Speaker Recognition using Neural Networks and MFCC Features

Hamza Zamani1, Taishi Kato1, David Rosenwasser1

1University of California, Los Angeles

hwzamani@g.ucla.edu, taishikato10@g.ucla.edu, dbrosenwasser@gmail.com

Abstract

This paper provides an introduction to the challenges of speaker recognition and offers an approach to form a text-independent speaker verification model. Described herein is an approach implementing mel-frequency cepstral coefficients (MFCC) as the main element of the feature vector to a K-Nearest Neighbors (KNN) algorithm in order to develop a speaker recognition model. Two sets of training data were trained against three different cases. The first set of training data set consisted of various speakers reading the same sentence (text-dependent) while the second set consisted of participants speaking in casual speech (text-independent). By obtaining a high identification rate in all cases, the model would provide evidence to being a considerably robust speaker recognition system. Under the optimized methods and parameters discussed herein, the model obtains an accuracy rate of <insert accuracy rates>

**Index Terms**: speech recognition, mel-frequency cepstral coefficients, neural networks, knn, feature extraction

# Introduction

This section provides an overview of the speaker recognition goal and challenges in addition to the theory behind the methods implemented in this project. The speaker recognition model will be given two speech segments and must determine if the samples are spoken by the same speaker or not. The feature vectors of the two samples will be extracted identically and the machine learning implementation will predict if they are the same speaker based on the model formed from the training data.

The ability to determine if a voice pattern is identical to another is useful for various purposes including authentication and surveillance speech recognition. Identification based on voice patterns can be used for password protection or creating vocal signature. Vocal signatures can be used to lock or unlocked access to areas of interest. Alternatively, a known voice pattern can be filtered through a large amount of speech data to target a unique person of interest for various purposes, like recovering missing persons or targeting military foes. [applications of speaker recognition, singh].

## Speaker Recognition Challenges

Speech recognition is an extremely popular field of speech processing and has vast amounts of research and data that can be provided for analysis when compared to speaker recognition, though many of the same principles and methods still apply. The dataset the speaker recognition model will be using provides challenges because the model must fit both text-dependent and text-independent cases. The text-dependent set consists of various female speakers reading the exact same sentence while the text-independent set involves those same speakers participating in unique casual speech (phone).

Additionally, the data is not a closed set meaning that some speakers in the testing phase may not necessarily have been part of the training phase which may dilute the classification accuracy. All of the data samples are within three to five seconds in length, which provides a limited amount of data that can be used for the model to make a determination. As detailed later in this paper with the Neural Network implementation, it is difficult to provide a model that is robust and accurate in both cases without a closed data set.

<any other challenges?>

## Mel-Frequency Cepstral Coefficients (MFCCs)

The linear time invariant (LTI) model of speech production is presented below in Figure 1.

Source

Transfer Function

Radiation

Sound

Figure 1: *LTI Model of Speech Production*

The cepstral domain is extraordinarily valuable for speech and speaker recognition due to its ability to separate the source, transfer function, and radiation effects of the speech signal being produced. This paper will use Mel-frequency cepstral coefficients (MFCCs) to obtain a representation of the transfer function of the speech signal under analysis. The transfer function contains a representation of the vocal tract information that is unique to the speaker and can thus be used for speaker identification purposes. [speaker identity and voice quality] [a new set of features for text-independ…].

The MFCC have become firmly established as an excellent feature vector for speech and speaker recognition problems. The MFCC approach is a frequency analysis based on a filter bank with approximately critical band spacing of the filters and bandwidths [book]. The MFCC is modeled after the human auditory system which contains proportionally large and narrow filters at the lower end of the frequency scale than the higher end. Because of this, the unit of the MFCC is the mel, which is a warped frequency representation of MFCC with the conversion shown below in equation (1).

(1)

Warping the frequency scale into mel scale allows for improved resolution at lower frequencies which contains the unique characteristics of speech that will be used as the feature vector.

## K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (K-NN) algorithm is a method used for classification of various systems. A typical binary K-NN classification is demonstrated in Figure 3.

