Google Duplex - A Big Leap in the Evolution of Artificial Intelligence

Parth Patel
Dept. of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Pratik Kanani
Dept. of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

ABSTRACT

Google CEO Sundar Pichai launched Google Du- plex on May 2018 at Google I/O developer conference. He demonstrated how the system functioned with an AI-driven voice to achieve its objective, which is to help individuals make appointments to various organizations via mobile phone, without any involvement from the client. Pichai's demo indicated that AI voice would comprehend the voice of the human on the opposite side of the call and react back with right responses to that individual's requests and inquiries. Google Duplex's voice would even use words like "um" and take breaks while answering to make it sound much more like a genuine human. This research is aimed at enhancing people's understanding of the Google Duplex and other similar and relevant research published in the domains of Human Computer Interaction and Artificial Intelligence.

General Terms

Human Computer Interaction, Artificial Intelligence, Neural Network, Machine Learning

Keywords

Google Duplex, WaveNet, Google Automatic Speech Recognition

1. INTRODUCTION

Human Computer Interactive systems are being limited by the time the system takes to respond to the user's requests. The optimal solution in this domain is not to develop better processing systems but to improve the user interface. Quicker, more natural, and more helpful methods for user and system to trade data is required. Currently, the technology is being limited by both the user's ability to communicate and the input/output devices and the methods that are used to communicate. The test before us is to invent new gadgets and type of interaction between system and user that would better fit and utilize the relevant characteristics of human communication. [1]

The main aim of human-computer interactive system is to empower individuals to have a casual discussion with computer, as one would have with another human being. Lately, we have seen an upheaval in the capacity of computers to comprehend and to produce regular discourse, particularly with the use of deep neural networks (e.g WaveNet, Google voice search). All things considered, even with the current cutting edge automated systems, it is irritating to converse with unnatural modernized voices that don't comprehend common language. Specifically, automated phone systems are still attempting to perceive straightforward words and orders. The flow of the conversation is not engaging forcing the caller to adapt to the system rather than the system adapting to the caller. [4,5]

Google Duplex is another innovation from Google that helps in having a normal discussion between a user and computer to do "real world" assignments via phone. The innovation is coordinated towards finishing explicit assignments, for example, scheduling specific kinds of reservations and appointments. For such errands, the framework makes the conversational experience as normal as could reasonably be expected, permitting individuals to talk ordinarily, similar as they would to someone else, without adjusting to a machine. One of the important aspects of Google Duplex is to train it extensively in closed domains. Duplex can only carry out normal conversation in this specific domain. It cannot carry out general conversations. [5]

2. METHODOLGY

The main objective of the research is to improve the reader's comprehension of the Google Duplex and other comparable and important exploration done in the areas of Human Computer Interaction and Artificial Intelligence. We did so by reviewing literature of Google Duplex, WaveNet and to do Google Voice Search technologies. Then we collected research publications and citation data. Previous research has shown that recurrent neural networks (RNNs) are designed to cope with these challenges. The search was restricted o journal and conference publications available in English language only to warrant the integrity of selected documents. The research focused on papers published with title and keywords involving "Human Computer Interaction", "Google Duplex", "Recurrent Neural Networks". The search provided 257 publications, covering wide range of publishing journals and research papers. To this information, we developed a qualitative measure of at least-one citation and extracted a data set of 11 publications.

3. OVERVIEW

To have a normal conversation with a human, Google Duplex system needs to be trained in that particular manner. There are a few difficulties in regulating normal conversation: natural language is difficult to comprehend, innate behavior is difficult to model, processing needs to be fast to handle latency issues, and creating speech that sounds natural and have a proper pitch is troublesome.

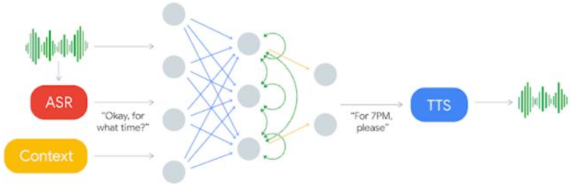

Fig. 1: Overview of Google Duplex [5]

In natural spontaneous speech, a person would be vague and quicker in conversation with another person when compared to the conversation they would have with an automated system. This makes speech recognition tough which causes an increase in word error rates. Loud surrounding noises and bad quality of sound exacerbate the problem.

