Brief Review on Text - to - Speech System

Petkar H. J

Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya, Wardha – 445001, Maharashtra, India

Abstract: ***Paper reviews the work on text - to - speech systems including its advances. Brief history along with its types viz articulatory synthesis, formant synthesis and concantenative synthesis is also discussed in the paper.***

Keywords: Text - to - Speech, articulatory synthesis, formant synthesis, concantenative synthesis

1. Introduction

Speech is the simplest modality of communication for human being. Due to advances in technology, speech can be processed by the computing devices. As a result communication tasks are easy to process and propagate. Speech Synthesis is a task to process and convert the text into speech. Main goal of the speech synthesis which is also known as speech - to - text is to produce the artificial speech but it should not be the robotic sound.

TTS is very popular and become a research area in the domain of artificial intelligence and natural language processing.

TTS which is also known as speech synthesis has wide range of applications such as reading text for visually impaired, call centre automation, weather announcement. TTS involves study of various disciplines such as core linguistics, acoustics and phonetics, digital signal processing, artificial intelligence, human - computer - interface and machine learning. Recent advances in the speech synthesis application are striving to produce real time speech in multilingual domain due to the development in the field of deep learning. It is equally notable that the synthesized speech has improved to the extent and suitable for real time applications.

This paper covers the review and recent advancements in speech synthesis including the history, approaches, fundamental concepts of TTS.

2. History of TTS

In late 1950s, electronic devices brought a resolution in speech synthesis system. It was the time when speech synthesis system built using computers [48]. The approaches and techniques used for the speech synthesis are articulatory synthesis [24, 25], formant synthesis [44, 45] and concantenative synthesis [40, 20, 19, 22]. Wide popularity and development in statistical machine learning, another approach comes into picture i. e. statistical parametric speech synthesis [37, 12, 10, 36]. SPSS works well as compared to other methods and approaches and able to predict the various parameter such as fundamental frequency, duration and spectrum.

Real time results are appreciated with neural network based speech synthesis from the year 2010 and it has become prominent method due to large processing power availability [26, 28, 29, 30, 33, 31, 32].

3. Articulatory Synthesis

Articulatory synthesis is a method where human articulators such as toungue, lips, glotis, velar, teeth along with the vocal tract is involved. Sychronised movement of these articulators [24, 25] produce the speech. Articulatory synthesis approach is based on the places of articulation and manner of articulation such as position of lips, teeth [2].

Data used for this model is from MRI or x - ray images. Collecting the data for articulatory synthesis is biggest challenge as heavy expenses are involved in purchasing the high precision MRI and x - ray machine [3]. It is also very difficult to model all the articulators with all positions which result in unsatisfactory quality of synthesized speech.

4. Formant Synthesis

Formant synthesis is based on source - filter model. Extraction of formant parameter and then mapping the parameter with the phoneme based on certain derived rules from the spectrogram. Due to limited memory this approach looses the naturalness. Compare to articulatory synthesis, formant synthesis is more intelligible and works well with low memory system.

5. Concantenative synthesis

Concantenative synthesis [19, 23, 21, 22, 20] is base on the concept of concantenation of chunks of speech stored in the database. In this approach database search is performed to match with input speech unit and thereafter the successfully matched speech units are concantenated. Concantanative approach produce natural speech as compared to articulatory and formant synthesis. It requires more memory as the variety of combination of recording of speech need to store in database. Diphone synthesis and unit selection synthesis are the two approaches to perform concantanative synthesis. Each possible phoneme is recorded and concantenated in diphone synthesis. After the concantenation in diphone synthesis. After the concantenation, diphones are modified so that prosody of the speech could be adjusted by the signal processing unit [1]. Unit selection approach doesnt require signal processing unit stores the speech units along with the prosodic features in the database [4].

Rule based implementation in the concantanative synthesis consume large amount of memory whereas statistical methods such as hidden markov model works well by retrieving the average of similar speech sound [5].

It is also known as statistical parametric synthesis. It overcomes the drawback of concantenative TTS [10, 12, 11, 13] viz naturalness. Audio is more natural. Huge amount of data require in concantenation synthesis whereas in parametric synthesis less data is needed. This method is laso flexible and can modify the parameters.

6. Advances in TTS

In the recent year, natural and accurate speech synthesis is achieved by deep neural network techniques and is proved to be superior than Hidden Markov Model [6].

Deep reinforcement learning (DRL) is investigated for speech synthesis. [7], graph neural network framework [8] is used to formulate the novel neural TTS architecture. This project is named as GraphSpeech. .