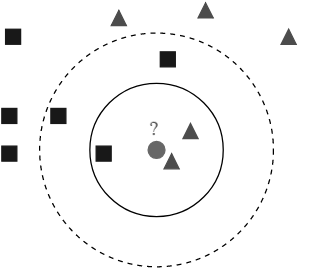


Figure 3: *Typical K-NN*

K-NN is a supervised learning approach, meaning that it takes the binary labels as inputs to develop a model for the purpose being implemented. Each feature vector is mapped across the K-NN feature space during the training phase as indicated by the square the triangle shapes in Figure 3. The solid and dotted circle corresponds to different scanning distances of the K-NN that will be used for the classification of subject under test, demonstrated by the circle. The distance between each shape and the test point is measured and the algorithm simply counts the shapes that are within the scanning range specified by the solid and dotted circles. The shape with the highest count in the respective scanning range is considered the shape that the subject under test belongs to. [source?] In section 2.3, this paper discussed the uses of K-NN for speaker recognition.

# Methodology

This section describes the methodology forming the speaker identification system. Computations were performed with MATLAB and functions from the Voicebox catalog were implemented in this project [cite voicebox]. The general block diagram of the speech diagram described in this project is shown in Figure 1. For our specific project, the features of every audio signal were first extracted and stored in a feature dictionary, but the result is the same.



Figure 1: *Block Diagram of Speech Recognition System*

The feature vector consisted mainly of Mel-Frequency Cepstrum Coefficients which were then fed into the K-Nearest Neighbors (K-NN) machine learning implementation to produce a model that can predict test data. The scoring was calculated using an Equal Error Rate (EER) which is the percentage of error when the threshold of your scoring function is set such that the False Positive Rate (FPR) is equal to the False Negative Rate (FNR).

## Pre-emphasis

During the pre-emphasis phase, a filter with numerator coefficients 1 and -.95 is applied to each audio sample. The pre-emphasis filter attempts to attenuate the lower frequencies and boost the high frequencies in an attempt to normalize the signal. Through various trial runs of the data, the pre-emphasis filter was determined to be beneficial in achieving a higher accuracy rate.

## MFCC Feature Extraction

This section describes the technique for extracting the MFCC coefficients from each audio sample and forming the feature vector used as the input to the classifier. Equation (2) describes the process for obtaining the MFCC.

To obtain the MFCC, the samples must first be divided in overlapping segments as described in Figure TBD in order for performing sampling of the signal. The segment under analysis, or frame, is multiplied by a Hamming window in the time domain before the Fourier transform is taken. <should I say how many frames per sample or the frame rate?> Next, the logarithm of the Fourier spectrum is computed before the mel-scale filter bank analysis is performed. The logarithm of the Fourier transform of a signal transforms the representation in the cepstral domain which, as described previously, offers valuable properties for speaker recognition because the transfer function contains the unique vocal tract properties and can be separated from the source and radiation elements. The mel-scale filter bank is demonstrated below in Figure TBD which is modeled after the human auditory system in the form of triangle filters with approximately critical band spacing of the filters and bandwidths [book]. The highest and lowest filters were tapered down to zero because the ends of the sampling produce higher error rates as they overlap with a zero value.

*f*

(k)

Figure TBD: *Mel-Scale Filter*

The filter bank analysis produces the cepstral energy in each channel representing different frequency bands. Finally, the discrete cosine transform (DCT) is performed on the filter-bank to improve the model efficiency as the sines are superfluous [speaker recognition by machines and human]. Twelve MFCCs were extracted from each frame to create a frame vector. One property of the cepstral coefficients is that the values approach zero rapidly, therefore no more than twelve coefficients were needed to produce meaningful values [reference for dis?]. Various test runs yielded nearly identical EER for twelve and twenty MFCCs and the less coefficients the faster the computations were performed. The mean across the frames were computed to form a single vector representation of the entire signal under test as described in Figure TBD. The standard deviation was calculated across the signals 12 MFCC coefficients and concatenated with the feature to form a 13 element feature vector per audio sample. The addition of the standard deviation improved the EER as the feature vector was represented in a more dynamic way. The standard deviation is a measure of the extent of the deviation of the MFCC coefficients as a whole, and logically this can contribute to the speaker recognition model as one speaker may have more variations in their speech than another speaker. The variations in speech become a useful parameter for the model to classify between speakers.