To cope with these challenges, Duplex uses a recurrent neural network (RNN) which is built using TensorFlow Extended (TFX). Duplex's RNN was trained on a corpus of anonymized phone conversation data to obtain its high accuracy. Output from Google's automatic speech recognition (ASR) technology, conversational history, audio features, conversation parameters (like what service is required, when is it required) are few of the information that is fed into recurrent neural network. The comprehensive model was trained separately for each task but utilized the collective corpus across multiple tasks. Finally, hyperparameter optimization was used from TFX to further improve the model. A concatenative text to speech (TTS) engine and a synthesis TTS engine (using Tacotron and WaveNet) are used together to control the pitch of the response which is based on the situation at hand. [5]

4. TEXT CONVERSION

A complete information about the text needs to be provided before WaveNet converts the text into speech. Information like the word, current phoneme, syllables are used to determine the linguistic and phonetic features of the text. RNN network predicts this information using earlier samples of audio and the response we want to provide. This synthesis process largely includes the conversion of a phonetic representation of the text into a number of speech parameters. The speech parameters are then converted into a speech waveform by a speech synthesizer (Dutoit, 1997). The problem with synthesis-by-rule (Mattingly, 1981) systems is that the transitions between phonetic representations are not natural due to transition rules tend to produce only a few styles of transition. In addition, a large set of rules must be defined. [7]

5. WAVENET

WaveNet is a dilated convolutional neural network which produces raw audio waveforms. The model is initially trained on huge corpus of real human voice samples (tens of thousands of samples per second of audio). This model makes prediction of each sample based on the output of previously conditioned audio samples, thus making the model both autoregressive and probabilistic. When used as TTS, it yields best in class execution, with human audience members grading it as essentially more human sounding than the best parametric and concatenative frameworks for both English and Mandarin. A WaveNet can comprehend the properties of various speakers with equal accuracy and can switch between them by the influence of the speaker character. It also generates unique and often highly realistic fragments of music after its trained to model the music. It tends to be utilized as a discriminative model, returning promising outcomes for phoneme acknowledgment.[6]

In general, the idea of WaveNet is to predict the audio sample xt based on previous samples x1, ..., xt1. For TTS task, the network input is audio waveforms (one audio sample per step) plus linguistic features generated from corresponding texts. The network output at each step xˆt is the prediction of the audio sample at next time step, given input audio sample from previous steps. The prediction is categorical, namely there is a probability corresponding to each possible next-step audio sample. The loss is the difference between predicted sample and real audio sample in next step. During test, the input is a initial audio sample and linguistic features from test texts. The output for each time step is used as the input for next step. Finally, the sequence of output samples is the generated speech for test text. [7]

5.1 Diluted Convolution

The structure of WaveNet is based on Dilated Convolution. We can think of the dilated convolution as convolution with filters having holes. The intuition is with dilated convolution, the output of a one time step is able to depend on a long sequence of inputs from previous time steps. This is how the network achieve P(xt —x1, ..., xt1)in practice. Figure below provides a visualization of dilated convolution with filter size 1 × 2. In the figure, we can see that 3 hidden layers are needed to gather information from 16 inputs. Without dilation, we need 15 hidden layers. [3,7]

5.2 Gated Activation and Residual Units

In the non-linearity part of network structure, Oord et al applied a gated activation unit similar to the activation in LSTM. In detail, the activation formula is the following: z = tanh(Wf,k × x) × (Wg,k × x) where represents sigmoid function, Ws are weights and × represents element wise multiplication. In the network structure Oord et al also applied residual net structure (He et al, 2016), which is shown to be useful for making deep network converge. [2,7]

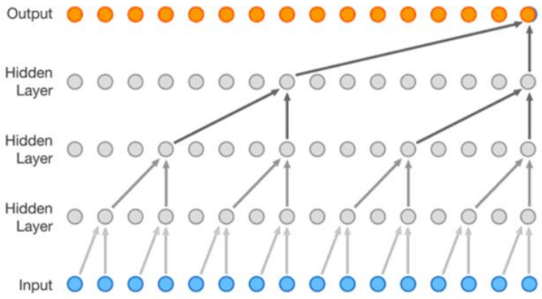

Fig. 2: Dilated Convolution [6]

