Assignment 1

Topic: Detailed analysis of speech synthesis (Question 1).
Date: 20-01-2025
Submitted By: Arjun Arora (M24CSA003, Prateek (M24CSA022))

---

1 Objective

Perform a detailed analysis on speech synthesis covering the following aspects: Prepare a detailed report and a short presentation by comparing and briefly describing the SOTA models available. These should include the following:

- Explain the task and its importance in the real world.
- Analyze the strengths and limitations of state-of-the-art models or tools in terms of the methods or models available.
- Discuss the results in terms of the metrics used to evaluate the task, including their strengths and limitations.
- Suggest what are the open problems and opportunities corresponding to that problem statement.
- Submission should include the presentation, report, and the codes/datasets you have worked with for this question.

2 Speech Synthesis

Speech synthesis is a technology that converts written text into spoken words using artificial intelligence and computational linguistics. It enables machines to produce human-like speech, which has several applications across various domains and offers substantial real-world benefits. Speech synthesis can support accessibility for the visually impaired, assist in providing accurate weather or navigation updates, and enhance customer service through automated voice assistants. Studies indicate that computer-generated speech can aid individuals with learning disabilities, helping them better comprehend written text. Students can instruct speech synthesis programs to read specific words, entire lines, or complete text selections. The immediate speech feedback enables learners to identify and correct reading mistakes, which in turn enhances their learning process [1].

In addition, speech synthesis plays a crucial role in enhancing the learning abilities of individuals with dyslexia. A study by Atkar et al. (2024) highlighted how deep learning-based speech synthesis techniques are being used to support people with dyslexia, ultimately aiding in their educational progress [2]. The application of speech synthesis extends beyond education. In autonomous systems, such as self-driving cars, delivery robots, and smart home devices, clear and natural communication is essential. The ability to generate articulate and expressive speech enhances the interaction between humans and machines, fostering trust and improving the user experience. Kuo and Tsai (2024) explored how synthesizing more expressive speech can make interactions between humans and robots more engaging and natural [3].

3 State of the Art Models

In this section, we explore several significant advancements in speech synthesis and text-to-speech (TTS) systems, highlighting the strengths, limitations, and innovations of each model. The models discussed here range from early efforts in waveform generation to state-of-the-art solutions in non-autoregressive speech synthesis.

3.1 WaveNet

WaveNet is a groundbreaking deep generative model introduced by Van Den Oord et al. in 2016. Unlike traditional TTS systems that rely on concatenative or parametric synthesis, WaveNet generates raw audio waveforms using a probabilistic autoregressive approach. This enables the model to capture fine-grained details of speech, resulting in highly natural and expressive synthetic voices. One of its key strengths lies in its ability to generate diverse audio characteristics such as different speakers, emotions, and languages. Additionally, WaveNet has shown versatility beyond TTS, including applications in music generation and audio compression. [4]

The model faces challenges particularly in its computational complexity. The high inference cost associated with WaveNet's autoregressive nature makes it impractical for real-time applications. Future improvements could focus on enhancing efficiency without compromising audio quality. The study used VCTK and LJSpeech for general speech synthesis, and Google's internal datasets for commercial applications. To assess the performance of WaveNets in the TTS task mean opinion score (MOS) tests were performed. In the MOS tests, participants rated the naturalness of each stimulus on a five-point Likert scale (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent).In the MOS test, each stimulus was shown to the participants individually. In the paired comparison test, eight subjects evaluated each pair, while in the MOS test, eight subjects rated each stimulus. The participants, who were paid native speakers, performed the task.

3.2 Tacotron

Tacotron, proposed by Wang et al. in 2017, revolutionized TTS by offering an end-to-end system that synthesizes speech directly from text. Unlike traditional systems that require manual feature extraction, Tacotron generates a mel-spectrogram, which can then be converted into raw audio. One of its major advantages is its ability to be trained from scratch with random initialization, removing the need for hand-engineered features. Additionally, Tacotron's frame-based synthesis method is faster compared to sample-level autoregressive models like WaveNet.[5]

Despite its success, Tacotron is not without its drawbacks. The Griffin-Lim algorithm used for waveform inversion can produce audible artifacts, and several parts of the model, such as the output layer and attention module, are still in need of refinement. Tactron was trained on LJSpeech (single speaker, English) and Google TTS datasets. Tacotron 2 was also trained on a combination of LJSpeech and proprietary datasets. The study also conducted mean opinion score tests where participants rated the naturalness of the stimuli. The tests were crowdsourced from native speakers. A total of 100 unseen phrases were used, with each phrase receiving 8 ratings. Only ratings from tests where headphones were used were included in the MOS calculation. The model was compared with a parametric system based on LSTM and a concatenative system, both of which are in production.

