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 implements mel-frequency cepstral coefficients (MFCC) as the main element of the feature vector to a neural network in order to develop a speaker recognition model. MFCCs are a commonly used speaker identification feature vector as they are modeled after the human auditory system. A data set consisting of 565 audio samples ranging from 3 to 5 seconds were using to form the model. Two sets of training data were trained against three different cases. The first set of training data consisted of various speakers reading the same sentence (read set) while the second set consisted of participants speaking in casual speech (phone set). The model will use the training data to form a model which will make predictions on the identity of the test data. The test data contains a read set, a phone set, and a mismatched set. By obtaining a high identification rate in all cases, the model would provide evidence to being a considerably robust speaker recognition system. The model obtains a <insert accuracy rates>

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

# Introduction

This section should talk generally about the MFCC, and signal processing stuff. Can talk about the problems with speaker recognition and be the literary survey part.

Speech can genrally be represented by the voice source, the vocal tract, and radiation. The voice source corresponds to the pitch of the speaker and radiation corresponds to the effects of the medium the sound is traveling in and are not relevant for our purposes. The vocal tract however is represented as the transfer function of the sound, which offer unique characteristics for speech and speaker recognition. Convolution in time is multiplation in frequency, thus the MFCC ffunction shown above takes the log yo.The MFCC is a warped frequency domain calls mels which is in the cepstrum spectrum.

This section provides an overview of the speaker identification tasks and methodologies that a user can take in order to solve the problem. The challenge in speaker recognition resides in the nature of human speech. The short duration of the samples of speech used to form the model and the stylistic mismatches of the read vs. read speech. The identity information of a speaker is an entity of how the speech is spoken in contrast to what is being said. Speech therefore possesses a large degree of variability between speakers. Humans have an uncanny ability to recognize the subtleties in human speech to provide an extremely accurate method of speaker identification. Machines rely on mathematical models based off of studies of human biology to provide an estimate of speaker recognition. Systems to detect speakers must develop a model of the speaker in the training stage. The features are extracted, and buh bam. The mel-frequency cepstral coefficients used in this project aim to model human auditory perception. [speaker recognition by machines and humans]. First describe what the LPC is, and why formants are important for recognizing vowels/speech. Taking the LPC of a signal and then doing the cepstrum of that gives you the LPCC. Talk about what the cepstrum is and why its important. Originally thought of as deconvolution for echo cancellation purposes, the cepstrum is used to provide robust uniqueness quantifiers for a sampled speech signal. The cepstrum can be used to reliably distinguish between the source, and the transfer function. The transfer function is provides the unique features that may be used for speaker identification like vocal tract length. The cepstrum is initially calculated by taking the logarithm of the spectrum, and then performing the inverse fourier transform to enter the cepstral domain. Talk about LPCCs. Then take the LPCC and then doing something gives you the MFCC I think. Mel-frequency cepstral coefficients function in the mel domain. The mel domain is a warped frequency domain meant to accurately model the human auditory system. The human auditory system works as a precise filter bank where the lower frequency filters are very sharp and close together while the larger frequencies become wider. The mel to frequency computation is as follows:

(1)

It’s the DFT, then warped frequency domain, then cepstrum, then inverse. The complex exponential can be expressed in terms of cosines and sines, but we only take the cosiens, DCT. The deltas/derivatives capture the dynamics of each frame to each frame and their formants, which is useful for considering the uniqueness of the speech. The deltas take the MFCC coefficient of frame n and subtract it from frame n-1 to form a new value. Challenges: short duration. Stylistic mismatch. Auditory filters in the ear are narrow as the frequency is low. This relates to the critical band filters in the Rabiner book. P. 144 has the mel-frequency reation. 464 has the MFCC eqn. Mel-scale imitates that. The mel domain allows for improved resolution at lower frequencies which comes into play during the mel-filter bank of the MFCC.

MFCCs implicitly capture speaker-specific ifnromation [a new set of features…] Cite voicebox.

# Methodology

This section describes the methodology forming the speaker identification system. 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 was first extracted and stored in a feature dictionary, but the result is the same.

A pre-emphasis filter attempts to improve the final score by providing the feature extraction mechanism a more ideal format. The feature vector consist of the Mel-Frequency Cepstrum Coefficients as the main quantity which is then fed into the K-Nearest Neighbors (K-NN) machine learning implementation to produce a model that can predict test data.

