Parameterizing Human Speech Generation

Nazariy Perepichka

Ukrainian Catholic University, Lviv 79007, Ukraine
perepichka@ucu.edu.ua

Abstract. In modern days, the synthesis of human images and videos is, arguably, one of the most popular topics in the Data Science community. The synthesis of human speech is less trendy but is deeply bounded to the mentioned topic. Since the publication of WaveNet paper in 2016, the State-of-the-Art transited from parametric and concatenative systems to the use of deep learning models. Each significant paper on the topic mentions the way to parameterize the output audio with different voices and sentiments, though parameterizing is not the major focus of those works. Most of the time-proven solutions require re-training of models for speech synthesis of unknown to the model voice. In my master's project, I aim to implement a competitive text-to-speech solution, enhance parameterization abilities, and improve the performance of current models.

Keywords: text-to-speech · deep learning · recurrent neural networks · audio generation

1 Introduction

The speech synthesis problem has a long research history. The desire to generate human speech from a written text is easy to understand. The potential areas of applications for such systems are enormous: generation of audio books, voice acting of films, making computer systems socially accessible.

In the last five years, the researchers have made significant progress. The big breakthrough happened with the applying of deep learning techniques to the task. With every year, the solutions diminish the difference between programmatically generated and human speech samples.

Though the quality of generated speech increased, human speech possesses multiple parameters: tone of speech, the mood of the speaker, melodic component. Replication of these parameters is still not a fully mature research area. The ability to generate natural-sounding, emotional speech can become the next big breakthrough in speech synthesis history.

During the work on my master's diploma, I plan to research possibilities for speech parameterization and present working solutions for such a task.

The rest of this paper is organized as follows. In Section 2, I describe the domain: evaluation metrics, history of algorithms development, results of text-to-speech (TTS)

In: Proceedings of the 1st Masters Symposium on Advances in Data Mining, Machine Learning, and Computer Vision (MS-AMLV 2019), Lviv, Ukraine, November 15-16, 2019, pp. 52–58

systems. In Section 3, I define the problem for my master's thesis and describe the proposed solution, along with a description of the datasets and timeline of the research.

2 Domain Description

2.1 Evaluation of Algorithms

The main two goals of the system are intelligibility (capability of being understood) and naturalness (ability to mimic human speech). Human perception of the output defines both of the evaluation parameters. Therefore, the evaluation of TTS systems requires subjective techniques for quality measurements. The most popular metric is the mean opinion score (MOS) - average grade given to the audio sample by respondents. MOS is arithmetic mean over user-given rating and depends on two parameters: quantity of respondents and grading scale:

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

2.2 Classical TTS Systems

Current State-of-the-Art solutions have two predecessors, which are considered as classical speech synthesis approaches: concatenative and parametric TTS.

Concatenative systems are the implementation of the intuitive idea to compose the final audio out of small pre-recorded samples. This system builds output by concatenating recording units (words, phonemes). Such an approach satisfies the intelligibility requirement, but it has multiple drawbacks: large, hard to collect unit database; artificial, so to say ”robotic”, sound; hardcoded rule-based programming.

Parametric systems (Fig. 1) exploit a statistical approach for speech generation. Such systems model synthesized speech based on acoustic and linguistic features. The mathematical model is called the vocoder. Parametric TTS requires feature engineering by hand, and it is the main drawback of approach. Hypothetically, with proper features selection, such systems should work on the same level as deep learning models, but practically such systems perform much poorly.

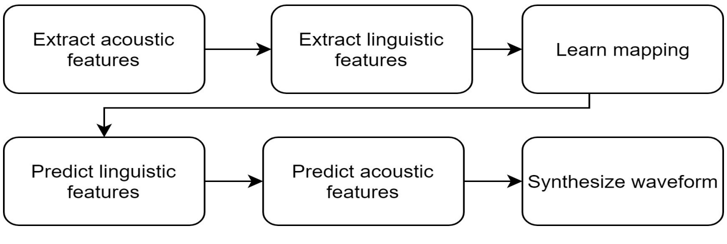
Extract acoustic features
Extract linguistic features
Learn mapping
Predict linguistic features
Predict acoustic features
Synthesize waveform

Fig. 1. A generic workflow of a parametric TTS system

2.3 Deep Learning TTS

Deep learning approach to TTS is the natural step forward from parametric systems. The main difference is the replacement of the features engineered by humans by the features learned by machine learning models. A breakthrough in speech synthesis happened with the publication from Google Deepmind in 2016 [1]. The researchers presented a new architecture, called WaveNet, which operates directly on the raw audio waveform and functions as a vocoder. The joint probability of a waveform is factorized as a product of conditional probabilities, as follows:

$$p(x) = \prod_{t=1}^{T} p(x_t | x_1, \dots, x_{t-1}) \tag{2}$$

Authors modeled the conditional probability distribution with a stack of convolutional layers (Fig. 2). On the output layer, we receive conditional probability distribution for x_t given by softmax function.

