Deep Learning Based Speech Synthesis

Subjects: Computer Science, Artificial Intelligence
Contributor: Yishuang Ning

Speech synthesis, also known as text-to-speech (TTS), has attracted increasingly more attention. Recent advances on speech synthesis are overwhelmingly contributed by deep learning or even end-to-end techniques which have been utilized to enhance a wide range of application scenarios such as intelligent speech interaction, chatbot or conversational artificial intelligence (AI). For speech synthesis, deep learning based techniques can leverage a large scale of <text, speech> pairs to learn effective feature representations to bridge the gap between text and speech, thus better characterizing the properties of events.

Keywords: deep learning ; speech synthesis ; end-to-end

1. Introduction

Speech synthesis, more specifically known as text-to-speech (TTS), is a comprehensive technology that involves many disciplines such as acoustics, linguistics, digital signal processing and statistics. The main task is to convert text input into speech output. With the development of speech synthesis technologies, from the previous formant based parametric synthesis [1][2], waveform concatenation based methods [3][4][5] to the current statistical parametric speech synthesis (SPSS) [6], the intelligibility and naturalness of the synthesized speech have been improved greatly. However, there is still a long way to go before computers can generate natural speech with high naturalness and expressiveness like that produced by human beings. The main reason is that the existing methods are based on shallow models that contain only one-layer nonlinear transformation units, such as hidden Markov models (HMMs) [7][8] and maximum Entropy (MaxEnt) [9]. Related studies show that shallow models have good performance on data with less complicated internal structures and weak constraints. However, when dealing with the data having complex internal structures in the real world (e.g., speech, natural language, image, video, etc.), the representation capability of shallow models will be restricted.

Deep learning (DL) is a new research direction in the machine learning area in recent years. It can effectively capture the hidden internal structures of data and use more powerful modeling capabilities to characterize the data [10]. DL-based models have gained significant progress in many fields such as handwriting recognition [11], machine translation [12], speech recognition [13] and speech synthesis [14]. To address the problems existing in speech synthesis, many researchers have also proposed the DL-based solutions and achieved great improvements. Therefore, to summarize the DL-based speech synthesis methods at this stage will help us to clarify the current research trends in this area.

2. An Overview of Speech Synthesis

2.1. Basic Concept of Speech Synthesis

Speech synthesis or TTS is to convert any text information into standard and smooth speech in real time. It involves many disciplines such as acoustics, linguistics, digital signal processing, computer science, etc. It is a cutting-edge technology in the field of information processing [15], especially for the current intelligent speech interaction systems.

2.2. The History of Speech Synthesis

With the development of digital signal processing technologies, the research goal of speech synthesis has been evolving from intelligibility and clarity to naturalness and expressiveness. Intelligibility describes the clarity of the synthesized speech, while naturalness refers to ease of listening and global stylistic consistency [16].

In the development of speech synthesis technology, early attempts mainly used parametric synthesis methods. In 1971, the Hungarian scientist Wolfgang von Kempelen used a series of delicate bellows, springs, bagpipes and resonance boxes to create a machine that can synthesize simple words. However, the intelligibility of the synthesized speech is very poor. To address this problem, in 1980, Klatt's serial/parallel formant synthesizer [17] was introduced. The most representative one is the DECtalk text-to-speech system of the Digital Equipment Corporation (DEC) (Maynard, MA,

USA). The system can be connected to a computer through a standard interface or separately connected to the telephone network to provide a variety of speech services that can be understood by users. However, since the extraction of the formant parameters is still a challenging problem, the quality of the synthesized speech makes it difficult to meet the practical demand. In 1990, the Pitch Synchronous OverLap Add (PSOLA) [18] algorithm greatly improved the quality and naturalness of the speech generated by the time-domain waveform concatenation synthesis methods. However, since PSOLA requires the pitch period or starting point to be annotated accurately, the error of the two factors will affect the quality of the synthesized speech greatly. Due to the inherent problem of this kind of method, the synthesized speech is still not as natural as human speech. To tackle the issue, people conducted in-depth research on speech synthesis technologies, and used SPSS models to improve the naturalness of the synthesized speech. Typical examples are HMM-based [19] and DL-based [20] synthesis methods. Extensive experimental results demonstrate that the synthesized speech of these models has been greatly improved in both speech quality and naturalness.