Figure 2: *Feature Extraction of MFCC*

The initial baseline of the project implemented using the pitch of the speaker as the main element of the feature vector, yet this approach yielded minimal success. The data set was entirely consistent of female speakers, thus the pitch does not possess as much variability as it would between males and females and produced a worse than 50% accuracy rate. The average pitch was concatenated with the MFCCs to provide additional features for the machine learning approach to compare against, yet empirical evidence produce no data to support its usefulness. Further, the calculation of the pitch performed on every frame of every sample was computationally expensive and produced no measureable improvements in ERR so it was struck from the feature vector.

During trial runs, the delta and delta deltas of the MFCC were implemented to increase the feature length and provide the model with more data to compare. The delta and delta delta MFCCs provide a dynamic representation of the signal as it passes through the frames. The delta and delta delta MFCCs track formants which are valuable for a speech recognition problem but did not provide a use for the speaker recognition problem that this paper presents. The results adding the extra elements to the feature vectors were nearly identical to those without them, so for computational reasons this paper did not include them.

## K-Nearest Neighbors

After the feature vectors for each signal are prepared, the machine learning approach attempts to determine if two speakers are the same or not. Figure TBD above describes a typical binary K-NN implementation which maps two classes across a feature space before a test point is introduced. The distance between the test point and the class representations are measured and the K-NN classifies the test point according to the count of the class representations in the scanning distance.

For our purposes, we only need to validate if two samples are spoken by the sample speaker so the feature space can be thought of as having one shape and one test point. The K-NN classifier was optimized in terms of neighbors and the distance method used. Through empirical analysis of the EER results, the ‘OptimizeHyperparameters’ parameter in MATLAB was chosen to be implemented. The standard distance used to measure the speaker under test to the other speaker models was calculated using a standard Euclidean geometry while the number of neighbors (dotted circle) was optimized to seventy-five. From the parameters used, if the scanning range of the test point does not detect the other speaker model the K-NN determines that the two speakers are not the same. Conversely, if the speaker model is within the scanning range of the speaker under test, the K-NN determines the two speakers are the same. <give results here or in conclusion?>

A Neural Network was experimented with, but due to empirical analysis of the EER results, the K-NN algorithm was determined to produce the most accurate results for our purposes. The Neural Network Implementation is described in Section 3.1. <or should it be described in 2.4?>.

# Discussion and Alternate Approaches

This paper presented an approach to implement MFCCs as the main element of the feature vector to a K-Nearest Neighbors (KNN) algorithm in order to develop a speaker recognition model. Each audio sample was represented by a 13 length feature vector consisting of twelve MFCC coefficients and the standard deviation of those values. The feature vector of was fed into a K-NN implementation in order to make a determination if the speaker under test is identical to the speaker provided. The results provided a <insert accuracy rates> success rate.

During the development of the speech recognition model presented herein, various approaches were explored that were not implemented in the final submittal. The following sections will describe and discuss alternate approaches that were researched to obtain the optimum speaker recognition model and ideas for future work.

## Neural Networks

Recent proliferation of neural networks in both academia and industry have made usage of the method widely available and simple to implement for research purposes. The popularity of the neural network has helped discover various applications of these algorithms in interesting and diverse fields, including speech and speaker recognition.

Modeled after the human neural system, neural networks are capable of classifying data in high dimensional features spaces. Rather than statistical pattern matching, neural networks invoke supervised learning to develop complex functions to model behavior and provide universal function approximation [ref1].

### Neural Network Architecture

Though implementation of a neural network is relatively straightforward, there is no method for determining *a priori*, or the optimal architecture, without human guidance. Parameters such as the number of hidden layers, neurons per layer, and more, are all subject to user tuning and is highly dependent on the nature of the input data. These parameters in addition to the algorithm settling at a local minima [ref2] can make optimization difficult and time consuming. The Figure below describes a basic model of a Neural Network.



Figure 1: *Binary classification neural network with a feature vector of 4 parameters with a single hidden layer.*

For the training sets discussed in this project, a network consisting of three hidden layers with thirty neurons each produced the best EER results out of the number configurations tested. The Figure below shows a representation of the implemented neural network.



Figure 2: *Block diagram representation of implemented network*

The selection of the input layer size was selected based on the feature vector chosen for testing. As described earlier, the feature consisted of twelve MFCC coefficients per audio sample concatenated with the standard deviation. Next, the hidden layers consisted of three 30 neuron layers until finally a single neuron output layer was configured for the binary classification of the speaker data. A single neuron output layer was chosen over a two neuron layer due to computation efficiency and insignificant changes in results.