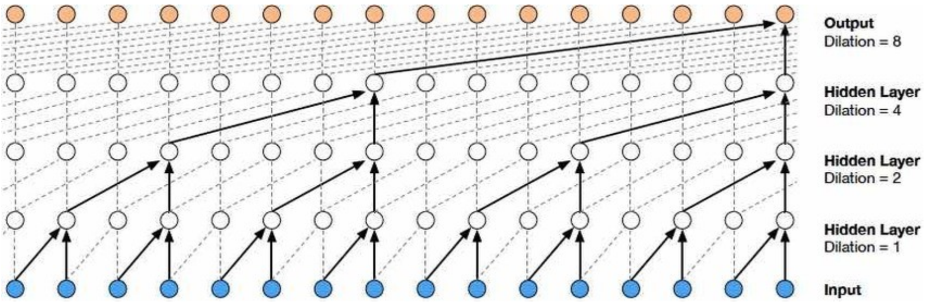

Fig. 2. Visual representation of WaveNet convolutional layers [1]

In [1], the authors mention the ability to parameterize output audio by incorporating additional parameters h in the joint probability, as follows:

$$p(x|h) = \prod_{t=1}^{T} p(x_t | x_1, \dots, x_{t-1}, h) \tag{3}$$

For their research, they parameterized the narrator – by passing encoding of the voice – and text – by passing linguistic features of the text. The TTS solution significantly outperformed all the previous benchmarks.

Since then, the WaveNet vocoder is used in the most of TTS research, which focus mostly on the representation of parameters. In 2018, Google researchers presented [2], which described Tacotron 2 architecture for speech synthesis (Fig. 4). The model projects text to MEL-scale spectrograms and then uses modified WaveNet vocoder to audio output itself. This architecture combines WaveNet and Tacotron 1 models. Tacotron 2 achieved a MOS score of 4.53, in comparison to 4.58 for professionally recorded human speech.

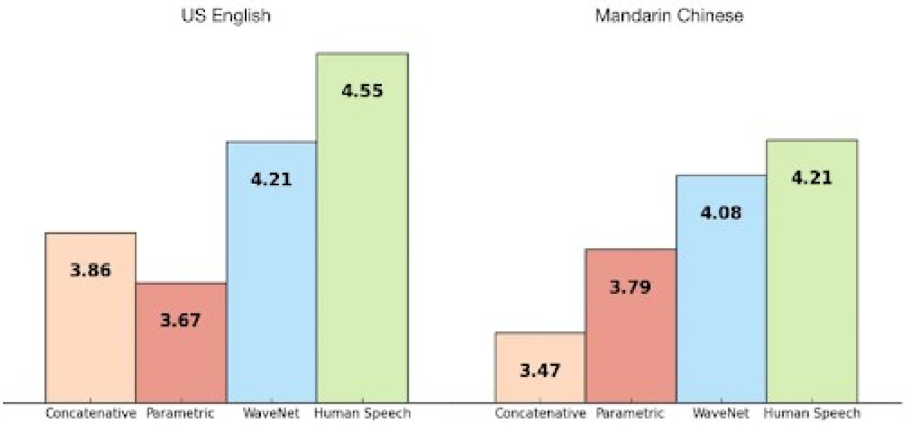

Fig. 3. MOS scores presented in WaveNet paper [1]

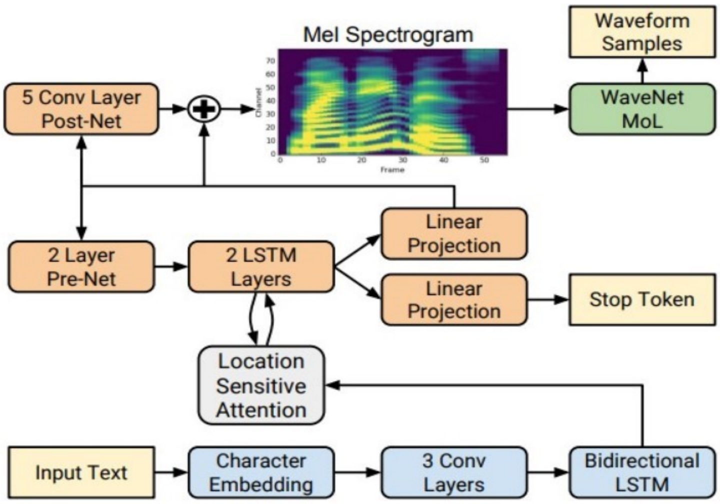

Fig. 4. Tacotron 2 architecture [2]

This year, the engineers from Dessa company presented the results of their new TTS model named Realtalk [6]. They published only a YouTube video presenting the generated speech of Joe Rogan (a famous podcast host) and wrote two small articles for the general public with minimal technical specifications. The output audio mimics the emotions and intonations of the narrator and synthesizes highly realistic output. They claim that their model has been trained on a significantly smaller dataset compared to the State-of-the-Art models (8 hours to 20 hours of audio) with no loss in quality of generated audio.