5.3 Local Conditioning and Global Conditioning

Oord et al also introduced the idea of conditioning to include more information in the model in order to achieve more meaningful tasks. Conditioning is implemented through adding more learn-able parameters. Currently there are two kinds of conditioning, global and local. Global conditioning records the information of different speakers such that during generation people can choose to generate voice from a specific speaker. Local conditioning records the text information of training data, so that people can choose to generate audio from a specific text file, which is exactly the goal of TTS. Following is the local conditioning version of activation function: z = tanh(Wf,k×x+Vf,k×y) × (Wg,k×x+Vg,k×y) where V s are newly introduced learnable variables and y is a function character from input text file. [7]

6. PERFORMANCE ANALYSIS

WaveNet was trained using some of Google's TTS data- sets to assess its execution. The figure below analysis WaveNet's performance on a range of 1 to 5 (with 1 being the lowest and 5 being the highest) using Mean Opinion Score (MOS), in comparison to Google's existing Text to Speech frameworks (both concatenative and parametric) and humans. MOS is a mean of all single ratings (1-5) given by human for a particular quality of a speech. The score in the diagram were collected from five hundred human evaluations on hundred sentences. The other applications included in this test are Multi-Speaker Speech Generation, Music Generation and Speech Recognition. As evident, the performance level of the machine generated speech increases when utilizing Wave-Nets, reducing the gap between machine and human speech by over 50% in language Mandarin Chinese and US English. Google's present Text To Speech frameworks are believed to be of upmost quality around the world, so improving it with a solitary model is a significant accomplishment.[8]

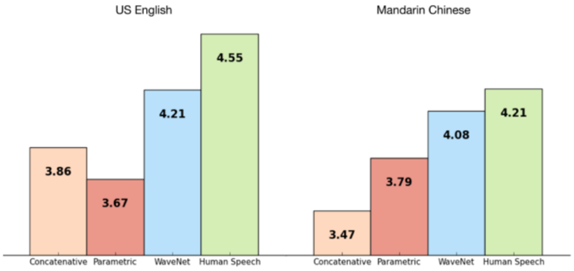

Fig. 3: Performance Analysis [6]

7. REFERENCES

[1] Aaron van den Oord and Sander Dieleman and Heiga Zen and Karen Simonyan and Oriol Vinyals and Alex Graves and Nal Kalchbrenner and Andrew Senior and Koray Kavukcuoglu WaveNet: A Generative Model for Raw Audio arXiv:1609.03499

[2] P Mulchandani, M.U. Siddiqui, P Kanani, "Real-Time Mosquito Species Identification using Deep Learning Techniques", International Journal of Engineering and Advanced Technology (IJEAT), 2019 pg 2000 - 2003.

[3] P Kanani, M Padole, "Deep Learning to Detect Skin Cancer using Google Colab", International Journal of Engineering and Advanced Technology (IJEAT), Blue Eyes Intelligence Engineering and Sciences Publication 2019 Pg 2176-2183

[4] M. Nagda, P. Mehta, S. Lamba, P. Kanani, "Gamification in Plant Education for Children", International Journal of Psychosocial Rehabilitation, Hampstead Psychological Associates (2020), pp.8845-8858

[5] Google AI Blog 'Duplex AI system for natural conversation' [Online]. Available: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

[6] Wavenet generative model Raw Audio Deepmind AI [Online]. Availabe: https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

[7] Yi Zhao1, (Student Member, IEEE), Shinji Takaki 2, (Member, IEEE), Hieu-Thi Luong2 (Student Member, IEEE), Junichi Yamagishi2,3,(Senior Member, IEEE), Daisuke Saito1 (Member, IEEE), Nobuaki Minematsu1, (Member, IEEE), 'Wasserstein GAN and Waveform Loss-based Acoustic Model Training for Multi-speaker Text-to-Speech Systems Using a WaveNet Vocoder'.

[8] Diemo Schwarz Current research in concatenative sound synthesis International Computer Music Conference (ICMC)