### Training

Implementation of a neural network for speech recognition has been researched extensively but many involve classification of text-dependent speech [ref 3] or text-independent speech with a closed speaker set, some achieving 100% classification accuracy [ref 4]. Other techniques include using a convolutional neural net to classify the spectrogram of a text-dependent closed set utterance [ref 5]. The unique challenge for this project is the combination of text-dependent and text-independent sample sets coupled with an open speaker set for the final classification.

An additional constraint implemented during Neural Network training is the ratio of excitatory classes in comparison of the overall training set. Table [tbd] shows the ratio of training vectors that have zero labels in comparison to vectors containing one label.

Table 1 : *Ratio of excitatory targets in read and phone training sets*

|  |  |  |  |
| --- | --- | --- | --- |
| Training | Vector Size | Excitatory | % Excitatory |
| READ | 9730 | 147 | 1.51 % |
| PHONE | 11175 | 150 | 1.34 % |

The nature of the training set allows for the neural network to classify all vectors as non-match speakers thereby artificially achieving a classification accuracy of 98.49% and 98.66% for read and phone training sets respectively given the training set. In reality, neural nets perform relatively poor given any excitatory test conditions. The issue of obtaining enough training data of all class types is an added difficulty of correctly training the classifier. Though not implemented for this design, there are methods to balance the training set which are further detailed in the next section.

The Receiver Operating Characterstic (ROC) plot and confusion matrix for the net trained with read data are shown in Figure tbd and Figure tbd respectively while the net trained with the phone data set are shown in Figure tbd and Figure tbd.

<figures>

It can be seen by the figures presented above that due to the lack of balance in classes, both confusion matrices deceptively report exceptionally high performance. This miscalculation is proven by simply inputting a feature vector that corresponds to a matching speaker set. Notice for both outputs the misclassified values almost exactly match the number of matching classes presented in the read and phone training sets.

### Neural Network Results

After implementation and tuning of the network, the results achieved were subpar in comparison to the K-NN implementation and were therefore disregarded for submittal. The calculated EER of the Neural Network model is shown below in Table tbd.

Table tbd : *Neural Network EER*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Read | Phone | Mismatch |
| Train Read | 30.98% | 29.94% | 46.67% |
| Train Phone | 21.67% | 26.67% | 47.59% |

Additionally, since the network must be trained in order for it to classify the run times of the design as a whole, the implementation negatively impacts the metrics used for selection of a neural net for the finalized approach. As the feature set grows, the network takes longer to converge. Run times for both training sets are displayed in the Table below. The fast run times suggested by the table provide evidence the neural network is not training properly to the data presented.

Table tbd : *Mean Script Runtime for Neural Network*

|  |  |  |
| --- | --- | --- |
|  | Read | Phone |
| Mean Runtime (sec) | 16.54 | 18.01 |

Additional challenges were encountered simply due to the nature of the training sets and structure. The project guidelines specified two separate training tasks and the neural network architecture that performed well with the read training set would not necessarily perform well with the phone training set and vice versa. This forces the architecture to be mediocre for both training sets as opposed to being optimized for one.

As mentioned in the previous section there are methods to balance the training data to include a more even distribution of classes to allow for proper training of the neural network. Other than collecting more matching cases future implementations could balance the data by pruning the overrepresented class, i.e. speech instances where the speaker is not the same, while keeping all the set where the speaker is the same [6]. Another possible method would be to artificially inject noise, or slightly perturb existing sets of data (with matching speakers) and add those to the training set, or more simple use mean shifting MFCC values

Though the KNN implementation outperformed the neural network given the project conditions, with an enhanced trainset and further tuning it is plausible that the neural network could outperform the KNN especially if enhanced high-dimensional features are extracted from the speech data.

## If I need more section

If I need more I’ll write some stuff about the following and say we wanna do this for future research. But we already have 5-6 pages!:

- Seeing how much improvement you can get with speech recognition if you use a speaker recognition model to detect a speaker, and then optimize your speech recognition parameters.

- Feature fusion to get better results

- GMMs

# Conclusions

Does this section contain the results or is this just for us to pat ourselves on the back?

# Acknowledgements

The project members would like to thank UCLA faculty and staff for their guidance and Mathworks for MATLAB and Voicebox for their tool suite and digital speech processing functions.

# References

1. TBD
2. TBD
3. TBD
4. TBD