Personalized Voice Recognition Project Proposal

Wisdom Boinde
wbboinde19@earlham.edu
Computer Science Department, Earlham College
Richmond, Indiana, USA

ABSTRACT

Speaker and Speech recognition systems are present to enable users to communicate via voice to computers. However, these systems use Machine Learning processes to achieve their desired functions. Due to the lack of diversity of datasets in all areas such as race and age, machine recognition of individual voices is biased toward the data used in training. These recognition systems are ubiquitous in several everyday applications including smart speakers, customer care centers, and other speech-driven analytics[14]. In such technology domains, questions like: How do we represent individuals from all demographic groups? What is the practical applicability of such a speaker recognition system? Is this system fair? How do we identify biases in this system?, and How might we mitigate these biases? are being addressed at a rapid pace. Making speaker recognition personalizable on a certain number of training data sets is a solution that could address most of these questions posed. To this end, my project is a personalized Speech recognition system that recognizes only the user. Based on existing open source code on Voice Assistant programs, I have implemented the wake word classification aspect of my speaker recognition which recognizes non-users all the time and the user 50% of the time.

1 KEYWORDS

Neural Network, Deep learning,Recurring Neural Network, Voice Classification,speaker Indentification, Speech recognition, Automatic Speech Recognition.

2 INTRODUCTION

Voice Recognition or speaker identification refers to identifying the speaker, rather than what they are saying.Regardless the two are quite related and recognizing the speaker can simplify the task of translating speech in systems that can be trained on a specific person's voice.

The Voice Recognition Market was valued at $9.56 billion in 2021. The global speech and Voice Recognition market was then projected to grow from $11.21 billion in 2022 to $49.79 billion by 2029 at a CAGR of 23.7% in the forecast period[1]. As we make our lives smart, so is the demand for Voice Recognition to make interacting with smart life easier.

In other domains such as cybersecurity, voice identification passwords are more secure than traditional passwords. Voice Recognition is also useful in situations where users are handicapped or unable to use the traditional typing and clicking method of communicating with smart devices.

Based on the demand and applicability of Voice Recognition, we get a measure of its importance in our daily lives as we advance technology.

Numerous things can interfere with Voice Recognition software. These could include Voices in the Background, Speed of communication, Distance from the microphone, and similar-sounding words. Another problem, one which led me to this project is the bias in speech recognition and the difficulty in personalizing software.

Thus I address the question, *How can I make speech recognition personalized?*

2.1 Automatic speech recognition(ASR)

The aim of ASR is to transcribe human speech into spoken words. This is based on identifying various speaker attributes. It also maps variable lengths of speech to varying lengths of words. ASR has advanced due to the advancement of deep learning and big data.

Some speech recognition systems require "training" (also called "enrollment") where an individual speaker reads text or isolated vocabulary into the system. The system analyzes the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy. Systems that do not use training are called "speaker-independent" systems. Systems that use training are called "speaker dependent".

Most ASR systems are statistical. Acoustic modeling and language modeling are two important parts of modern statistically-based speech recognition algorithms. The acoustic model is used to relate the audio signal to the phonemes. Here, the model learns from sets of audio recordings and their corresponding transcripts. Language Modeling uses a probability distribution over sequences of words. Given any sequence of words of length m, a language model assigns a probability to the whole sequence.

To fully personalize a speech recognition system, we need the system to respond only to one voice or a set of voices. To do this, speaker identification has to be embedded in the ASR system to classify sounds into user vs non-user of the system.

2.2 Speaker identification Systems

Speaker identification software uses audio analysis techniques to identify and verify the identity of a speaker.No two individuals sound identical because parts of their voice production organs are different. In addition to these physical differences, each speaker has his or her characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, pronounciation pattern, choice of vocabulary and so on[10]. With the Speaker Identification systems, there program usually involves an enrollment and recognition process. During the enrollment process, features of the speaker are extracted and used to train a speaker model. In the recognition process, we take features of an unknown, and compare it to the speaker model's database to give a similarity score. The similarity score is then used to make a decision about the identity of the unknown speaker.

Thus, making a software capable of extracting and recognizing the feature vectors of any target speaker(user) can make speech recognition and personalization easy. The features will have to be extracted from a dataset which will be used in training and testing process of program.

The rest of the paper talks about the dataset, Neural Networks and Related Work, as well as the design, implementation and evaluation of the program.

