Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/2023310

Analysis of the survey of voice synthesis technology

Zhangyao Zhou

Oregon State University, 1500 SW Jefferson Way,Corvallis, OR 97331

zhouzha@oregonstate.edu

Abstract. The original purpose of speech synthesis was to complete the task of converting from text to speech (TTS). With the application of deep learning models in this field of speech synthesis, the results of speech synthesis have gradually reached the level of human voice, which makes it widely used for voice assistant, navigation, reading, intelligent customer service, and many other aspects. In order to help readers expand more research ideas and understand the development process of speech synthesis technology and the ethical choices faced, this article introduces as much as possible in the development process of speech synthesis technology, several frameworks for optimization, as well as interpreting their respective advantages and disadvantages, summarize their approximate results obtained in the MOS test method and analyze the ethical issues that may be brought about by voice cloning technology.

Keywords: Voice Synthesis, Voice Clone, Speech Generation, Speech Synthesis Models.

1. Introduction

One study showed that composite speech scores were positively correlated with agreeableness and neuroticism in participants, with humans preferring real human voices over synthesized ones, but manipulating diagnostic speech features may increase acceptance of synthesized voices, supporting Human-Computer Interaction [1]. This means that humans are very sensitive to sound characteristics, so with the goal that humans cannot distinguish between synthetic voices and real human voices, the model used to generate speech becomes particularly important. The performances of the generation of a speech synthesis model, including prosody, timbre, naturalness, emotion, etc., largely determine the degree of human satisfaction with it.

This paper focuses on the currently popular voice synthesis models of WaveNet, Parallel WaveNet, Tacotron, Tacotron 2, DurIAN, Transformer, FastSpeech, FastSpeech, and analyzes their characteristics, which will help those who are new to the field of voice synthesis to quickly understand this field. Various speech synthesis models will be further researched with interest and appropriate models will be applied in future researches.

2. Application of technology

With the breakthrough of speech synthesis technology, the TTS system has been able to synthesize speech quickly and accurately, which makes it suitable for daily use. Speech synthesis technology is mainly used in the field of human-computer interaction and helping people with disabilities.

DOI: 10.54254/2755-2721/4/2023310

2.1.Human-computer interaction

Nowadays, smartphones (iPhone, Samsung, etc.), smart speakers (Google Home, etc.), and navigation devices have become popular, and many of them are equipped with virtual personal assistants that use speech synthesis to help interact with users. The representative one is Siri, users can ask Siri to answer or complete various things by talking to Siri, and Siri's voice comes from speech synthesis technology [2]. In addition, speech synthesis technology is also applied to AI customer service, which makes the voice of AI customer service more natural and smooth, so that users can get better service.

2.2.Helping people with disabilities

For people who have lost their voice due to some accident or illness, speech synthesis technology can help them regain a voice similar to their original voice [3], they can also choose their favorite voices on the VocaliD website (collecting the voices donated by people in a crowdsourced way), and these voices can be made to sound with emotion through the adjustment of the algorithm. For blind people, speech synthesis can help them read books and other reading materials, which greatly improves their reading efficiency.

3. Spech synthesis strategies background

There are two traditional speech synthesis strategies, namely cascade waveform synthesis and the statistical parameter method. The cascaded waveform synthesis method generates new speech by concatenating pre-recorded speech fragments in a large database. The resulting speech is of high quality but lacks the rhythm and timbre characteristics of the sound. Another method is the statistical parameter method, which provides a new voice by adjusting various voice parameters of the recorded human voice. The voice generated by this method is smoother and more natural, but the voice quality is poor. In recent years, the development of deep learning technology has been applied to the field of speech synthesis, and new breakthroughs have been made in speech synthesis technology. The following will summarize and introduce various speech synthesis models brought by deep learning.

4. Common technical models applied in speech synthesis

4.1.WaveNet

WaveNet [4], which is an early waveform-based deep neural network model capable of generating subjective natural raw audio signals, is based on the PixelCNN architecture, which enables it to analyze thousands of random variables distribution is modeled. WaveNet combines causal filters and dilated convolutions to complete training in order to deal with long-range temporal dependencies needed for raw audio generation. At the same time, CNN requires less time to train than RNN, and the use of dilated convolutions also reduces the cost of training. Expensive computational cost.

