Integrated Smartphone and Cloud Speech Recognizer

Shenghao Lu, Zhengyi Hu
Supervisor: Dr. Catherine Watson
The University of Auckland

Introduction

Producing a speech recognition system which will run on a range of smart devices requires a flexible split in the processing load between the device and cloud, which is dependent on the processing power of the device and network speed. This project is to develop the architecture of the speech processing application that can be flexibly allocated between the device and the cloud. To this purpose, an integrated speech recognition system that supports both online and offline recognition has been developed.

Voice Recognition

The common way to recognize speech is the following: we take waveform, split it on utterances by silences then try to recognize what's being said in each utterance. To do that we want to take all possible combinations of words and try to match them with the audio. We choose the best matching combination. According to the speech structure, three types of model are used in speech recognition to do the match.

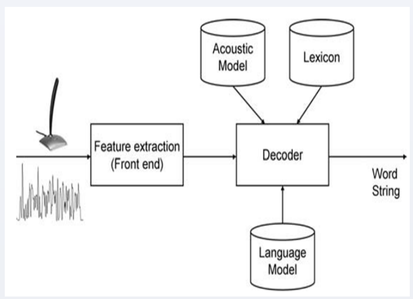

Figure 1: Speech Recognition Process

Acoustic Model: An acoustic model is a file that contains statistical representations of each of the distinct sounds that makes up a word. Each of these statistical representations is assigned a label called a **phoneme**.
The English language has about 40 distinct sounds that are useful for speech recognition, and thus we have 40 different phonemes.

•**Phonetic Dictionary:** A Phonetic Dictionary is a **mapping** from words to phones.

•**Language Model:** Language models help a speech recognizer figure out how likely a word sequence is, independent of the acoustics. This lets the recognizer make the right guess when two different sentences sound the same.

```
brogan B R OW G AH N
brogden B R AA G D AH N
brogdon B R AA G D AH N
brogna B R OW G N AH
broich B R OY CH
broil B R OY L
broiled B R OY L D
```

Figure 2: Phonetic Dictionary

Sphinx Toolkit

Sphinx is the leading speech recognition toolkit with various tools used to build speech applications. Two different Sphinx tools are used for the development of the integrated speech recognition system.

Sphinx4:

•Pure Java speech recognition library

•Provides a quick and easy API to convert the speech recordings into text

•Used to develop the server part of the system

PocketSphinx:

•A library that depends on another library called SphinxBase which provides common functionality across all CMUSphinx projects

•Used to develop the client part of the system

NZ Acoustic Model

Sphinx only provides US English acoustic models, so we have trained **our own NZ Acoustic Model**.

•Using SphinxTrain, SphinxBase and PocketSphinx as training tools

•Requires hours of speech audio in NZ accent with its corresponding transcription, a phoneset file, a phonetic dictionary and a language model.

Integrated Speech Recognition System

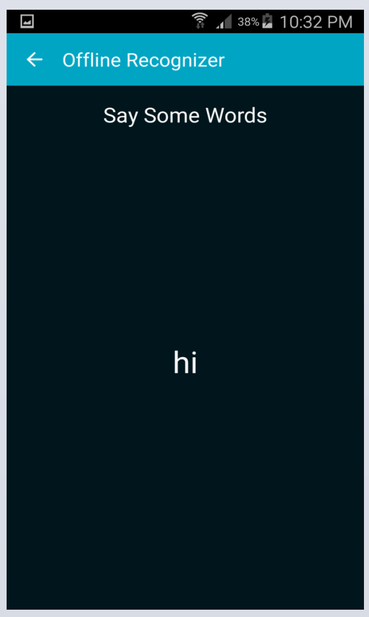

Figure 3: Offline Mode UI

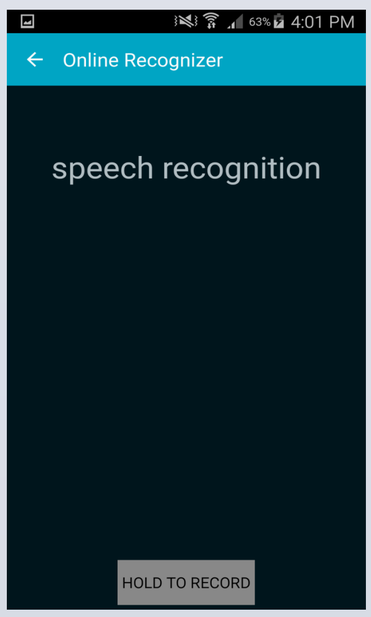

Figure 4: Online Mode UI

The offline mode is implemented using PocketSphinx. It takes voice stream as input and continuously generate transcript as new voice send into the microphone. However, online mode transmit a complete piece of recording to the server, and then send back the decoded text just for this piece of recording.

Since there is no language model in PocketSphinx, the offline mode can only recognise a single word at a time. However, the online mode has a language model on the server side, so it can recognize sentences. The reason that the offline mode does not support speech recognition is because the memory acquired to use a language model is around 2G, smartphones are not capable for this operation.

