced emotional expression, fine-grained emotional control, and natural emotional transitions will make synthesized speech more humanized [11]; improved acoustic modeling and rhyme generation techniques will further enhance the naturalness and fluency of speech; and combining multimodal information, such as the fusion of vision and speech, will provide a more immersive user experience. Personalized and customized speech generation will provide more personalized interactions based on user preferences, while continuously optimizing through adaptive learning. Multi-language and cross-cultural adaptability will enable the system to better support users from different linguistic and cultural backgrounds. At the same time, technologies such as model compression and acceleration, cloud-side co-computing, etc. will optimize computational efficiency and enable speech synthesis to run efficiently even on resource-constrained devices.

5 Conclusions

The existing applications of speech synthesis technology have greatly improved the human-computer interaction experience. In the field of intelligent voice assistants and smart homes, voice synthesis technology enables Apple Siri, Amazon Alexa and Chat-GPT to respond to user commands and queries with natural and smooth voice. In addition, in the automatic response system of the shopping platform, the IVR system (interactive voice response) is widely used in the automatic chatbot, which can basically achieve the automatic voice response, thus improving the efficiency of and serving customers. Speech synthesis technology is also being used in the education industry, such as Duolingo's speech teaching feature, which helps students learn different languages and dialects in a personalized way. The main intention of this paper is to introduce the development history of the field of speech synthesis and the major synthesis techniques in recent years. This article focuses on the modeling methods of acoustic modeling and waveform synthesis according to the process of speech synthesis. At the same time, it also provides a clear introduction to the evaluation indicators of synthesized speech and the development direction of future work.

References

1. Y. Tabet, M. Boughazi, Speech synthesis techniques. A survey, in Proceedings of the International Workshop on Systems, Signal Processing and their Applications, IEEE, (2011), 67-70
2. K. Tokuda, H. Zen, A.W. Black, An introduction of trajectory model into HMM-based speech synthesis, in Proceedings of the ISCA SSW5, (2004)
3. K. Tokuda, H. Zen, A.-W. Black, Hidden semi-Markov model based speech synthesis, in Proceedings of the Interspeech, (2004), 1185-1180
4. Y. Wang, R.J. Skerry-Ryan, D. Stanton, et al., Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135 (2017)
5. J. Shen, R. Pang, R.J. Weiss, et al., Natural TTS synthesis by conditioning Wavenet on mel spectrogram predictions, in Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, (2018), 4779-4783
6. W. Ping, K. Peng, A. Gibiansky, et al., Deep Voice 3: Scaling text-to-speech with convolutional sequence learning. arXiv preprint arXiv:1710.07654 (2017)

7. Y. Ren, Y. Ruan, X. Tan, et al., FastSpeech: Fast, robust and controllable text to speech. Adv. Neural Inf. Process. Syst. **32**, (2019)
8. A.-V. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
9. Z. Mu, X. Yang, Y. Dong, Review of end-to-end speech synthesis technology based on deep learning. arXiv preprint arXiv:2104.09995 (2021)
10. X. Tan, T. Qin, F. Soong, A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561 (2021)
11. C.C. Lee, T. Chaspari, E.-M. Provost, An engineering view on emotions and speech: From analysis and predictive models to responsible human-centered applications. Proc. IEEE. **111**, 1142-1158 (2023)