In the first test, without text conditioning, WaveNet was able to model speech from any speaker by conditioning on the speaker's one-hot vector dataset, capturing characteristics of each speaker, producing realistic intonation, but for longer sentences, it lacks coherence and can only remember the last 2-3 phonemes produced. In the second test, when it is applied to generate speech from the text (TTS), conditioned on language features, WaveNet can synthesize speech samples with natural segment quality, but due to the receptive field of the WaveNet is not good enough to capture this long-term dependency, and it sometimes has an unnatural sound when it emphasizes the wrong word in a sentence.

4.2.Parallel waveNet

Due to the excellent performance of WaveNet in the quality of real speech synthesis, people developed Parallel WaveNet [5] based on it using Probability Density Distillation for high-fidelity speech synthesis. Since the samples of WaveNet's convolutional structure are essentially generated using the previous samples, the generation speed is slow and cannot be processed in parallel. Parallel WaveNet accelerates speech generation by three orders of magnitude without losing quality by using a new form of neural network distillation, Probability Density Distillation, that combines the properties of WaveNet's efficient training and the efficient sampling of IAF networks to allow parallel processing.

DOI: 10.54254/2755-2721/4/2023310

4.3.Tacotron

Tacotron [6] is also an early TTS model. In order to solve the sequence-to-sequence prediction problem, it uses a recurrent neural network Encoder-Decoder as an architecture, which consists of an encoder, an attention-based decoder, and a post-processing network. Among them, the encoder is responsible for reading English sentences word by word and encoding the input sequence as an internal state vector for pre-training, and the CBHG module acts as a pre-processing net to extract valuable features from the pre-trained output to help improve the generalization ability of the model, the decoder can predict that multiple non-overlapping mel-spectrogram frames correspond to one output letter from the encoder. The number of mel-spectrogram frames is inversely proportional to the number of decoder steps required by the model to produce an output of the same length, thus reducing the complexity of the model, after which the attention mechanism will help align the input text with the predictions. Finally, the prediction results are passed to the post-processing network through the CBHG module as a post-processing net to synthesize audio through the Griffin-Lim algorithm.

Compared with WaveNet, Tacotron belongs to audio synthesis, and its generation speed is faster, but since it does not require phoneme-level alignment, it can be found by the Griffin-Lim algorithm without messing up the left and right adjacent amplitude spectrum and its own amplitude spectrum. The approximate phase of the input spectrogram is used to reconstruct the wave, which is generated with some noise and false articulations.

4.4.Tacotron 2

Based on the excellent performance of WaveNet and Tacotron in the field of speech synthesis, Google Brain proposed the Tacotron 2 framework in 2017. Tacotron 2 [7] combines the methods of Tacotron and WaveNet, which consists of a Tacotron-style model of Seq2seq, intermediate connection layers (Mel-wave spectrograms), and a modified WaveNet vocoder. The way Tacotron 2 works is to use vanilla LSTM instead of Tacotron's RNN model, first using three convolutional layers to perform feature extraction (preprocessing) on the text through the LSTM model and attention mechanism, and then input the features into the model and generate Mei. The Mel-wave spectrogram is then input into the modified WaveNet vocoder waveform to synthesize speech. Compared to the original Tacotron, the Tacotron 2 model uses simpler building blocks, replacing the "CBHG" stack and GRU recurrent layers with vanilla LSTMs and convolutional layers in the encoder and decoder, and in doing so overcomes the vanishing gradients of RNNs and short-term memory problem thereby improving the quality of the generated speech, more say, each decoder step of Tacotron 2 will correspond to a single spectrogram frame, rather than a decoder predicting multiple non-overlapping Mel spectrogram frames. In addition, Tacotron 2 modified WaveNet to generate time-domain waveform samples in order to predict Mel spectral frames. As a result, Tacotron 2 has become a speech synthesis system that combines the prosody of Tacotron-generated audio and the quality advantages of WaveNet-generated audio. It is currently one of the best speech synthesis systems in the field of speech synthesis. Its drawback is that it cannot generate audio in real-time, the voice lacks emotion, and in particular cases, the pronunciation of words is difficult and accompanied by murmurs.

4.5.DurIAN

DurIAN [8] developed by Tencent AI Lab is a multi-modal synthesis system of highly natural speech and facial expressions. DurIAN is similar in architecture to Tacotron. It consists of an encoder, an alignment model that aligns the input phoneme sequence with the target acoustic frame at the frame level, a decoder, and a post-processing network. DurIAN's encoder is the same as that used in Tacotron 1, the post-processing network is exactly the same as that used in Tacotron 2, and the decoder is similar to that used in Tacotron 1. However, its biggest innovation over Tacotron is to replace the attention context mechanism with a simple and powerful alignment model.