3 DATASET

What a NN can learn depends on the data used to train it. For the purpose of this work, we need the Speech Recognition System to learn nuances of speech. Common Voice, an open-source speech data set initiative led by Mozilla, will be used as a dataset as it represents a variety of ages, gender, and so on. The data set used is the Common Voice Delta Segment 12.0. It has a size of 1.22 GB with 63 recorded hours, 64 Validated Hours, and 1,152 voices.

Besides Common Voice as a dataset, recordings of my voice and ambient noise will also be used to train the model. Since we want the system to be personalized and we want correct representations of the accuracy of the Voice Classification NN, I have trained the model with 200 2 seconds data files for 0 classification and 100 2 seconds data files for 1 classification. This ratio is good enough to keep all non-users out and give only user access to the program. Also, 2 seconds is the average time it should take for the user to call the wake word of the program. It should also be the maximum length of audio utterance needed for the program to identify the speaker.

4 NEURAL NETWORKS AND RELATED WORK

Neural Networks is a structural program that is made up of nodes and layers that interact to receive inputs and work toward a desired output. The outmost layers are the input and output layers and the middle layers are what we call hidden layers. These hidden layers have activation functions that takes in input and provides an output based on the activation function type. It could be sigmoid function, softplus, or ReLU function. The process by which an input reaches the activation function is guided by weights and bias which influence the range for the activation function. In general, NN use weights and bias generated from a method called back propagation to stretch and fit data inputs to a certain range in an activation function.

4.1 Deep Learning

Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans: learn by example. It is the key to voice control in consumer devices like phones, tablets, TVs, and hands-free speakers. In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance.Most Models are trained by using a large set of labeled data and neural network architectures that contain layer(s).

Several Neural Network models are used in the processing of speech. In my speech processing system, there is a Convolutional NN for speech feature extraction, and a Recurring NN (Long Short Term Memory -LSTM- NN) for Speaker Identification and linguistic modeling. Pytourch, as my Deep Learning framework of choice, will be used to implement the Speech Recognition model.

4.2 Convolutional Neural Network

A CNN convolves learned features with input data, and uses 2D convolutional layers, making this architecture well-suited to processing 2D data. A convolutional neural network is more specific in the tasks that it accomplishes. The CNN is designed to have multiple layers; including a convolutional layer, non-linearity layer, pooling layer, and fully-connected layer. CNN has an excellent performance in machine learning problems. Especially the applications that deal with image data, such as the largest image classification data set (Image Net), computer vision, and natural language processing (NLP).

Convolutional Neural Networks are relevant to my work because CNNs are primarily used to solve difficult image-driven pattern recognition tasks and with their precise yet simple architecture, offer a simplified method of getting started with Artificial Neural Networks. Thus, in the context of my work, CNN will be used to analyze spectrograms generated from audio data for easy classification and phonemes recognition. This will be in the speech recognition aspect of the program.

4.3 Recurring Neural Network

Recurrent neural networks, also known as RNNs, are a class of neural networks that allow previous outputs to be used as inputs while having hidden states. RNNs have a long history of applications in various sequence learning tasks Werbos (1988); Schmidhuber (2015); Rumelhart et al. (1985). RNNs have the possibility to process an input of any length, their model size not increasing with the size of the input, and their computation takes into account historical information, and weights shared across time. Long Short-Term Memory NN, a variant of RNN can in principle store and retrieve information over long time periods with explicit gating mechanisms and a built-in constant error carousel.This makes them capable of learning order dependence in sequence prediction problems. Due to these features, LSTM NN will be the layer for language processing and speaker identification.

5 DESIGN AND IMPLEMENTATION

5.1 Data Processing Pipeline

As shown in fugure 1, while the program runs, it will listen for the wake word or the uttered speech. The program receives audio input from the microphone and analysis the audio data to make meaning out of it. Since the data input is in the form of sound, numerous nuances exist to be considered in analyzing the data. For example, a word could be said in different pitches(frequencies), intensities(loudness), and lots of background noise, which makes it difficult to discern the truth from the data. Thus, we have to extract the important features and discard the noise. Generally, audio files are treated as wav or mp3 files and while reading an audio file we get the sampling frequency and the audio waveform. The properties of audio could be divided into two properties, the physical properties, and the linguistic properties.