Model based on the encoder decoder architecture with self - attention or bi - directional long short - term units are capable of synthesing high quality speech waveform from linguistics features are generated in WaveNet [18] which is termed as first modern neural TTS model.

DeepVoice is the model which follow the statistical parametric syntheis but gradually upgrade the mode with neural network. There are other several models such as Tacatron1/2 [17], DeepVoice3 [14], FastSpeech1/2 [15, 16] are more sophisticated version of speech synthesis. It uses mel - spectrogram to simplify acoustic features.

As compared to concantanative synthesis and statistical parametric synthesis, neural network approach is superior in terms of intelligibility and naturalness. It does not require much human processing and feature could be learnt automatically.

7. Conclusion

Paper reviewed the research work for speech synthesis. It has major two divisions, one is to introduce the history of TTS systems and the gradually moving to starting from electronic era to digital era where the various approaches including neural network approach is also reviewed

References

[1] Tabet, Y.; Boughazi, M. Speech Synthesis Techniques. A Survey. In Proceedings of the International Workshop on Systems, Signal Processing and Their Applications (WOSSPA), Tipaza, Algeria, 9–11 May 2011.

[2] Kaur, G.; Singh, P. Formant Text to Speech Synthesis Using Artificial Neural Networks. In Proceedings of the 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), Gangtok, India, 25–28 February 2019.

[3] Tsukanova, A.; Elie, B.; Laprie, Y. Articulatory Speech Synthesis from Static Context - Aware Articulatory Targets. In International Seminar on Speech Production; Springer: Berlin/Heidelberg, Germany, 2018; pp.37–47. Available online: https: //hal. archivesouvertes. fr/hal - 01937950/document (accessed on 30 September 2020).

[4] Jurafsky, D.; Martin, J. H. Speech and Language Processing, 2nd ed.; Prentice Hall: Hoboken, NJ, USA, 2008; pp.249–284.

[5] Jeon, K. M.; Kim, H. K. HMM - Based Distributed Text - to - Speech Synthesis Incorporating Speaker - Adaptive Training.2012. Available online: https: //www.researchgate. net/publication/303917802_HMM - Based_Distributed_Text - to - Speech_Synthesis_ Incorporating_Speaker - Adaptive_Training (accessed on 30 September 2020). Symmetry 2021, 13, 819 12 of 12

[6] Qian, Y.; Fan, Y.; Hu, W.; Soong, F. K. On the Training Aspects of Deep Neural Network (DNN) for Parametric TTS Synthesis. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014. Available online: https: //ieeexplore. ieee. org/document/6854318 (accessed on 30 September 2020).

[7] Latif, S.; Cuayahuitl, H.; Pervez, F.; Shamshad, F.; Ali, H. S.; Cambria, E. A survey on deep reinforcement learning for audio - based applications. arXiv 2021, arXiv: 2101.00240.

[8] Liu, R.; Sisman, B.; Li, H. Graphspeech: Syntax - aware graph attention network for neural speech synthesis. arXiv 2020, arXiv: 2104.00705.

[9] Takayoshi Yoshimura. Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for hmm - based text - to - speech systems. PhD diss, Nagoya Institute of Technology, 2002.

[10] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Simultaneous modeling of spectrum, pitch and duration in hmm - based speech synthesis. In Sixth European Conference on Speech Communication and Technology, 1999.

[11] Heiga Zen, Keiichi Tokuda, and Alan W Black. Statistical parametric speech synthesis. Speech communication, 51 (11): 1039–1064, 2009.

[12] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Speech parameter generation algorithms for hmm - based speech synthesis. In 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), volume 3, pages 1315–1318. IEEE, 2000.

[13] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura. Speech synthesis based on hidden markov models. Proceedings of the IEEE, 101 (5): 1234–1252, 2013.

[14] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan

Raiman, and John Miller. Deep voice 3: 2000 - speaker neural text - to - speech. Proc. ICLR, pages 214–217, 2018.

[15] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie - Yan Liu. Fastspeech: Fast, robust and controllable text to speech. In NeurIPS, 2019.

[16] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie - Yan Liu. Fastspeech 2: Fast and high - quality end - to - end text to speech. In International Conference on Learning Representations, 2021. URL https: //openreview. net/forum?id=piLPYqxtWuA.

[17] Yuxuan Wang, RJ Skerry - Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end - to – end speech synthesis. Proc. Interspeech 2017, pages 4006–4010, 2017.

[18] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. arXiv preprint arXiv: 1609.03499, 2016.

[19] Joseph Olive. Rule synthesis of speech from dyadic units. In ICASSP'77. IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 2, pages 568–570. IEEE, 1977.