3.3 HMM-driven Concatenative Synthesis

The HMM-driven concatenative speech synthesis model, as presented by Gonzalvo et al., relies on unit selection, where speech units are pre-recorded and concatenated based on a probabilistic model. Improvements in this system have focused on reducing latency and maintaining high-quality output, especially when handling large databases in real-time applications. A novel voice-building strategy was proposed to reduce building time without sacrificing quality.[6].This model used CMU Arctic, Blizzard Challenge datasets, and VCTK for building unit-selection and statistical parametric models. Concatenative systems suffer from limitations, including the need for large datasets and the difficulty of capturing the full variability of human speech. While latency has been reduced, the system still struggles to generate expressive and dynamic speech. The mean opinion score (MOS) tests were conducted with a specific experimental setup. A total of 173 test utterances were used, with each subject evaluating up to 30 stimuli. Each pair was assessed by three subjects in the MOS tests, and headphones were used by all participants.

3.4 Deep Voice 3

Deep Voice 3, introduced by Ping et al., is a fully-convolutional neural TTS model that outperforms its predecessors in both naturalness and training speed. By scaling up to a large dataset with over eight

hundred hours of audio and two thousand speakers,[7] it demonstrates robust multi-speaker capabilities. The model uses a convolutional architecture with an attention mechanism and can be trained efficiently. Deep Voice 3 is also agnostic to the waveform synthesis method, supporting Griffin-Lim, WaveNet, and WORLD vocoder.Deep voice 3 got trained on LJSpeech, VCTK, and Blizzard Challenge datasets. It also supports multi-speaker datasets like LibriTTS. One limitation of Deep Voice 3 is that it still faces challenges related to generating the full variability of human speech and accents. Further work is needed to refine its grapheme-to-phoneme model and explore more diverse datasets.

The models were trained using the VCTK and LibriSpeech datasets. Ground-truth samples were intentionally included in the evaluation set, as the accents in these datasets may be unfamiliar to the North American crowdsourced raters. Batches of samples from these models were presented to the raters via Mechanical Turk.

3.5 FastSpeech 2

FastSpeech 2[8], introduced by Ren et al., improves upon the original FastSpeech model by addressing several limitations related to duration prediction and training complexity. Unlike its predecessor, FastSpeech 2 directly uses ground-truth data, reducing the information loss that was present when using teacher-student distillation. It introduces additional input features such as pitch and energy to improve the accuracy of speech synthesis, and it can generate speech waveforms directly, providing faster inference times. It Used LJSpeech, LibriTTS, and AISHELL-3 (Chinese TTS dataset), along with internal datasets for improving prosody and expressiveness.

The model's reliance on external alignment tools for pitch extraction and other tasks remains a challenge for full end-to-end training, though the results have shown significant improvements in voice quality and synthesis speed. To assess the perceptual quality, a mean opinion score (MOS) evaluation was conducted on the test set. Twenty native English speakers were asked to evaluate the quality of the synthesized speech samples. The text content remained consistent across different systems, ensuring that the testers focused solely on the audio quality, without other influencing factors, and compared the results to the ground truth.

3.6 VITS2

VITS2 represents the next evolution in single-stage TTS models[9], offering significant improvements in the naturalness and computational efficiency of speech synthesis. It builds upon the earlier version of VITS, addressing limitations such as intermittent unnaturalness and phoneme conversion dependency. The model introduces enhanced structures that allow it to generate high-quality speech more efficiently than its predecessors. The model was trained on LJSpeech, VCTK, LibriTTS, and JSUT (Japanese TTS dataset). VITS models benefit from large-scale multi-speaker datasets for robust synthesis.

Though VITS2 shows promise, it still faces challenges in fully eliminating unnaturalness and improving computational efficiency for real-time applications. Further development is needed to address these aspects while ensuring high-quality speech synthesis. To verify that the proposed model synthesizes natural speech, crowdsourced mean opinion score (MOS) tests were conducted. Raters assessed the naturalness of the audio samples on a 5-point scale (1 to 5) after listening to randomly selected samples from the test sets. Since previous work has already shown similar quality to human recordings, a comparative mean opinion score (CMOS) test was also performed, which is suited for evaluating high-quality samples through direct comparison. Raters indicated their relative preference in terms of naturalness on a 7-point scale (-3 to 3) after listening to randomly selected audio samples. Each raters evaluated each sample only once. To eliminate the impact of amplitude differences on the scores, all audio samples were normalized.

4 Performance Metrics

There are various objective and subjective performance metrics used for evaluating speech synthesis. Objective evaluation metrics offer automated and reproducible assessments like Mel-Cepstral Distortion (MCD), which quantifies the spectral distance between synthesized and reference speech. This metric reflects the degree to which the generated audio aligns with the target in terms of acoustic features.

For assessing intelligibility, the Word Error Rate (WER) is another metrics that can be employed for comparing the transcriptions of synthesized speech to the input text using automated speech recognition. Another widely-used objective metric is the Perceptual Evaluation of Speech Quality (PESQ) [10], which evaluates speech quality by comparing degraded audio with a clean reference. PESQ models human auditory perception and provides a score between 1 and 4.5, reflecting intelligibility and distortion under various conditions, including noise or compression.