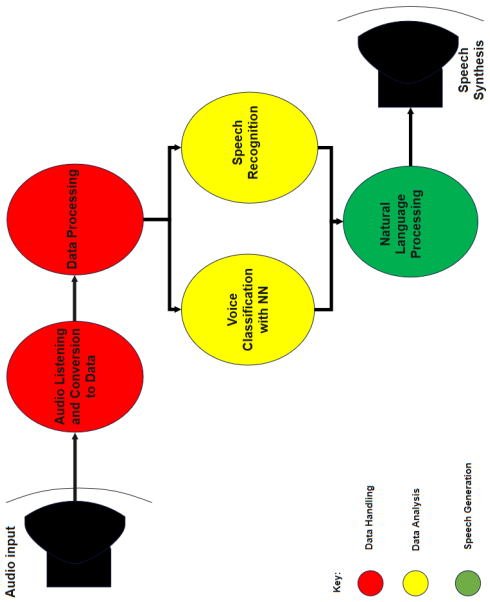

Figure 1: Data Architecture Diagram of my personalized ASR system.

As per the physical properties of sound, we produce different intensities of sounds at different frequencies. The human vocal tract gives out an envelope of a short-time power spectrum and so in certain instances, we can have data for the intensity of sound produced. Since speech is sequential, we can break the recording of it into data points along the time axis. To accurately represent the voice envelope, I use the Mel Frequency Cepstral Coefficient (MFCC). 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, I feed the Mel Spectrogram into our NN as input.

An Acoustic Neural Network model handles the physical properties of sound in my program. Here, the Neural Network model takes in an audio file input and extracts the features. Analyzing the features in very short intervals of sequential data, we can store the data of the phonemes of the audio. As phonemes are the smallest units of sound in a language, the ASR transcription is more similar to the correct utterance at the phoneme level than at the character or word levels.

During the acoustic model implementation, the audio is converted into samples of probable texts for linguistic analysis. For a personalized ASR system, we need our models to be lightweight in terms of memory and computing and we want them to run on everyday consumer machines. Since we are dealing with feature extraction and sequential data, both a convolutional NN and a Recurring NN(LSTM) is used for the acoustic model.

For processing the linguistic properties of sound, the raw samples of probable texts of phenomes are processed in the context of linguistics to make sense of text samples.

Since there exist pre-build systems of NN for speech recognition, we can utilize these NN models for speech recognition and language processing. Thus, to personalize the system, during data processing, we need to perform a speaker identification with NN to determine the validity of the speaker. This is particularly where my senior capstone project is mainly based.

5.2 Speaker Identification Neural Network Model

Several Neural Network models are used in the processing of speech. In my speech processing system, there is an LSTM for Voice Classification(speaker identification). For the speech recogniton, the model is based on a speech recognition system consisting of a CNN layer, two dense layers, a Long Short Term Memory (LSTM) NN variate of an RNN, and an extra dense layer with a softmax activation classifier. Between each layer, there is layer Normalization, Gelu Activation, and Dropout to make the network more generalizable. These models for speaker identification and speech recognition are modified from an open source Voice Assistant Program[11].

The speaker identification was implemented via Keras. Keras is a deep learning API written in Python, running on top of the machine learning platform TensorFlow. Keras is used to build machine learning models. Keras' models offer a simple, user-friendly way to define a neural network, which will then be built for you by TensorFlow. TensorFlow is an open-source set of libraries for creating and working with neural networks, such as those used in Machine Learning (ML) and Deep Learning projects. Keras is perfect for those that do not have a strong background in Deep Learning but still want to work with neural networks. Using Keras, you can build a neural network model quickly and easily using minimal code, allowing for rapid prototyping.

I used the sequential class of the Keras Models API to help create and train a binary classification RNN for my wake word detection. The sequential class is used to create a layer for Mel Frequency Cepstral Coefficients(MFCC) and spectrum augmentation. The spectrum augmentation takes out portion of the spectrum for increased real world application. This NN created by the sequential class takes in the waveform and converts it into an MFCC. This MFCC is then fed into the speaker recognition model to be classified. The model assigns 1 to "wake" and 0 to "not wake" from the sea of sound. The model has two NNs. An LSTM and a linear NN. Input and hidden layer information are passed onto the LSTM NN and the output of this is then fed into the classifier(linear NN) for a final output of either 1 or 0.

