Edge Compute Voice Assistant

Gurshish Dhupar[1], Saurabh Godbole[2], Indrajeet Kulkarni[3], Prof. Anjali Purohit[4]

[1,2,3,4] *Student, Dept. of E&TC Engineering, MITCOE Pune, Maharashtra, India*
[4] *Professor, Dept. of E&TC Engineering, MITCOE Pune, Maharashtra, India*

Abstract - *Recognition of human voice and extraction of meaning from it is a current challenge. The objective of this paper is to provide a reasonable speech cognition solution which understands and executes human issued commands. Our idea draws on the Wavenet architecture by Deepmind for Speech-to-Text and Text-to-Speech operation. We implement Natural Language Processing (NLP) to extract meaning from the text input. We use Spacy - A Python library, and sequence-to-sequence mapping for Natural Language Processing of the data.*

Key Words: **Wavenet, NLP, Spacy, Sequence-to-sequence, GRU, LSTM.**

1. INTRODUCTION

Virtual assistants are increasingly in demand today. Most of the existing voice assistants have cloud based models, with little to none edge compute ability. We decided to make an efficient voice assistant/ automation agent, which would perform recognition and interpretation on the user device.

A voice assistant has two primary tasks:

1. Accept user input in the form of speech and convert it to a processable format.

2. Extract 'meaning' and perform actions accordingly (including direct speech response).

In most of the current public voice assistants, the second task and its related processing are carried out in the cloud.

Keeping in mind the erratic network conditions in India and the security concerns of users, our voice assistant offers a completely offline functionality. It is designed to be deployable in the most unconnected of places. All the stages of processing, including the Natural Language Processing take place on the user's device and do not require a server database or model to be functional.

2. LITERATURE REVIEW

In the paper published by A. Graves, A. R. Mohamed and G. Hinton [4], they investigated the working of deep recurrent neural networks, which combine the multiple levels of representation that have proved so effective in deep networks with the flexible use of long range context that empowers RNNs. When properly trained end-to-end their model achieved a test set error of 17.7% on the TIMIT benchmark. Thus the combination of LSTM with RNN proved to work more efficiently than the working of RNN independently.

K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink and J. Schmidhuber [6], in their paper presented a large scale analysis of eight LSTM variants on three representative tasks: Speech recognition, handwriting recognition, and polyphonic music modeling. They used random search for optimizing hyper parameters of all LSTM variants and assessed their importance using powerful functional Analysis of Variance framework. They summarized 5400 experimental runs and concluded that the forget gate and the output activation function to be its most critical components. The studied hyper parameters are virtually independent and derive guidelines for the efficient adjustments.

The paper published by N. Naderi and B. Nasersharif [5], showed the implementation multiple CNN for robust speech recognition. In the proposed model CNN has noisy speech spectrum as input and its outputs are denoised logarithm of Mel filter bank energies (LMFBs) and convolution filter size is fixed. They implemented a method named Multiresolution CNN which proposes use of multiple CNNs with different sizes of convolution filter to provide different frequency resolutions for feature extraction. The model operates in 2 manners, in the first manner all the outputs are concatenated to construct feature vectors; while in the second manner some outputs are selected from each CNN based on the filter size and are concatenated to obtain feature vectors.

S. Basu, J. Chakraborty, A. Bag and M. Aftabuddin [8], published a comparative study on various techniques used for emotion detection using speech recognition. This paper stated the main challenge in detection of emotion through speech is the selection of speech corpora. The paper presented a detailed description of a prime feature extraction technique named Mel Frequency Cepstral Coefficient (MFCC) and brief description of the working principle of some classification models.

Q. Zhang and A. Benveniste [10] published a paper explaining the applications of 'Wavelet networks'. They proposed an idea of replacing neurons by 'wavelons' which are computing units obtained by cascading an affine transform and a multidimensional wavelet. Then these affine transforms and the synaptic weights must be identified from possibly noise corrupted input/output data. An algorithm of backpropagation type is proposed for wavelet network training.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior and Koray Kavukcuoglu[2] of Deepmind published a paper proposing a generative model for raw audio named 'WaveNet'. This is a deep neural network which

is fully probabilistic and autoregressive with the predictive distribution for each audio sample conditioned on all previous ones. A single WaveNet can capture the characteristics of many different speakers with equal fidelity, and can switch between them by conditioning on the speaker identity.

