Speech Recognition and Speech Synthesis Models for Micro Devices

Bismark Asiedu Asante[1]*, and *Hiroki* Imamura[1]

[1]Information Systems Department, Faculty of Science and Engineering, Soka University, Japan

Abstract. With the advent and breakthrough of interaction between humans and electronic devices using speech in communication, we have seen a lot of applications using speech recognition and speech synthesis technology. There are some limitations we have identified to these applications. Availability of a lot of resources and internet connectivity have made it possible in making case but with limited resources it is quite difficult to achieve this feat. As a result, it limits the application of the technology into micro devices and deploying them into areas where there are no internet connectivity. In this article, we developed a smaller Deep Neural Network models for Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) for communication on micro devices such as Raspberry Pi. We tested and evaluated the models of the system. The accuracy and the performance of the models to be implemented on micro devices shows that they are good for application development in micro devices.

1 Introduction

One of the formidable problems that have an active research going on is computer generation and recognition of speech [1]. Same can be said for the speech synthesis.

Most systems for speech recognition are often referred to as Automatic Speech Recognition (ASR) and the system for speech synthesis are referred to as the Text-to-speech systems (TTS).

A lot of organizations and institutions have found many uses for ASR and TTS in their applications. Siri® and Cortana® are being used in operating systems. Google translate® also employs this technology in translating from one language to another through speech.
ASR and TTS systems have been deployed in computers, smart phones and electronic devices that have been connected to the internet. Most of the state-of-the-art Human Computer Interactions systems for speech communications are deployed on servers accessed using Application Programming Interfaces (API) and the client access them via a request to the server. The systems achieve good results and have a very high accuracy from the responses the clients receives.

However, the API based ASR and TTS systems can be used in micro devices only when there is internet connection. Thus, implementing the API based ASR and TTS for outdoor usage without internet comes with a lot of challenges.

The APIs are only available and consumed over the internet when the request is made to the server. The computations are performed on the server and the results are presented as respond to the client that made the call.

Even though it is very good to use the API approach because the client devices uses less resources and a centralized powerful server with a lot of resource and state of the art trained models are used for the computations. Since all the clients are sending requests with data, they can be accumulated to further fine tune the outcome of the prediction of the servers.

One challenge with the API is that a constant and reliable internet to access is required.

In this work, we consider building a deep neural network models to automatically recognize speech and synthesize speech in micro devices. The deep neural network chosen for the development of the model is Convolutional Networks and Bidirectional Long Short-Term Memory (LSTM) is known as the state or the art deep learning architecture for sequence to sequence classification.

In addition to the deep neural networks, we implement some acoustic phonetic techniques for feature extraction. The audio signals are converted to MFCC spectrograms which are the magnitudes of the Short-Time Fourier Transform Modulus (STFTM) for recognition of the speech and using the fast Griffin-Lim algorithm [2] we reconstruct the audio signal from the spectrograms recognized.

We believe that the proposed models will be small enough to be implemented in micro devices such as PCI boards, micro controllers and electrical gadgets which do not have access to internet.

* Corresponding author: e18d5201@soka-u.jp

2 Model Architecture

In this paper, we introduce two deep neural networks for an Automatic Speech Recognition (ASR) and Text-to-speech (TTS) system using speech synthesis in micro devices. The main structure of the models dwells on the seq2seq models designed with bidirectional LSTM and convolutional neural networks. One of the models will be responsible for the speech recognition while the other will be responsible for the speech synthesis. Even though ASR and TTS have similar approach but opposing processes. This is due to the differences in the techniques employed at some stages of the processing and training [3]

The main approach is to have an encoder for either encoding characters to spectrograms or encoding waveforms to spectrograms.

Since there are two models, which are speech recognition model and speech synthesis model built for the ASR system and the TTS system respectively, we will give brief description of the models and their constituents in following subsection.

2.1. Automatic Speech Recognition (ASR) system

The speech recognition model is a model with transforming audio signals or waveforms into a sequence of characters representing a sentence in a given language. There are several approach to developing ASR systems but our model is isolated words type of speech recognition where the words spoken a segmented by the region of no speech or where there are no utterances.

There are a number of classes of speech recognition approach but this paper implements the Artificial Intelligence approach. The Artificial Intelligence approach combines the acoustic phonetic and pattern recognition. Based on this approach our model comprise acoustic phone model for feature extraction, a deep neural network made up of a Convolutional Layer and a Recurrent Neural Network for recognition and time distributed dense to predict the corresponding text for a given audio.

