

Personalized Voice Recognition Program

Wisdom Boinde

Department of Computer Science, Earlham College



I. Introduction

Speaker and Speech Recognition(SSR) systems

- SSR systems enable users to communicate via voice to computer. Real world examples:
 - smart speakers, customer care centers and other speech-driven analytics
 - cybersecurity, voice identification passwords
- These systems/programs use Machine Learning processes to achieve their desired functions. The success is dependent on-
 - Datasets
 - Machine Learning model
- Thus, the performance is biased toward the data used in training

My project is a personalized Speech recognition system that recognizes only the user.

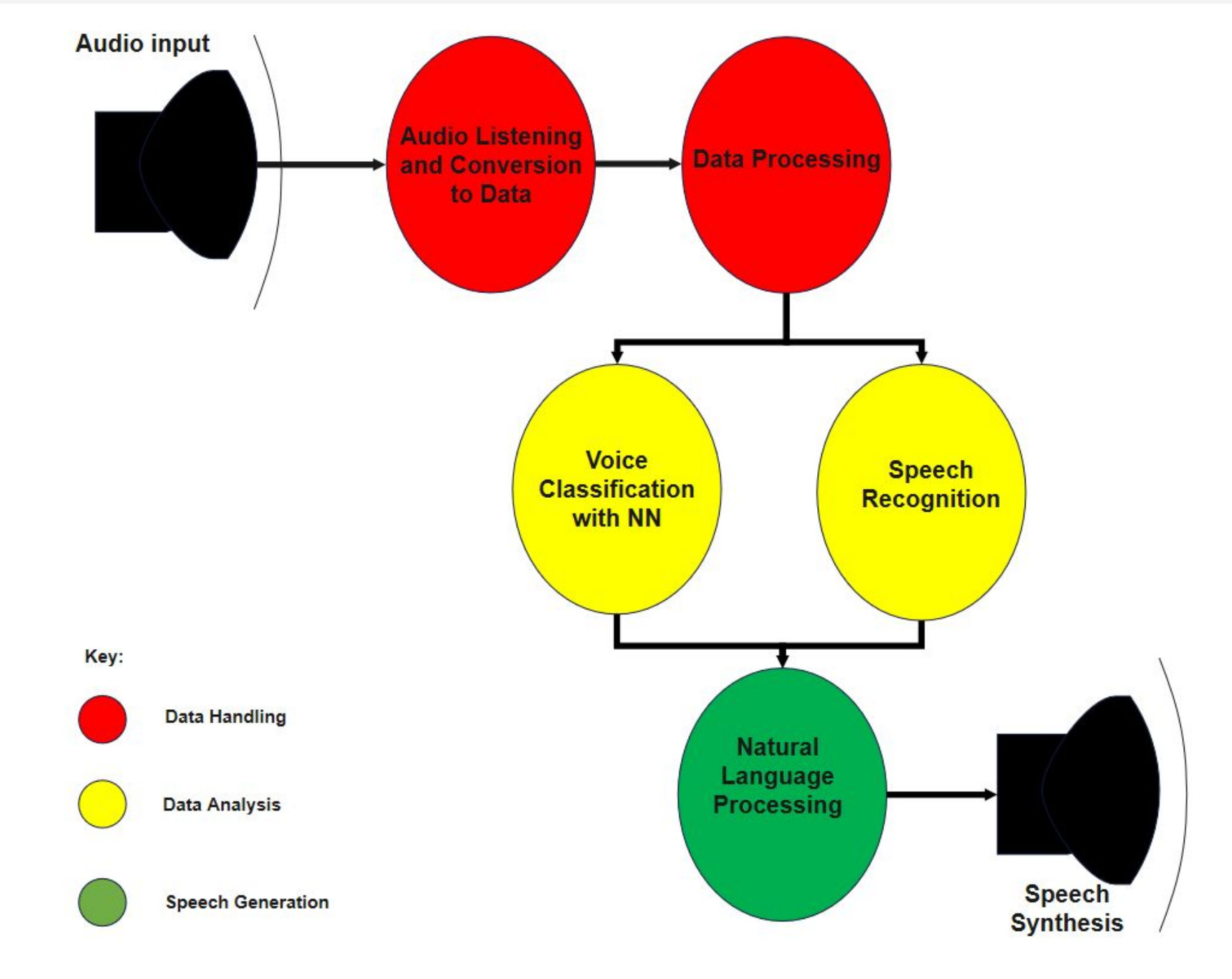
Machine Learning Process

- Deep learning is a Machine Learning technique that learns by example
- It does so via Deep Learning models which learns to perform classification tasks directly from images, text, or sound
- They just need to be trained with Data and Neural Network Architectures. My program utilizes:
 - Convolutional Neural Network
 - Recurrent Neural Network

II. Program Design and Implementation

The Flow of data in ASR system

- Sound is processed as **Audio input** to the program
- Generally, audio files are read to get the sampling frequency and the actual audio.
- The vocal tract gives out an envelope of a short-time power spectrum
- and so in instances, data for the intensity of sound can be recorded.
- Since speech is sequential, we can break the recording of it into data points along the time axis.
- To accurately represent the voice envelope, we use the Mel Frequency Cepstral Coefficient (MFCC).



Voice classification and Speech Recognition

- The Mel scale relates the perceived frequency, or pitch, of a pure tone to its actual measured frequency.
- This enables easy extraction of features of the audio file and thus, we
- feed the Mel Spectrogram into our NN as input into the **Voice classification** and **Speech Recognition**
- The Voice Classification aspect analysis features of data to identify user and the Speech Recognition aspect maps the sound features to text. Natural Language Processing is then done to make meaning of text as they are generated.
- Several Neural Network models are used in the processing of speech.
- My program has a NN for voice classification, a Convolutional NN for speech feature extraction, and a Recurring NN (Long Short Term Memory -LSTM- NN) for linguistic modeling

III. Program Evaluation

The Voice Classification report

- **Precision:** The ability of the classifier not to label a negative sample as positive.
- **Recall:** The ability of the classifier to find all positive samples.
- **F1-score:** The harmonic mean of precision and recall. It combines both precision and recall into a single metric.
- **Support:** The number of actual occurrences of the class in the dataset.
- **Weighted Average:** Useful when the class distribution is imbalanced. It provides a comprehensive evaluation of the model's performance on each class

Train Report	precision	recall	f1-score	support
0.0	0.79	0.94	0.86	265
1.0	0.63	0.29	0.40	93
accuracy			0.77	358
macro avg	0.71	0.61	0.63	358
weighted avg	0.75	0.77	0.74	358

Test Report	precision	recall	f1-score	support
0.0	0.89	1.00	0.94	16
1.0	1.00	0.33	0.50	3
accuracy			0.89	19
macro avg	0.94	0.67	0.72	19
weighted avg	0.91	0.89	0.87	19

Speech Recognition report-TBD

IV. Conclusion

- Program was precise at not mistaking wake word for a non wake work
- Program was able to predict non wake words 89% of the time
- Program detected 50% of the wake words
- Program wake detection performed well on average but can be improved

V. Future Developments

- Complete the Speech Recognition and Natural Language Processing aspect of Program
- Incorporate Speech Synthesis in Program
- Make Program Interface with device microphone
- Assemble a product that compiles and begins training when program is setup