3. Statistical Parametric Speech Synthesis

A complete SPSS system is generally composed of three modules: a text analysis module, a parameter prediction module which uses a statistical model to predict the acoustic feature parameters such as fundamental frequency (F0), spectral parameters and duration, and a speech synthesis module. The text analysis module mainly preprocesses the input text and transforms it into linguistic features used by the speech synthesis system, including text normalization [21], automatic word segmentation [20], and grapheme-to-phoneme conversion [22]. These linguistic features usually include phoneme, syllable, word, phrase and sentence-level features. The purpose of the parameter prediction module is to predict the acoustic feature parameters of the target speech according to the output of the text analysis module. The speech synthesis module generates the waveform of the target speech according to the output of the parameter prediction module by using a particular synthesis algorithm. The SPSS is usually divided into two phases: the training phase and the synthesis phase. In the training phase, acoustic feature parameters such as F0 and spectral parameters are firstly extracted from the corpus, and then a statistical acoustic model is trained based on the linguistic features of the text analysis module as well as the extracted acoustic feature parameters. In the synthesis phase, the acoustic feature parameters are predicted using the trained acoustic model with the guidance of the linguistic features. Finally, the speech is synthesized based on the predicted acoustic feature parameters using a vocoder.

4. Deep Learning Based Speech Synthesis

It is known that the HMM-based speech synthesis method maps linguistic features into probability densities of speech parameters with various decision trees. Different from the HMM-based method, the DL-based method directly perform mapping from linguistic features to acoustic features with deep neural networks which have proven extraordinarily efficient at learning inherent features of data. In the long tradition of studies that adopt DL-based method for speech synthesis, people have proposed numerous models. To help readers better understand the development process of these methods (Audio samples of different synthesis methods are given at: http://www.ai1000.org/samples/index.html.), this paper gives a brief overview of the advantages and disadvantages in Table 1 and makes a detailed introduction in the following.

Table 1. The advantages and disadvantages of different speech synthesis methods, including hidden Markov model (HMM), restrictive Boltzmann machine (RBM), deep belief network (DBN), deep mixture density network (DMDN), deep bidirectional long short-term memory (DBLSTM), WaveNet, Tacotron and convolutional neural network (CNN).

4.1. Restrictive Boltzmann Machines for Speech Synthesis

In the field of speech synthesis, Boltzmann machine (RBM) is usually regarded as a density model forgenerating the spectral envelope of acoustic parameters. It is adopted to better describe the distribution of high-dimensional spectral envelopes to alleviate the over-smooth problem in HMM-based speech synthesis [23].

4.2. Multi-distribution Deep Belief Networks for Speech Synthesis

Multi-distribution deep belief network (DBN) [24] is a method of modeling the joint distribution of context information and acoustic features. It models the coutinuous spectral, discrete voiced/unvoiced (V/UV) parameters and the multi-space F0 simultaneously with three types of RBMs. In DBNs, the visible unit can obey different probability distributions, therefore, it is possible to characterize the supervectors that are composed of these features.

4.3. Deep Bidirectional LSTM-based Speech Synthesis

BLSTM-RNN is an extended architecture of bidirectional recurrent neural network (BRNN) [25]. It replaces units in the hidden layers of BRNN with LSTM memory blocks. With these memory blocks, BLSTM can store information for long and short time lags, and leverage relevant contextual dependencies from both forward and backward directions for machine learning tasks.

When using deep BLSTM-based (DBLSTM) model to predict acoustic parameters for speech synthesis, first we need to convert the input text prompt into a feature vector, and then use the DBLSTM model to map the input feature to acoustic parameters. Finally, the parameter generation algorithm is used to generate the acoustic parameters and a vocoder is utilized to synthesize the corresponding speech.

4.4. End-to-End Speech Synthesis

A TTS system typically consists of a text analysis front-end, an acoustic model and a speech synthesizer. Since these components are trained independently and rely on extensive domain expertise which are laborious, errors from each component may compound. To address these problems, end-to-end speech synthesis methods which combine those components into a unified framework have become the main stream of speech synthesis field. In the following we will give a brief introduction to the end-to-end speech synthesis methods.

4.4.1. Speech Synthesis Based on WaveNet

WaveNet [26] is a complete probabilistic autoregressive model that predicts the probability distribution of the current audio sample based on all samples which have been generated before. As an important component of WaveNet, dilated causal convolutions are used to ensure that WaveNet can only use the sampling points from 0 to t - 1 while generating the tth sampling point. The original WaveNet model uses autoregressive connections to synthesize waveforms one sample at a time, with each new sample conditioned on the previous ones.

4.4.2. Speech Synthesis Based on Tacotron

Tacotron [27][28] is a fully end-to-end speech synthesis model. It is capable of training speech synthesis model given <text, audio> pairs, thus alleviating the need for laborious feature engineering. And since it is based on character level, it can be applied in almost all kinds of languages including Chinese Mandarin. Tacotron uses seq2seq model with attention mechanism to map text to spectrogram which is a good representation of speech. Since spectrogram doesn't contain phase information, the system uses Griffin-Lim algorithm [29] to reconstruct the audio by estimating the phase information from the spectrogram iteratively.

5. Conclusions

Deep learning that is capable of leveraging large amount of training data has become an important technique for speech synthesis. Recently, increasingly more researches have been conducted on deep learning techniques or even end-to-end frameworks and achieved state-of-the-art performance. This paper gives an overview to the current advances on speech synthesis and compare both of the advantages and disadvantages among different methods, and discusses possible research directions that can promote the development of speech synthesis in the future.