[20] Eric Moulines and Francis Charpentier. Pitch - synchronous waveform processing techniques for text - to - speech synthesis using diphones. Speech communication, 9 (5 - 6): 453–467, 1990.

[21] Rui Liu, Berrak Sisman, Guanglai Gao, and Haizhou Li. Expressive tts training with frame and style reconstruction loss. arXiv preprint arXiv: 2008.01490, 2020.

[22] Andrew J Hunt and Alan W Black. Unit selection in a concatenative speech synthesis system using a large speech database. In 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, volume 1, pages 373–376. IEEE, 1996.

[23] Alan Black, Paul Taylor, Richard Caley, and Rob Clark. The festival speech synthesis system, 1998.

[24] Cecil H Coker. A model of articulatory dynamics and control. Proceedings of the IEEE, 64 (4): 452–460, 1976.

[25] Christine H Shadle and Robert I Damper. Prospects for articulatory synthesis: A position paper. In 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis, 2001.

[26] Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing, pages 7962–7966. IEEE, 2013.

[27] Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In 2013 ieee international conference on acoustics, speech and signal processing, pages 7962–7966. IEEE, 2013. .

[28] Heiga Zen and Ha¸sim Sak. Unidirectional long short - term memory recurrent neural network with recurrent output layer for low - latency speech synthesis. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4470–4474. IEEE, 2015.

[29] Wenfu Wang, Shuang Xu, and Bo Xu. First step towards end - to - end parametric tts synthesis: Generating spectral parameters with neural attention. In Interspeech, pages 2243–2247, 2016.

[30] Hao Li, Yongguo Kang, and Zhenyu Wang. Emphasis: An emotional phoneme - based acoustic model for speech synthesis system. Proc. Interspeech 2018, pages 3077–3081, 2018.

[31] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. arXiv preprint arXiv: 1609.03499, 2016.

[32] Yuxuan Wang, RJ Skerry - Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end - to – end speech synthesis. Proc. Interspeech 2017, pages 4006–4010, 2017.

[33] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Simultaneous modeling of spectrum, pitch and duration in hmm - based speech synthesis. In Sixth European Conference on Speech Communication and Technology, 1999

[34] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Speech parameter generation algorithms for hmm - based speech synthesis. In 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), volume 3, pages 1315–1318. IEEE, 2000.

[35] Heiga Zen, Keiichi Tokuda, and Alan W Black. Statistical parametric speech synthesis. speechcommunication, 51 (11): 1039–1064, 2009.

[36] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura. Speech synthesis based on hidden markov models. Proceedings of the IEEE, 101 (5): 1234–1252, 2013.

[37] Joseph Olive. Rule synthesis of speech from dyadic units. In ICASSP'77. IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 2, pages 568–570. IEEE, 1977.

[38] Eric Moulines and Francis Charpentier. Pitch - synchronous waveform processing techniques for text - to - speech synthesis using diphones. Speech communication, 9 (5 - 6): 453–467, 1990.

[39] Yoshinori Sagisaka, Nobuyoshi Kaiki, Naoto Iwahashi, and Katsuhiko Mimura. Atr _ - talk speech synthesis system. In Second International Conference on Spoken Language Processing, 1992.

[40] Alan Black, Paul Taylor, Richard Caley, and Rob Clark. The festival speech synthesis system, 1998.

[41] Alan W Black, Heiga Zen, and Keiichi Tokuda. Statistical parametric speech synthesis. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP'07, volume 4, pages IV–1229. IEEE, 2007.

[42] Jonathan Allen, Sharon Hunnicutt, Rolf Carlson, and Bjorn Granstrom. Mitalk - 79: The 1979 mit text - to -

speech system. The Journal of the Acoustical Society of America, 65 (S1): S130–S130, 1979.

[43] Dennis H Klatt. Software for a cascade/parallel formant synthesizer. the Journal of the Acoustical Society of America, 67 (3): 971–995, 1980.

[44] Dennis H Klatt. Review of text - to - speech conversion for english. The Journal of the Acoustical Society of America, 82 (3): 737–793, 1987

[45] Cecil H Coker. A model of articulatory dynamics and control. Proceedings of the IEEE, 64 (4): 452–460, 1976.

[46] Christine H Shadle and Robert I Damper. Prospects for articulatory synthesis: A position paper. In 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis, 2001.

[47] Wikipedia. Speech synthesis —Wikipedia, the free encyclopedia. http: //en. Wikipedia.org/w/index. php?title=Speech%20synthesis&oldid=1020857981, 2021.