Starting from a baseline using the average pitch frequency of every signal, a robust speech recognition system was developed. This section should show what we actually do. The general block diagram is as follows:



Figure 1: *Block Diagram of Speech Recognition System*

## 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 high accuracy rate.

## Feature Extraction

After reading the signal and pre-emphasis filtering, the feature vector is developed as described in Figure 2. For each audio sample of *fs* = 22050Hz, the signal was sampled <x> times. Twelve MFCC coefficients were extracted from each frame creating a frame vector. One property of the cepstral coefficients is that the values approach zero rapidly, therefore no more than twelve coefficients are needed to produce meaningful values [gotta be a reference for that somewhere]. In test runs, extracting twenty coefficients and extracting twelve coefficients yielded nearly identical results, therefore for computation reasons the MFCC calculation only consisted of the twelve coefficients. The MFCC computation in each frame vector implements triangular shaped filters in the mel domain, as is customary for MFCC implementations (why?). At the ends of each sample, the error is high because the filter catches zeroes in its calculation. To counteract this, the highest and lowest filters are tapered down to zero. Additionally, the filters act in the absolute magnitude domain. For every signal, n frames created a frame vector of 12 MFCC coefficients which were averaged across every sample to obtain the average 12 MFCC coefficients of a signal. The standard deviation was calculated across the signals 12 MFCC coefficients and concatenated with the feature to form a 13 valued feature vector per audio sample. The addition of the standard deviation improved the EER results as the data feature vector, which consisted of the mean of raw data, was represented in a more dynamic way. <need an equation?> 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.

The initial baseline used a feature vector consisting of the average pitch frequency of the sample. In this scenario, each frame resulted in an average pitch, which was then averaged across all the frames in the signal to achieve the average frame of the signal. The average pitch was initially concatenated with the MFCC coefficients to provide additional features for the machine learning approach to build the model. Not only is pitch not optimal for speaker recognition of the same gender[cite something?] but providing the computing resources for the calculations heavily impact run-time measurements. During trial runs, it was determined that the average pitch frequency feature produced negligent improvements in accuracy, yet significant increases in run-time so the feature was removed from further testing. <provide run-time measurements>.

Other trials runs implemented the delta of the MFCC, which acts to provide a dynamic representation of the signal as it passed through the frames. In other words, the delta MFCCs describe the dynamics of the formants as the audio signal develops which is highly useful for speech recognition but not necessarily speaker recognition, as is the aim of this project. As another layer of abstraction, the delta-delta MFCCs were tested, which give a representation of the dynamics of the signal across multiple frames. Both the delta MFCCs and delta-delta MFCCs produced identical results to the prior implementations and were therefore left out for computation reasons.

A Gaussian Mixture Model (GMM) was attempted briefly after MFCC feature vector extraction, but… (maybe don’t include…)



Figure 2: *Feature Extraction of MFCC*

## K-Nearest Neighbors

After the feature vectors for each audio signal are extracted, the binary labels are fed into the K-Nearest Neighbors machine learning algorithm to develop a model that can be used to predict the speakers for the testing phase. 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.

Using the labels as classifiers, speaker models are formed during the training phase. The speaker models are mapped across the feature spaces as demonstrated by the unique shapes in Figure 3. During the testing phase, the feature vector is plotted across the K-NN feature map and K-NN parameters determine where the speaker under test is classified. The dotted circle in Figure 3 corresponds to the scanning distance of the algorithm. The distance between each shape and the data point under test is recorded, and the algorithm simply counts the ‘shapes’ that are closest to the data point within the scanning range. Whichever speaker model is closest to the speaker under test, the model predicts as belonging to that group. In other words, the model will predict the ‘shape’ of the speaker under test according to a certain set of parameters and feature vectors from the training phase.



Figure 3: *K-NN*

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. 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 were optimized to seventy-five.

All figures must be centered on the column (or page, if the figure spans both columns). Figure captions should follow each figure and have the format given in Figure 1.



Figures should be preferably line drawings. If they contain gray levels or colors, they should be checked to print well on a high-quality non-color laser printer.

Graphics (i. e., illustrations, figures) must not use stipple fill patterns because they will not reproduce properly in Adobe PDF. Please use only SOLID FILL COLORS.

Figures which span 2 columns (i. e., occupy full page width) must be placed at the top or bottom of the page.

## Tables

An example of a table is shown in Table 1. The caption text must be above the table.

Table 1: *This is an example of a table.*

|  |  |
| --- | --- |
| Ratio | Decibels |
| 1/1 | 0 |
| 2/1 | ≈ 6 |
| 3.16 | 10 |
| 1/10 | 20 |
| 10/1 | -20 |
| 100/1 | 40 |
| 1000/1 | 60 |

Figure 1: *Schematic diagram of speech production.*

## Equations

Equations should be placed on separate lines and numbered. Examples of equations are given below. Particularly,

 (1)

where  is a special warping function

 (2)

A residue theorem states that

 (3)

Applying (3) to (1), it is quite straightforward to see that

 (4)

Finally we have proven the secret theorem of all speech sciences. No more math is needed to show how useful the result is!

## Information for Word users only

For ease of formatting, please use the styles listed in Table 2. The styles are defined in this template file and are shown in the order in which they would be used when writing a paper.

When the heading styles in Table 2 are used, section numbers are no longer required to be typed in because they will be automatically numbered by Word. Similarly, reference items will be automatically numbered by Word when the “Reference” style is used.