To align text letters and Mel spectrogram frames, end-to-end systems typically use an attention mechanism to help with this task. However, existing end-to-end attention mechanisms are not stable and often produce unpredictable artifacts (unwanted sounds), and some words are skipped or repeated in the generated speech. The alignment model predicts the duration of each phoneme through a separate phoneme duration model and then infers the alignment between the phoneme sequence and the target acoustic sequence based on the duration of each phoneme, which minimizes alignment

DOI: 10.54254/2755-2721/4/2023310

errors. In addition, a fine-grained style control algorithm with supervised style labels is proposed in DurIAN to achieve fine-grained control of spoken style. The learned style embedding is represented as a vector in the latent space concatenated with the corresponding phoneme to control their style. As a result, when the DurIAN system is used to synthesize speech, the naturalness and quality of the generated results are comparable to those of Tacotron 2, solving the possible problems of artifacts, word skipping and repetition in the generated speech, and also allowing the generation of speech Perform style control to show the voice style of neutral, happy, exciting, angry.

4.6.Transformer

Transformer [9] was originally used for language translation, it uses an encoder-decoder structure with an attention mechanism and contains a fully connected Feedforward network. Later, because of its excellent performance was applied to end-to-end speech synthesis, Transformer TTS [10] combines the advantages of Tacotron2 and Transformer, it uses parallel training in Transformer The multi-head attention mechanism replaces the RNN in Tacotron2. Compared with bidirectional RNN, the multi-head attention mechanism splits attention into multiple ones, making it possible to establish long-term dependencies between frames in many different aspects, and each attention considers the context of all sequences globally. This can better help align phoneme sequences and target acoustic frames, which solves the problem of word skipping and repetition that Tacotron2 may have when sentences are too long. In addition, employing multi-head attention can also improve the training speed through parallel computing.

4.7.FastSpeech & fastspeech 2

FastSpeech [11], like DurIAN, focuses its research on solving the generated speech containing unpredictable artifacts (unwanted sounds) due to mel-spectrograms, where some words are skipped or repeated in the generated speech Stability issue. FastSpeech is developed on the basis of Transformer TTS, which ensures the alignment of text letters and mel-spectrogram frames through a phoneme duration predictor, which is the same alignment model used by DurIAN, and interestingly, both coincide in the same The alignment problem was solved using the same method. It also uses a length modifier to increase the controllability of synthesized speech by extending or shortening phoneme durations and adding breaks between adjacent phonemes. In addition to improving the robustness and controllability of synthesized speech, the highlight of FastSpeech is that it greatly improves the speed of speech generation by generating mel-spectrograms in parallel.

As a result, the Transformer TTS network can speed up training by ~4.25 times compared to Tacotron 2 [10], while FastSpeech accelerates mel-spectrogram generation by 269.40 compared to Transformer TTS model times [11]. If both use WaveGlow as a vocoder, FastSpeech can achieve a 38.30x speedup in audio generation with similar voice quality. However, its disadvantage is that the training process is complicated and time-consuming, the phoneme durations obtained from the teacher model are inaccurate and the simplified output of the teacher model leads to the loss of some mel-spectrogram frames, which is optimized by FastSpeech 2 [12], which directly makes The ground training model is used to replace the simplified output of the teacher model to simplify the training process and introduce more variables (pitch, energy), which improves the accuracy of phoneme duration while improving the training speed, achieving 3 times more than FastSpeech training speed boost and better voice quality.

5. Model evaluation method

The most widely used method for evaluating the quality of the results generated by different speech synthesis methods is the mean opinion score (MOS), which is done manually. The person evaluating the speech will rate the audio, and eventually, the mean of all the scores will be taken as the MOS result. The average of the MOS results, with scores ranging from 1 to 5 (1 being the lowest quality and 5 being the highest quality), is approximately between 4.5 and 4.8 for true human audio quality.

Figure 1 shows the MOS scores obtained by the speech synthesis method mentioned in this paper, however, it is worth noting that the speech quality evaluation of the MOS method is subject to the subjectivity of the assessor and there is no accurate evaluation standard. The scores in the chart are

DOI: 10.54254/2755-2721/4/2023310

from different papers, so the score in the graph differs from how it actually performs, and the difference in actual performance between different synthesis methods differs from the difference in scores between them. In fact, to reduce this discrepancy, crowdMOS [13] engages people through crowdsourcing Speech MOS assessment, and has been tested, it provides accurate and repeatable results.