Our modelled architecture is shown Fig 1. The acoustic phonetic approach is used to generate the Mel Frequency Cepstral Coefficient (MFCC) as a feature of the wave form. The MFCC are based on the identified variations of the human ear's critical bandwidth frequencies which are under 1000 Hz [4] which is an amplitude spectrum of the sampled frequency of the wave form is passed as an input to the Deep Neural Network for classification and scoring. The Softmax classifier connected to the network is used to decode and determine the text that matched the produce or given wave form.

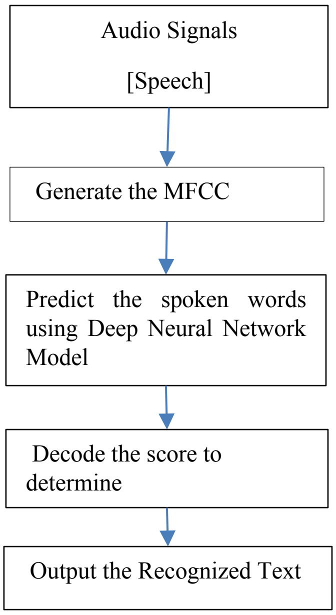
Audio Signals
[Speech]
Generate the MFCC
Predict the spoken words using Deep Neural Network Model
Decode the score to determine
Output the Recognized Text

Figure 1 Modelled flow of the Automatic Speech Recognition (ASR) system

2.2 Text to Speech (Speech Synthesis)

At a higher level, most of the speech synthesis convert sentences or a set of input characters into speech or audio signal. The process involves generating spectrogram for the given characters and then constructing them into wave forms.

With our system, the modelled approach for the deep learning model is similar to [5]. The modifications made to this model is to ensure that it runs faster with minimum resources for prediction.

The architecture of the model for speech synthesis is graphically presented in Fig 2. The flow shows that given a text as an input, the first process in synthesizing the wave form is to formulate a phoneme for the given text the phoneme is then used to predict the spectrogram for the text. The next stage involves the use of the popular Griffin-Lim reconstruction technique [2] to generate the expected wave form.

‘

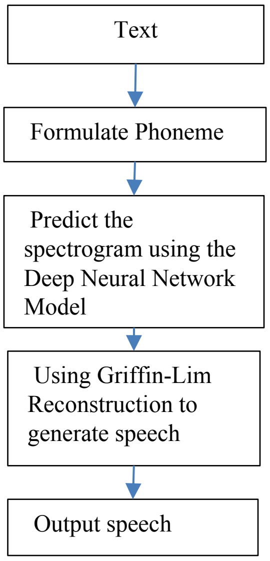
Text
Formulate Phoneme
Predict the spectrogram using the Deep Neural Network Model
Using Griffin-Lim Reconstruction to generate speech
Output speech

Figure 2 Modelled flow of the Text-to-Speech (TTS) system

3 Our Approach

In this section, we describe the approach or the method followed for the development of both the ASR and the TTS system. The systems after development would be implemented in Raspberry Pi and the performance will be evaluated.

3.2 Experimental Setup

We develop the ASR system and the TTS system using Deep Neural Networks for micro controllers like Arduino and Raspberry Pi. The resource for these devices is limited in terms of computational power and available memory for storage and computations hence the models will need to be small in size or space. Also, these models need to be highly optimized to operate with the limited resources of the micro controllers. Therefore, we need to carefully plan and implement a robust and efficient model to be deployed in the micro devices. We deployed the models on a Raspberry Pi Model 3B+ and used it to predict some recorded speech. Also, the TTS system was also implemented on the Raspberry Pi to generate some speech from the text.

In the following sub sections, we will present the experimental stages targeted at achieving the development of small model for micro devices.

3.1 Dataset Descriptions

In this work, we used the LibriSpeech [6] training the speech recognition ASR system and also for the speech synthesis in the TTS system. The LibriSpeech is collection audio books that are part of the LibriVox project that contains 1000 hours of sampled speech and their text. Also, the English read speech from audiobooks have been aligned and segment to match the corresponding book text making it easier and more useful to our research.

3.2 Feature Extraction

Feature extraction for speech signals takes a different approach than the visual. There are quite a number of feature extractors for speech processing such as Linear Predictive Coding (LPC), Mel-frequency Cepstrum (MFCCs), RASTA filtering and Probabilistic Analysis (PLDA). In this research, the MFFCs have been chosen for extracting the useful information from the wave forms as images for classification and predicting features by the deep learning models.

3.3 Training Details

To train the two networks separately, we use the LibriSpeech dataset, an open source dataset of 1000hours of read audio books with segmented and aligned with the corresponding text.