Table 2: *Main predefined styles in Word.*

|  |  |
| --- | --- |
| Style Name | Entities in a Paper |
| Title | Title |
| Author | Author name |
| Affiliation | Author affiliation |
| Email | Email address |
| AbstractHeading | Abstract section heading |
| Body Text | First paragraph in abstract |
| Body Text Next | Following paragraphs in abstract |
| Index | Index terms |
| 1. Heading 1 | 1st level section heading |
| 1.1 Heading 2 | 2nd level section heading |
| 1.1.1 Heading 3 | 3rd level section heading |
| Body Text | First paragraph in section |
| Body Text Next | Following paragraphs in section |
| Figure Caption | Figure caption |
| Table Caption | Table caption |
| Equation | Equations |
| • List Bullet | Bulleted lists |
| [1] Reference | References |

If your Word document contains equations, you must not save your Word document from “.docx” to “.doc” because when doing so, Word will convert all equations to images of unacceptably low resolution.

## Hyperlinks

For technical reasons, the proceedings editor will strip all active links from the papers during processing. Hyperlinks can be included in your paper, if written in full, e. g. “http://www.foo.com/index.html”. The link text must be all black. Please make sure that they present no problems in printing to paper.

## Multimedia Files

The INTERSPEECH organizing committee offers the possibility to submit multimedia files. These files are meant for audio-visual illustrations that cannot be conveyed in text, tables and graphs. Just like you would when including graphics, make sure that you have sufficient author rights to the multimedia materials that you submit for publication. The proceedings media will NOT contain readers or players, so be sure to use widely accepted file formats, such as MPEG, Windows WAVE PCM (.wav) or Windows Media Video (.wmv) using standard codecs.

Your multimedia files must be submitted in a single ZIP file for each separate paper. Within the ZIP file you can use folders and filenames to help organize the multimedia files. In the ZIP file you should include a TEXT or HTML index file which describes the purpose and significance of each multimedia file. From within the manuscript, refer to a multimedia illustration by its filename. Use short file names without blanks for clarity.

The ZIP file you submit will be included as-is in the proceedings media and will be linked to your paper in the navigation interface of the proceedings. The publisher and the conference committee will not check or change the contents of your ZIP file.

Users of the proceedings who wish to access your multimedia files will click the link to the ZIP file which will then be opened by the operating system of their computer. Access to the contents of the ZIP file will be governed entirely by the operating system of the user's computer.

## Page Numbering

Page numbers will be added later to the document electro­nically. *Don't make any footers or headers*!

## References

The reference format is the standard IEEE one. References should be numbered in order of appearance, for example [1], [2], [3, pp. 417–422], and [4].

## Abstract

The total length of the abstract is limited to 200 words. The abstract included in your paper and the one you enter during web-based submission must be identical. Avoid non-ASCII characters or symbols as they may not display correctly in the abstract book.

## Author affiliation

Please list country names as part of the affiliation for each country.

## Number of authors in the author list

The maximum number of authors in the author list is twenty. If the number of contributing authors is more than twenty, they should be listed in a footnote or in acknowledgement section, as appropriate.

## Submitted files

Authors are requested to submit PDF files of their manu­scripts. You can use commercially available tools or for in­stance http://www.pdfforge.org/products/pdfcreator. The PDF file should comply with the following requirements: (a) there must be no PASSWORD protection on the PDF file at all; (b) all fonts must be embedded; and (c) the file must be text searchable (do CTRL-F and try to find a common word such as ‘the’). The proceedings editors will contact authors of non-complying files to obtain a replace­ment. In order not to endanger the preparation of the pro­ceedings, papers for which a replacement is not provided timely will be withdrawn.

# Discussion and Alternate Approaches

During the development of the speech recognition model presented herein, various approaches were explored that are not included in this report. This section will describe and discuss alternate approaches that were researched to obtain the optimum speaker recognition model.

## 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

-- start here

This discussion is a summary and talks about other stuff we did.

Fusion of features

GMM

Neural Networks

# Conclusions

Authors must proof read their PDF file prior to submission to ensure it is correct. Authors should not rely on proof­reading the Word file. Please proofread the PDF file before it is submitted.

# Acknowledgements

The ISCA Board would like to thank the organizing committees of the past INTERSPEECH conferences for their help and for kindly providing the template files.

Note to authors: Authors should not use logos in acknowledgement section; rather authors should acknowledge corporations by naming them only.

# References

1. S. B. Davis and P. Mermelstein, “Comparison of parametric representation for monosyllabic word recognition in con­ti­nuous­ly spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.
2. L. R. Rabiner, “A tutorial on hidden Markov models and selec­ted applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
3. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statis­tical Learning – Data Mining, Inference, and Prediction*. New York: Springer, 2009.
4. F. Lastname1, F. Lastname2, and F. Lastname3, “Title of your INTERSPEECH 2019 publication,” in *INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association, September 15-19, Graz, Austria, Proceedings*, 2019, pp. 100–104