When applied to text-to-speech, it yields state-of-the-art performance, with human listeners rating it as significantly more natural sounding than the best parametric and concatenative systems. When trained to model music, it generates novel and often highly realistic musical fragments. It can also be employed as a discriminative model, returning promising results for phoneme recognition.

3. ARCHITECTURE

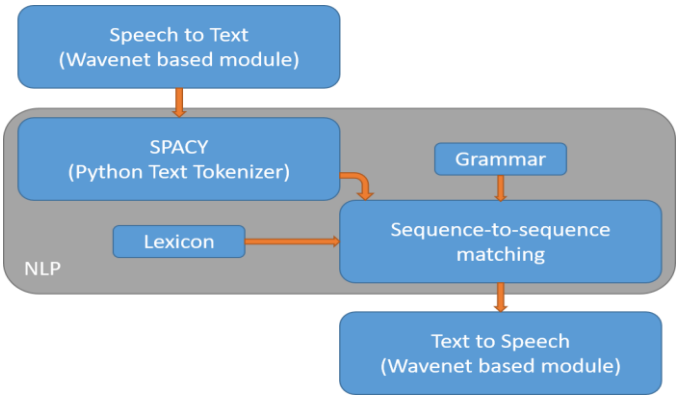

Fig -1: System architecture

The system consists of three major modules:

1. Speech-to-Text (STT) module
2. Natural Language Processing module
3. Text-to-speech (TTS) module

The STT module works as the input block of the system and takes input from the user as speech. This data is then passed to the NLP module which extracts meaning from the data obtained. NLP module infers the task to be performed from the input speech data. Lastly the TTS module converts the text output to voice based response. Thus, the system works as a full-fledged voice assistant.

3.1 Speech-to-Text

We use the Wavenet architecture designed by DeepMind to perform speech to text conversions. Wavenet is a generative model which operates directly on raw audio waveforms. At its core Wavenet implements causal convolutional neural networks. Thus, Wavenet is a fully probabilistic and autoregressive model. The predictive distribution for each audio sample is conditioned on the previous ones. Since Wavenet uses causal convolutions it does not require recurrent connections. This increases the processing speed for long sequences.

Thus, this module converts the speech input to equivalent text format with a good accuracy. This text data is passed on to the NLP module of the system.

3.2 Natural Language Processing

The NLP module extracts the 'meaning' and performs actions. To achieve this functionality we have implemented the GRU cells architecture.

Gated Recurrent Unit (GRU) is a variant of LSTM networks. The GRU is similar to LSTM in its basic functionality, however it has few key differences:

1. GRU does not have an output gate and also does not possess internal memory that is different from the exposed hidden state.

2. The Input and Forget gates are coupled together by an Update gate (z). The reset gate(r) is directly connected to the previous states.

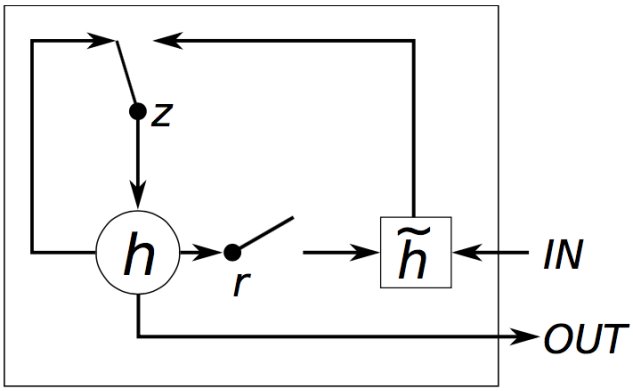

Fig -2: GRU architecture

Thus GRU have fewer parameters which leads to faster training and lesser data for generalization.

Firstly, in our module we remove any unwanted data, which might be generated due to noise in the input speech signal. For this purpose we use SPACY which is a Python Text Tokenizer. It splits the sequential input data into tokenized sentences and then into smaller tokens. SPACY also provides Part of Sentence (POS) tags for each word which are defined by the usage of the word in grammatically correct sentences.

The module consists of a word lexicon which contains a list of logical words. The Grammar block dictates the rules for processing of the module.

Secondly, we use sequence-to-sequence matching to map input statements to expected outputs. For example, sequence-to-sequence matching will map "How are you" to "Fine, thank you". In short, it trains the model by teaching it the expected response given a certain input. This can be said to be the core of the NLP output stage.