For the speech recognition, the spectrograms were generated and the corresponding text for the spectrograms were matched. This was the input for the deep neural network. The neural network comprises of an input layer with output shape of 161 dimensions. A 1D convolutional layer and a recurrent neural network for extraction of features from the MFCC. The trainable parameters of the model were 906,617 parameter with 9, 407non-trainable parameters. The network was train with 150 Epoch

For the speech synthesis, the approach proposed by [5] is implemented. The model consisted of two components. The first component is responsible for converting the given text is

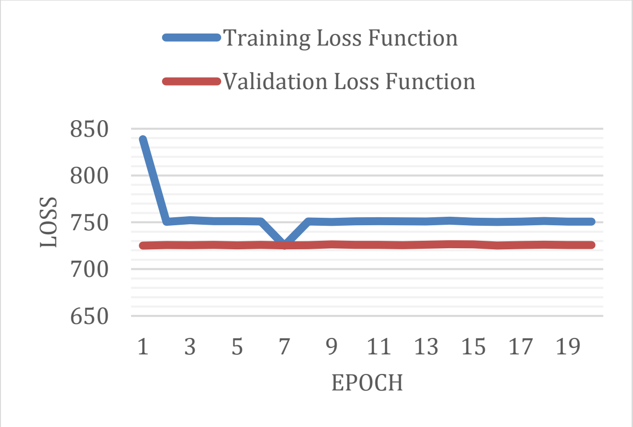

Figure 3 A graph of the training and the validation losses.

In figure 3, the validation loss is slightly better than the validation but not very appreciable and required further improvement. The improvement could be achieved by increasing the number of iterations.

After training the deep neural network, the saved model for the speech recognition is used to recognize the speech made in a given audio file.

4 Results and Discussion

In this section, we report the results from the training of the ASR system and the TTS system, implementation, testing and evaluation of the systems on a Raspberry Pi device.

4.21 Performance Accuracy of the ASR component

The accuracy and the performance of ASR system is often measured in terms of accuracy of the predicted words and how fast the systems is able to predict the words. The metric often used is called the Word Error Rate which is often computed at

$$WER = \frac{S+D+1}{N} \qquad (1)$$

Where is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the number of words in the reference.

The calculated accuracy of the speech recognition model is 60 %. This is as a result of no language model included in the prediction of the model.

4.2 Mean Opinion Score Test for the TTS component

We conducted mean opinion score (MOS) tests to rate the naturalness and the intelligibility. This approach of measuring speech quality is a modification of the Mean Opinion Scale by [7]. The modified measure which appraise the synthesis systems with the addition of psychometrics techniques.

In the determining the Mean Opinion score we conducted a survey. 10 participants were selected to listen the speech generated by our model and measuring the quality using the specify criteria for MOS. The data analysed and used to determine the MOS score for the two datasets that was used to train the models for speech synthesis.

Speech Samples	**Subjective 5-scale MOS in naturalness and Intelligibility**	
	LibriSpeech	**VCT-K**
DC + SSRN model	3.47 ± 0.094	3. 56 ± 0.093

The MOS score showed the models performed moderately as a score around 3.5 shows a fair quality of sound. We believe a that a larger dataset and more iterations of training could improve the quality of the speech synthesized

4 Conclusion

In this work, we studied the development of speech recognition and speech synthesis in micro devices with no access to internet. This prevents such systems to have access state of the arts system that use APIs since they are accessible via internet. We developed speech recognition and synthesis model using deep neural networks.

In the future, we will work on reduction of the size of the models and also make the models update with new data that is provided. The future research work will focus on further reducing and optimizing the models to achieve a higher accuracy.

References

[1] S. W. Smith, “Audio Processing,” *Sci. Eng. Guid. to Digit. Signal Process.*, no. 1, pp. 351–372, 1997.

[2] N. Perraudin, P. Balazs, and P. L. Sondergaard, “A fast Griffin-Lim algorithm,” *IEEE Work. Appl. Signal Process. to Audio Acoust.*, no. August 2015, 2013.

[3] M. Eichner, M. Wolff, and R. Hoffmann, “a Unified Approach for Speech Synthesis and Speech Recognition Using Stochastic {M}Arkov Graphs,” *Proc. ICSLP*, no. Icslp, pp. 701–704, 2000.

[4] S. Narang and M. Divya Gupta, “Speech Feature Extraction Techniques: A Review,” *Int. J. Comput. Sci. Mob. Comput.*, vol. 4, no. 3, pp. 107–114, 2015.

[5] H. Tachibana, K. Uenoyama, and S. Aihara, “Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4784–4788.

[6] P. S. K. Vassil, Panayotov; Guogo, Chen; Daniel, “LIBRISPEECH: AN ASR CORPUS BASED ON PUBLIC DOMAIN AUDIO BOOKS,” vol. 108, no. 2, pp. 581–583, 1995.

[7] M. Viswanathan and M. Viswanathan, “Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale,” *Comput. Speech Lang.*, vol. 19, no. 1, pp. 55–83, 2005.