4.1 Mean Opinion Score (MOS)

Mean Opinion Score (MOS) is a key metric in the field of telecommunications and Quality of Experience (QoE), used to measure the perceived quality of synthesized speech.[11] All od the SOTA models discussed in this report are evaluated based on this metric. MOS is the average rating given by a set of listeners, typically on a scale from 1 to 5, where higher scores correspond to better quality. MOS captures human perception of speech quality, although it requires substantial resources for large-scale evaluations. MOS can be represented mathematically as:

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N}$$

where R_n is the rating given by the n^{th} listener, and N is the total number of listeners.

MOS is commonly used for evaluating all state-of-the-art (SOTA) models in speech synthesis due to its ability to reflect overall quality as perceived by humans. Despite its usefulness, MOS is subject to some issues, such as variability in individual listener preferences and the resources required for large-scale evaluations. Nonetheless, MOS remains the standard for subjective evaluation in the industry due to its straightforwardness and direct correlation with human perception.

Rating	Label
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 1: MOS Rating Scale

5 Comparison of Different State-of-the-Art Models

As discussed in the previous sections, we examined six different state-of-the-art (SOTA) models developed by researchers worldwide, including contributions from R&D teams of major technology companies. In all these studies, the Mean Opinion Score (MOS), a reference-based metric, was used to evaluate synthesized speech quality. Note that the MOS scores were reported in comparison to their respective reference audio samples and specifications of their experimental setup which are discussed in the papers cited.

Model	Language	MOS Score
WaveNet	North American English	4.21 ± 0.081
Tacotron	North American English	3.82 ± 0.085
HMM-driven concatenative	North American English	3.86 ± 0.137
Deep Voice 3 (WaveNet)	North American English	3.78 ± 0.30
FastSpeech 2	English	3.71 ± 0.09
VITS2	English	4.47 ± 0.06

Table 2: Comparison of MOS scores for different SOTA models. The reported values are based on the specifications of their respective studies.

6 Open Problems and Opportunities

Many state-of-the-art (SOTA) models primarily rely on the Mean Opinion Score (MOS) for evaluating speech synthesis quality. However, as synthetic speech increasingly resembles human speech, the reliability of MOS as a metric is being questioned. A recent study has highlighted its limitations, emphasizing the need to refine MOS or develop novel evaluation metrics to ensure more robust and consistent assessments [12].

Another open problem in speech synthesis is the lack of precise definitions for key terminologies such as "expressive," "emotion," "prosody," and "style." The absence of standardized definitions leads to inconsistencies across research efforts, making it difficult to compare different approaches. Establishing clear and universally accepted definitions would help unify research efforts and improve reproducibility [13].

Beyond evaluation and terminology, there are several critical challenges in advancing speech synthesis. Developing models that are highly efficient, robust, and adaptable to low-resource environments remains a significant research direction. Current speech synthesis systems also face ethical and social concerns, particularly in their potential misuse for deepfakes, misinformation, and privacy violations. Addressing these issues requires not only technical solutions but also policy frameworks to ensure responsible deployment.

The field faces challenges in context-aware speech synthesis, where models must adapt their output based on conversational context, speaker intent, and emotions. Another major challenge is cross-lingual and accent adaptation, as current models struggle with generating high-quality speech for low-resource languages and diverse accents. Lastly, personalized speech synthesis, which aims to generate speech tailored to individual user preferences, remains an ongoing research problem. Advancing in these areas will be crucial for the next generation of speech synthesis systems.

References

[1] K. Forgrave, "Assistive technology: Empowering students with learning disabilities," *The Clearing House*, vol. 75, no. 3, pp. 122–126, 2002.

[2] G. Atkar, A. Gaikwad, S. More, S. Hatte, and M. Naimuddinquadri, "Deep learning-based speech synthesis for enhancing the learning ability of individuals with dyslexia," in *2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pp. 1–4, IEEE, August 2024.

[3] Y. Kuo and P. Tsai, "Enhancing expressiveness of synthesized speech in human-robot interaction: An exploration of voice conversion-based methods," in *2024 10th International Conference on Control, Automation and Robotics (ICCAR)*, pp. 1–4, IEEE, April 2024.

[4] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[5] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, and Q. Le, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[6] X. Gonzalvo, S. Tazari, C. Chan, M. Becker, A. Gutkin, and H. Silen, "Recent advances in google real-time hmm-driven unit selection synthesizer," *Google Research*, 2016.

[7] W. Ping, K. Peng, A. Gibiansky, S. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[8] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[9] J. Kong, J. Park, B. Kim, J. Kim, D. Kong, and S. Kim, "Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design," *arXiv preprint arXiv:2307.16430*, 2023.

[10] T. Xie, Y. Rong, P. Zhang, and L. Liu, "Towards controllable speech synthesis in the era of large language models: A survey," *arXiv preprint arXiv:2412.06602*, 2024.

[11] Wikipedia contributors, "Mean opinion score — Wikipedia, The Free Encyclopedia," 2024. [Online; accessed 2-February-2025].

[12] S. Le Maguer, S. King, and N. Harte, "The limits of the mean opinion score for speech synthesis evaluation," *Computer Speech & Language*, vol. 84, p. 101577, 2024.

[13] H. Barakat, O. Turk, and C. Demiroglu, "Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 11, 2024.