6 RESULTS AND CONCLUSION FOR SPEAKER IDENTIFICATION NN

Precision of the classifier is the ability not to label a negative sample as positive. The Recall is the ability of the classifier to find all positive samples. The F1-score is the harmonic mean of precision and recall. It combines both precision and recall into a single metric. The support gives the number of actual occurrences of the class in the dataset. Finally, the weighted average is useful when the class distribution is imbalanced. It provides a comprehensive evaluation of the model's performance on each class. The Speaker Identification in the wake work model is analysed based on the metrics defined above the yield the following output:

Table 1: Train Report

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

Table 2: Test Report

Train 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

We see from the report that the program was 100% precise at not mistaking wake word for a non wake work. It was able to predict non wake words 89% of the time. The program detected 50% of the wake words. On average, the wake word detection performed well for a first time implementation. It can however be improved based on number of hidden layers, and better datasets.

7 FUTURE DEVELOPMENTS

To the end of a complete program I still need to implement the Speech Recognition and Natural Language Processing aspect of the program, incorporate Speech Synthesis in program, make the program interface with device microphone, and assemble a product that compiles and begins training when program is setup with given arguments.

To improve the speaker classification, I will find the optimal hidden layer number, number of data sets, and the activation function type to train the program to an optimal output.

REFERENCES

[1] 2023. Speech and Voice Recognition Market Size: Report [2029]. https://www.fortunebusinessinsights.com/industry-reports/speech-and-voice-recognition-market-101382

[2] ActiveState. 2022. What Is A Keras Model. https://www.activestate.com/resources/quick-reads/what-is-a-keras-model/#:~:text=Keras%20is%20a%20neural%20network,built%20for%20you%20by%20TensorFlow

[3] Google AI. 2020. How speaker recognition works. https://www.youtube.com/watch?v=CqOfi41LfDw

[4] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. 2017. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*. 1–6. https://doi.org/10.1109/ICEngTechnol.2017.8308186

[12]]https://stanford.edu/ shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks Afshine Amidi and Shervine Amidi. [n. d.]. CS 230 - Deep Learning. https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks

[6] Practical Cryptography. 2023. Mel Frequency Cepstral Coefficient (MFCC) tutorial. http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[7] Indra den Bakker. 2017. *Python Deep Learning Cookbook*. Number 9781787125193. Packt. 330 pages. The Python Deep Learning Cookbook presents technical solutions to the issues presented, along with a detailed explanation of the solutions. Furthermore, a discussion on corresponding pros and cons of implementing the proposed solution using one of the popular frameworks like TensorFlow, PyTorch, Keras and CNTK is provided. The book includes recipes that are related to the basic concepts of neural networks. All techniques s, as well as classical networks topologies. The main purpose of this book is to provide Python programmers a detailed list of recipes to apply deep learning to common and not-so-common scenarios..

[8] Anjie Fang, Simone Filice, Nut Limsopatham, and Oleg Rokhlenko. 2020. Using Phoneme Representations to Build Predictive Models Robust to ASR Errors. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 699–708. https://doi.org/10.1145/3397271.3401050

[9] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and Understanding Recurrent Networks. https://doi.org/10.48550/ARXIV.1506.02078

[10] Tomi Kinnunen and Haizhou Li. 2010. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication* 52, 12 (2010), 1059–1072.

[11] LearnedVector. 2021. A-Hackers-AI-Voice-Assistant. https://github.com/LearnedVector/A-Hackers-AI-Voice-Assistant/blob/master/VoiceAssistant/speechrecognition/neuralnet/model.py

[12]]medvedovsky Vlad Medvedovsky. [n. d.]. A DETAILED GUIDE TO CREATING A VOICE RECOGNITION APPLICATION. https://proxet.com/blog/a-detailed-guide-to-creating-a-voice-recognition-application/

[13] Keiron O'Shea and Ryan Nash. 2015. An Introduction to Convolutional Neural Networks. https://doi.org/10.48550/ARXIV.1511.08458

[14] Raghuveer Peri, Krishna Somandepalli, and Shrikanth Narayanan. 2023. A study of bias mitigation strategies for speaker recognition. *Computer Speech Language* 79 (2023), 101481. https://doi.org/10.1016/j.csl.2022.101481

[15] PyTorch. 2021. torch.nn.LSTM. https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html

[16] sentdex. 2017. Speech Recognition with Python Introduction. https://www.youtube.com/watch?v=AsNTP8Kwu80&t=42s