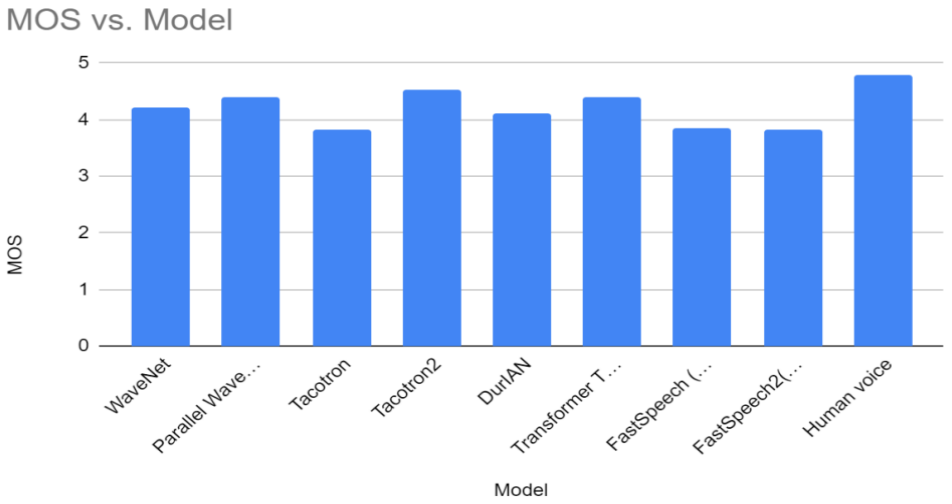

Figure 1. The MOS scores obtained by the speech synthesis method.

6. Ethical issues brought by new technologies

With the development of technology, people are faced with the ethical issues brought about by new technologies. Speech synthesis technology is breaking through and getting closer to one of the ultimate goals of all artificial intelligence technology, making it impossible for humans to distinguish between human agents and the artificial intelligence with which they are interacting. However, due to the deceptive nature of the generated results, it is easy to be used in unethical ways.

6.1.Violate personal privacy

A person's appearance and voice are part of personal privacy, and artificial intelligence technology makes it possible to imitate a person's appearance and voice characteristics. It can generate a Deepfake video by using other people's appearance and voice data so that you can make almost the same person in the video do anything you want including scenes with insults, pornography, violence, etc. There have been many Deepfake videos with celebrity faces and voices on the Internet. Although most DeepFake videos are created for the purpose of spoofing, some videos are still likely to involve immoral behaviors that violate the privacy of others. However, there are only a few local governments around the world that clearly stipulate that the use of others without their permission will be used against others. It is against the law to clone people's appearance and voice.

6.2.Fraud issues

In addition, there have been many cases of fraudulent use of voice clones. Criminals gain the trust of victims by forging Deepfake voices or videos of people they know and then defraud people's property under false identities. Since speech and video carry more recognizable features of a person than text, and in the past, it was nearly impossible to perfectly replicate so many features of a person, only limited people realized that creating an almost identical clone in voice or video is no longer so far away, which makes it easy for others to believe that these Deepfakes are real.

6.3.Misuse of the speech cloning algorithm

Due to concerns about the misuse of the speech cloning algorithm, and the fact that it is increasingly difficult to identify accurately, deception detection systems have been developed [14] to help identify

DOI: 10.54254/2755-2721/4/2023310

malicious attacks using speech synthesis techniques to protect speech-based authentication systems. A wide variety of methods are available for speech synthesis. The detection system uses the ASVspoof 2019 dataset to train and evaluate the model in order to be suitable for detecting speech signals generated by various speech synthesis techniques. ASVspoof 2019 includes 17 text-to-speech types. (TTS) and Voice Conversion (VC) technologies [15]. This system detects target audio using a classic supervised learning pipeline, first using a source-filter model to extract a set of features from the audio, Afterwards, the audio is classified into real speech and unknown classification (synthetic speech) according to the extracted features by training a supervised classifier.

7. Conclusion

This paper introduces the application of speech synthesis technology in people's lives and summarizes several popular deep learning speech synthesis models in the process of speech synthesis technology development. At present, Tacotron 2 is the most advanced speech synthesis model, and a large number of models have been developed and researched on its basis. The generated results have excellent speech quality and prosody and have obtained a MOS score close to the human voice.