3.3 Text-to-Speech

The human speech production process first translates a text (or concept) into movements of muscles associated with articulators and speech production-related organs. Then using air-flow from the lungs, vocal source excitation signals, which contain both periodic (by vocal cord vibration) and aperiodic (by turbulent noise) components, are generated. By filtering the vocal source excitation signals by time-varying vocal tract transfer functions controlled by the articulators, their frequency characteristics are modulated. Finally, the generated speech signals are emitted. The aim of the Text-to-Speech module is to mimic this process in some manner. Thus its goal is to render naturally sounding speech signals given a textual data to be synthesized.

To achieve the above mentioned goal we again use the WaveNet architecture.

4. RESULTS

We trained the Speech-to-Text block was trained on VCTK corpus. Whereas the NLP block was trained on Reddit database.

While training the system we used CTC as an error correction parameter. The system stabilized at a CTC value below 20. The system was trained on NVIDIA GPUs for 8 days.

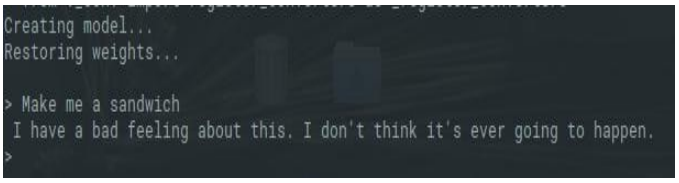

Fig -3: Result (1)

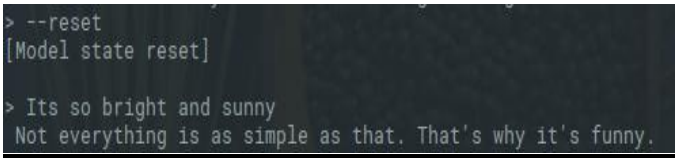

Fig -4: Result (2)

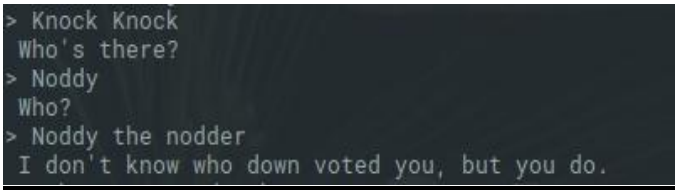

Fig -5: Result (3)

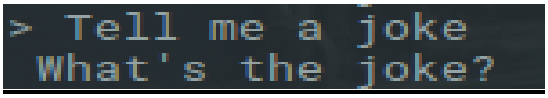

Fig -6: Result (4)

5. COUNCLUSION

We have made a completely offline, Edge Compute Voice Assistant. A careful study of existing methods and techniques has allowed us to pick and choose from available efficient techniques while crafting our version of a NLP for the Voice Assistant. As a proof of concept, our model is able to demonstrate basic functionalities of a conversation bot.

REFERENCES

[1] N. Jamil, F. Apandi and R. Hamzah, "Influences of age in emotion recognition of spontaneous speech: A case of an under-resourced language," 2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 2017, pp. 1-6

[2] Aaron van den Oord, Sander Dieleman, Heiga Zen

Karen Simonyan, Oriol Vinyals, Alex Graves,Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, “WaveNet: A Generative Model for Raw Audio”

[3] [Parikh et al., 2016] Ankur P Parikh, Oscar Tackstr ¨om, Di- ¨ panjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference

[4] A. Graves, A. r. Mohamed and G. Hinton, "Speech recognition with deep recurrent neural networks," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 6645-6649.

[5] N. Naderi and B. Nasersharif, "Multiresolution convolutional neural network for robust speech recognition," 2017 Iranian Conference on Electrical Engineering (ICEE), Tehran, Iran, 2017, pp. 1459-1464

[6] K. Greff; R. K. Srivastava; J. Koutník; B. R. Steunebrink; J. Schmidhuber, "LSTM: A Search Space Odyssey," in IEEE Transactions on Neural Networks and Learning Systems , vol.PP, issue 99, July 2016.

[7] Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups, IEEE Signal Processing Mag., vol. 29, issue 6, Oct 2012.

[8] S. Basu, J. Chakraborty, A. Bag and M. Aftabuddin, "A review on emotion recognition using speech," 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2017, pp. 109-114

[9] A Review on Speech Recognition Technique, International Journal of Computer Applications Volume 10– No.3, Nov 2010.

[10] Q. Zhang and A. Benveniste, "Wavelet networks," in IEEE Transactions on Neural Networks, `vol. 3, no. 6, pp. 889-898, Nov 1992. doi: 10.1109/72.165591.