References

1. Klatt, D.H. Review of text-to-speech conversion for English. J. Acoust. Soc. Am. 1987, 82, 737–793.

2. Allen, J.; Hunnicutt, M.S.; Klatt, D.H.; Armstrong, R.C.; Pisoni, D.B. From Text to Speech: The MITalk System; Cambridge University Press: New York, NY, USA, 1987.
3. Murray, I.R.; Arnott, J.L.; Rohwer, E.A. Emotional stress in synthetic speech: Progress and future directions. Speech Commun. 1996, 20, 85–91.
4. Festival. Available online: http://www.cstr.ed.ac.uk/projects/festival/ (accessed on 3 July 2019).
5. Chu, M.; Peng, H.; Zhao, Y. Microsoft Mulan. A bilingual TTS system. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03), Hong Kong, China, 6–10 April 2003; pp. 264–267.
6. Tokuda, K.; Nankaku, Y.; Toda, T. Speech synthesis based on hidden Markov models. Proc. IEEE 2013, 101, 1234–1252.
7. Murray, I.R. Simulating Emotion in Synthetic Speech; University of Dundee: Dundee, UK, 1989.
8. Tokuda, K.; Yoshimura, T.; Masuko, T.; Kobayashi, T.; Kitamura, T. Speech parameter generation algorithms for HMM-based speech synthesis. In Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 5–9 June 2000; Volume 3, pp. 1315–1318.
9. Ratnaparkhi, A. A Simple Introduction to Maximum Entropy Models for Natural Language Processing; University of Pennsylvania: Philadelphia, PA, USA, 1997.
10. Yang, J.A.; Wang, Y.; Liu, H.; Li, J.H.; Lu, J. Deep learning theory and its application in speech recognition. Commun. Countermeas. 2014, 33, 1–5.
11. Graves, A. Supervised Sequence Labelling with Recurrent Neural Networks; Springer: Berlin, Germany, 2012.
12. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the Annual Conference on Neural Information Processing Systems, Montreal, QB, Canada, 8–13 December 2014; pp. 3104–3112.
13. Graves, A.; Mohamed, A.R.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649.
14. Zen, H.; Tokuda, K.; Alan, W.B. Statistical parametric speech synthesis. Speech Commun. 2009, 51, 1039–1064.
15. Xu, S.H. Study on HMM-Based Chinese Speech Synthesis; Beijing University of Posts and Telecommunications: Beijing, China, 2007.
16. Sotelo, J.; Mehri, S.; Kumar, K.; Santos, J.F.; Kastner, K.; Courville, A.; Bengio, Y. Char2wav: End-to-end Speech Synthesis. In Proceedings of the International Conference on Learning Representations Workshop, Toulon, France, 24–26 April 2017.
17. Klatt, D.H. Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Am. 1980, 67, 971–995.
18. Moulines, E.; Charpentier, F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphone. Speech Commun. 1990, 9, 453–456.
19. Yoshimura, T.; Tokuda, K.; Masuko, T.; Kobayashi, T.; Kitamura, T. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In Proceedings of the Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99), Budapest, Hungary, 5–9 September 1999; pp. 2347–2350.
20. Zen, H.; Senior, A.; Schuster, M. Statistical parametric speech synthesis using deep neural networks. In Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7962–7966.
21. Lin, Z.H. Research on Speech Synthesis Technology Based on Statistical Acoustic Modeling; University of Science and Technology of China: Hefei, China, 2008.
22. Fan, Y.; Qian, Y.; Xie, F.L.; Soong, F.K. TTS synthesis with bidirectional LSTM based recurrent neural networks. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
23. Kang, S.Y.; Qian, X.J.; Meng, H. Multi-distribution deep belief network for speech synthesis. In Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8012–8016.
24. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM networks. In Proceedings of the IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July–4 August 2005; pp. 2047–2052.

25. Seltzer, M.L.; Droppo, J. Multi-task learning in deep neural networks for improved phoneme recognition. In Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6965–6969.
26. Oord, A.V.D.; Li, Y.; Babuschkin, I.; Simonyan, K.; Vinyals, O.; Kavukcuoglu, K.; Casagrande, N. Parallel wavenet: Fast high-fidelity speech synthesis. arXiv 2017, arXiv:1711.10433.
27. Barron, A. Implementation of Google's Tacotron in TensorFlow. Available online: https://github.com/Kyubyong/tacotron (accessed on 20 October 2018).
28. Ito, K. Tacotron Speech Synthesis Implemented in TensorFlow, with Samples and a Pre-Trained Model. Available online: https://github.com/keithito/tacotron (accessed on 20 October 2018).
29. Yamamoto, R. PyTorch Implementation of Tacotron Speech Synthesis Model. Available online: https://github.com/r9y9/tacotron_pytorch (accessed on 20 October 2018).