However, speech synthesis technology still faces many challenges. If you want to truly make the synthesized voice sound like a human voice, you need to let the voice carry emotions, not just the elements of appearance such as timbre. If you want to complete this challenge It requires a large amount of data such as emotion recognition, automatic emotion annotation, and tone adjustment. However, the current emotional expressions that voice clones can perform are very limited and not ideal, and they are all obtained by adjusting certain parameters of the generated audio.

In addition, communication between humans is usually fast and natural, and people tend to give timely feedback after receiving messages. In order to have a smooth conversation with the user, the voice of the AI assistant needs to be synthesized quickly, which has certain requirements for the speech synthesis model.

Currently, Tacotron 2, Durian, and FastSpeech are able to generate high-quality and natural speech, these models usually rely on a large amount of training data from a single speaker, but in the real world each speaking There is only one or a few samples of the human head for cloning, so if you want to clone the speaking style of the target speaker and apply it in the real world, you need to let the model use few samples to complete the training. Currently, few-shot voice cloning has achieved quite good performance, but one-shot voice cloning is still an open problem [16]. The widely circulated 5s clone [17] uses the SV2TTS deep learning framework. It can clone speech from 5 seconds of sample audio, however, the speech it generates lacks naturalness, is mechanical, and is clearly different from real human voices.

In terms of synthetic speech detection, due to the wide variety of synthetic speech generation methods, it is still challenging to accurately detect some families of synthetic speech tracks in open-set scenarios.

Despite the challenges, due to the development of artificial intelligence technology, it is foreseeable that speech synthesis will be widely used in human-computer interaction in the future. As the same as natural language processing, its potential is huge and needs to be developed by people.

Acnkowledgement

I would like to thank my advisor Miss Alisa Wang for her encouragement and support, who helped me adjust the structure of my paper and corrected some small mistakes in detail. When I almost gave up writing, the encouragement from her and my parents was what motivated me to continue my research.

References

[1] K. Kühne, Fischer MH and Zhou Y, The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study. Front. Neurorobot. 14:593732, 2020. doi: 10.3389/fnbot.2020.593732.

[2] Siri Team. (n.d.). Deep learning for siri's voice: On-device deep mixture density networks for hybrid unit selection synthesis. Apple Machine Learning Research. Retrieved April 5, 2022,

DOI: 10.54254/2755-2721/4/2023310

from https://machinelearning.apple.com/research/siri-voices.

[3] J. D. Gray, (n.d.). On a mission to help people sound like themselves. The ASHA Leader. Retrieved April 5, 2022, from https://leader.pubs.asha.org/doi/full/10.1044/leader.LML.24072019.28.

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

[5] A. Oord, Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., ... & Hassabis, D. Parallel wavenet: Fast high-fidelity speech synthesis. In International conference on machine learning (pp. 3918-3926, July, 2018. PMLR.

[6] Y. Wang, Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. Tacotron: Towards end-to-end speech synthesis, 2017. arXiv preprint arXiv:1703.10135.

[7] J. Shen, R. Pang, Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4779-4783, April. IEEE, 2018.

[8] Yu, C., Lu, H., Hu, N., Yu, M., Weng, C., Xu, K., ... & Yu, D. Durian: Duration informed attention network for multimodal synthesis, 2019. arXiv preprint arXiv:1909.01700.

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[10] N. Li, S. Liu, Liu, Y., Zhao, S., & Liu, M. Neural speech synthesis with transformer network. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 6706-6713, July, 2019.

[11] Y. Ren, Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. Fastspeech: Fast, robust and controllable text to speech. Advances in Neural Information Processing Systems, 32, 2019.

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

[13] F. Ribeiro, D. Florêncio, C. Zhang and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 2416-2419, doi: 10.1109/ICASSP.2011.5946971.

[14] C. Borrelli, Bestagini, P., Antonacci, F. et al. Synthetic speech detection through short-term and long-term prediction traces. EURASIP J. on Info. Security 2021, 2 (2021). https://doi.org/10.1186/s13635-021-00116-3.

[15] T. Chen, A. Kumar, et al., E. Generalization of Audio Deepfake Detection. Proc. The Speaker and Language Recognition Workshop (Odyssey 2020), 132-137, 2020, doi: 10.21437/Odyssey.2020-19.

[16] Q. Xie, X. Tian, et al., The multi-speaker multi-style voice cloning challenge 2021. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8613-8617, June, 2021.

[17] Y. Jia, Y. Zhang, R. Weiss, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. Advances in neural information processing systems, 31, 2018.