3 Proposal

3.1 Problem Statement

For my master's, I plan to implement the end-to-end TTS system, which generates the speech specific to speaker voice and tone.

3.2 Proposed Solution

The system will have three main components. The first component will encode sentiment and voice and pass the representation to the second component – the sequential encoder. I want to experiment with possible implementations of sequential encoding. You can see possible solutions and forms of representation in Fig. 5.

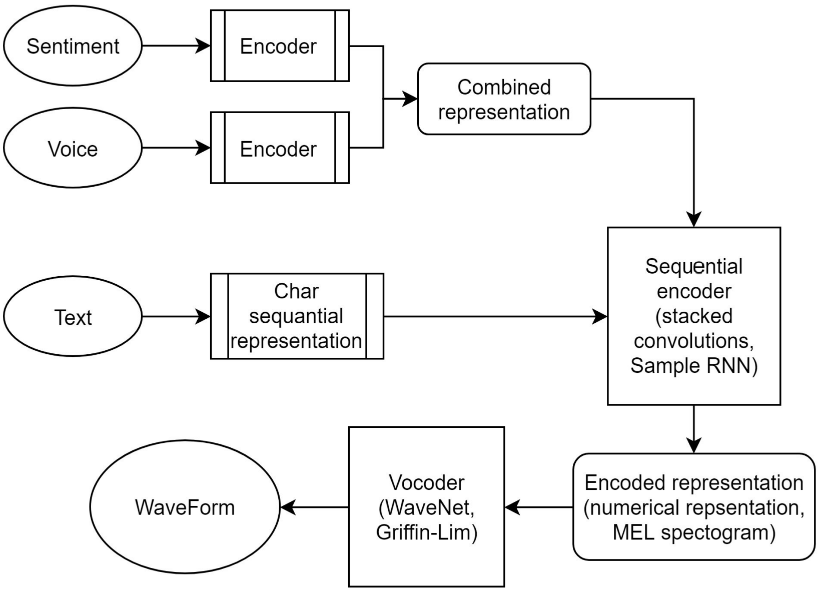
Sentiment
Encoder
Voice
Encoder
Combined representation
Text
Char sequantial representation
Sequential encoder (stacked convolutions, Sample RNN)
Encoded representation (numerical repsentation, MEL spectogram)
Vocoder (WaveNet, Griffin-Lim)
WaveForm

Fig. 5. Proposed architecture

For the third component – the vocoder – I want to experiment with wave generation algorithms: starting from more classical algorithms, like Griffin-Lim, to that more frequently used in current research, like WaveNet. The model will consist of two parts: the first one will represent input features: voice, sentiment, text; the second model will be vocoder for audio generation.

3.3 Datasets Description

In my research, I plan to use the following datasets:

VoxCeleb [3] – the dataset contains more than two thousand hours of speech from seven thousand of different speakers.

Toronto emotional speech set (TESS) [4] – the dataset contains 2800 audio samples recorded by two female speakers aged 26 and 64 years. Each sample has a length of two seconds

The LJ speech dataset [5] – the dataset contains 13,100 audio clips of a single speaker with transcriptions of the text. The total length of clips is approximately 24 hours.

Self-collected dataset – I plan to collect a transcripted dataset with different voices from YouTube videos. I implemented the script, which downloads audio with subtitles, generated by the platform. By downloading data from specific channels, I can ensure speaker identity.

3.4 Timeline

Below I describe the timeline of research:

Present state

- researched most significant papers on the topic;
- build the script for extracting audio and description from YouTube videos;
- wrote the general audio preprocessing scripts;
- worked with pre-trained models;

October 2019

- collecting the dataset;
- modeling;
- experiments with vocoder input representation;

November 2019

- experimenting with vocoder;
- hyper-parameters tuning;
- building the pipeline; packaging the model;

December 2019

- analyzing the results;
- evaluating the model using MOS;
- working on the diploma text;

January 2019

- bringing everything together.

4 Conclusion

The goal of my work is to build a working TTS-system with the ability to parameterize audio output. By incorporating current knowledge in the domain and experimenting with new approaches, I want to present a working solution for my thesis defense.

References

1. Oord, A.V.D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: WaveNet: a generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
2. Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang,Y., Wang, Y., Skerrv-Ryan, R., Saurous, R.A.: Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4779–4783. IEEE Press, New York (2018)
3. Nagrani, A., Chung, J.S., Zisserman, A.: VoxCeleb: a large-scale speaker identification dataset. arXiv preprint, arXiv:1706.08612 (2017)
4. Kate Dupuis, M. Kathleen Pichora-Fuller, University of Toronto, Psychology Department. https://tspace.library.utoronto.ca/handle/1807/24487
5. Keith Ito: The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/
6. RealTalk: We Recreated Joe Rogan’s Voice with AI. https://dessa.com/realtalkwerecreated-joe-rogans-voice-with-ai/