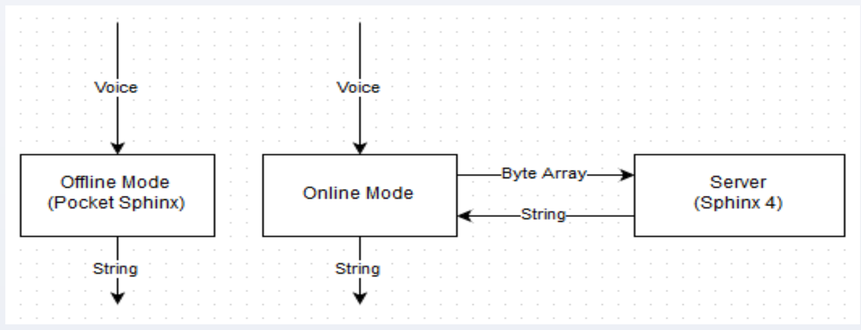

Figure 4: The Integrated Speech Recognition System

The voice recorded from the smartphone microphone is stored as a pcm file. A pcm file is a byte array, so it is transmitted to the server by using the socket. When server receives the pcm file, it is then converted into a wav file. The wav file can be recognised by Sphinx4 to transcribe the voice, then generate the text transcription. Finally, server sends the text back to the client, and the result text will be displayed on the screen.

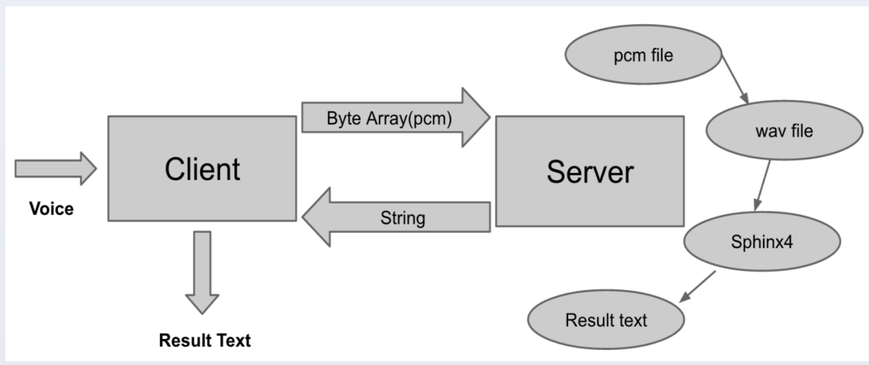

Figure 5: The Online Mode Flow Diagram

Methodology and Results

The data is generated by testing 100 times of prerecorded testing voice, and compare the result manually with the correct translation. This test has been carried out in an extremely quiet place.

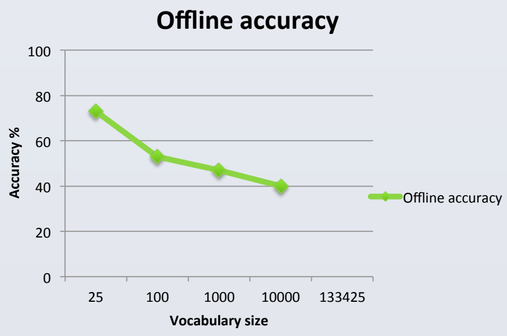

Figure 5: Accuracy analysis for offline mode

The result chart shows that the accuracy decreases as the vocabulary size increases. The accuracy of word recognition for the online mode is 71%. It is estimated that the offline mode will have the same accuracy with only 50 vocabularies. However, the online mode contains 133425 vocabularies. It clearly indicates that the online mode has a lower error rate than the offline mode.

The result chart for the online mode shows the relationship between accuracy and word size. When user speaks more word, the accuracy decreases slightly. But overall, it still maintain a high accuracy compare to the offline mode.

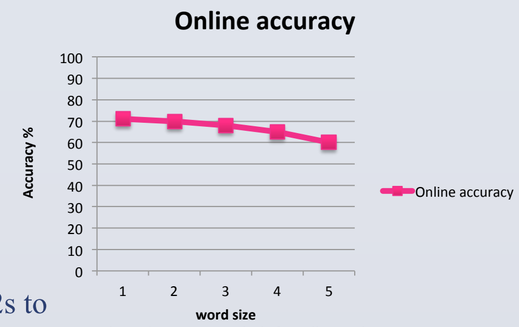

Figure 6: Accuracy analysis for online mode

Execution time for the offline mode varies from 0.2s to 0.6s depends on the vocabulary size. Execution time needed of word recognition for the online mode is around 7s and speech recognition will take a few seconds longer.

Conclusion

Generally, the Online Mode provides a better recognition accuracy while the Offline Mode has a shorter recognition time. The recognition accuracy for the two modes are close when the vocabulary size of the Offline mode is less than 50. However, due to the fact that it is still not practical to perform offline speech recognition on modern smartphones, the offline mode should only be used for word recognition. Hence, the Offline Mode is suitable for small-vocabulary word recognition where the Online Mode is capable of large-vocabulary speech recognition with a relatively long recognition time needed.

Acknowledgement

We would like to express our sincerely thanks to Dr. Watson for her continuous support to us.