應用於台灣腔中文之自動語音辨識

Automatic Speech Recognition for Taiwanese Mandarin

組別:A100　組員:莊裕嵐　指導老師:劉奕汶 教授

Introduction

Automatic speech recognition has been widely researched for recent decades. It is the technology that allows human beings to use their voices to speak with a computer interface in a way that, in its most sophisticated variations, resembles normal human conversation. This research mainly focusses on the speech recognition model accustomed to the Taiwanese accent Mandarin.

the paper provides a way to introduce and explore the professional field of speech recognition. This recognition model is not just for the research usage. In fact, the speech recognition model could be well applied into several field in our daily life to increase the common benefit for out human beings.

Methodology

- Hidden Markov Model (HMM)

Allowing us to talk about both observed events Hidden Markov model (like words that we see in the input) and hidden events (like part-of-speech tags) that we think of as causal factors in our probabilistic model

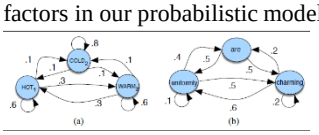

Fig. HMM interpretation

- Gaussian Mixture Model (GMM)

Successfully applied in the speech recognition field. Due to its mathematical feature, it could use a single vector to represent the probability of the word sequence. That is, it is more useful and practical than the probability simulation by the HMM method since the HMM takes more processes to simulate the probability

- Time Delay Neural Network (TDNN)

a multilayer artificial neural network architecture whose purpose is to classify patterns with shift-invariance, and model context at each layer of the network.

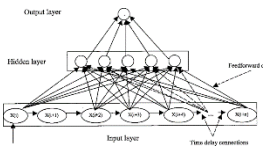

Fig. TDNN structure

Model Realization

- Kaldi

Kaldi is a state-of-the-art automatic speech recognition (ASR) toolkit , containing almost any algorithm currently used in ASR systems.. It's being used in voice-related applications mostly for speech recognition but also for other tasks — like speaker recognition and speaker divarication.

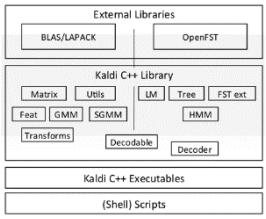

Fig. Kaldi Structure

- Dataset

Online open-source data (30hr)

Self-prepared data (8hr)

Results

Once the all processes are finished, the system would have a file recording the quality of the model building. The file would store the recognition correction rate, several key parameters in the final model and the duration of the model training. The table shows the word error rate and the total training duration.

Word Error Rate	25.14%
Training Duration	46hours45mins

Testing sentences

A02_04:晃動杯子，使沙子填滿所有的縫隙，最後又倒進水，

A02_05:做完這些，教授對學生們

A02_06:說道：「現在，我想讓大家把這個杯子理解為生活。

Testing results

晃動 杯子 使 砂子 添滿 所有的 縫隙 最後 又 到進水
作完 這些 教授 對 學生 們
說 到 現在 我 想讓 大家 把 這個 杯子 理解 為 昇華

Conclusion

In this research, I have performed how the speech recognition operate and build from scratch. More importantly, the speech recognition system is not built from some common programming languages. The Kaldi toolkit provides a more plausible and more clear instruments to implement the sophisticated model. The main difficulty is how I realize the original features of the Taiwanese Mandarin and how I could attain the recognition features applied to the speech recognition model to get a better result from the model. Despite the basic algorithm knowledge in the machine learning field, I should also review features the language embedded. In other words, this project is not only based on the mathematic operations but human beings perform the natural languages.