**ABSTRACT**

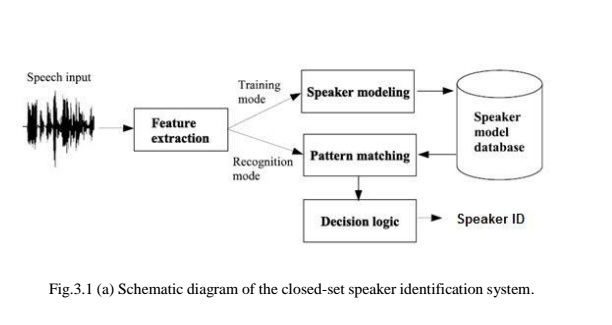
Audio information plays a rather important role in the increasing digital content that is available today; resulting in a need for methodologies that automatically analyze such content. Speaker Identification is one of the vital field of research based upon Voice Signals. Its other notable fields are: Speech Recognition, Speech-to-Text Conversion, and vice versa, etc. Mel Frequency Cepstral Coefficient (MFCC) is considered a key factor in performing Speaker Identification. But, there are other features lists available as an alternate to MFCC; like- Linear Predictor Coefficient (LPC), Spectrum Sub-band Centroid (SSC), Rhythm, Turbulence, Line Spectral Frequency (LPF), ChromaFactor, etc. Gaussian Mixture Model (GMM) is the most popular model for training on our data. The training task can also be executed on other significant models; viz. Hidden Markov Model (HMM). Recently, most of the model training phase for a speaker identification project is executed using Deep learning; especially, Artificial Neural Networks (ANN). In this project, we are mainly focused on implementing MFCC and GMM in pair to achieve our target.

We have considered MFCC with “tuned parameters” as the primary feature and delta- MFCC as secondary feature. And, we have implemented GMM with some tuned parameters to train our model. We have performed this project on two different kinds of Dataset; viz. “VoxForge” Dataset and a custom dataset which we have prepared by ourselves. We have obtained an outstanding result on both of these Datasets; viz. 100% accuracy on VoxForge Dataset and 95.29 % accuracy on self prepared Dataset. We demonstrate that speaker identification task can be performed using MFCC and GMM together with outstanding accuracy in Identification/ Diarization results.

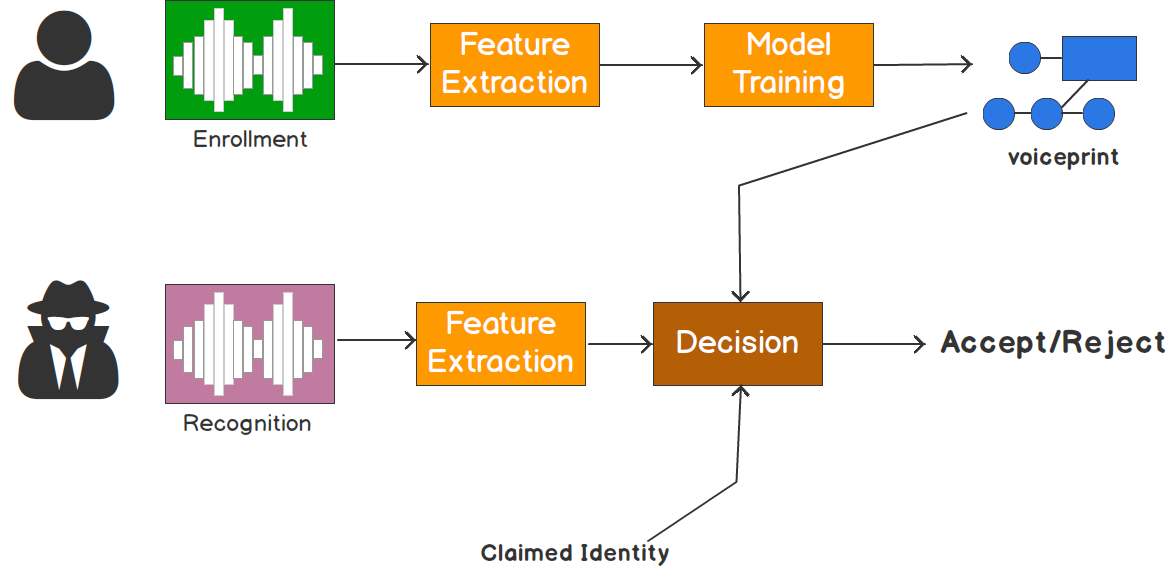
**Keywords:** *Speaker Identification, LPC, LSF, Speaker Diarization, MFCC, HMM, GMM, VoxForge, ANN, SSC, Turbulence, ChromaFactor, etc.*

*Speaker recognition* is the process of automatically recognizing who is speaking on the basis of individual information included in speech signals. It can be divided into *Speaker Identification* and *Speaker Verification*. **Speaker identification** determines which registered speaker provides a given utterance from amongst a set of known speakers. **Speaker verification** accepts or rejects the identity claim of a speaker - is the speaker the person they say they are? Speaker recognition technology makes it possible to a the speaker's voice to control access to restricted services, for example, phone access to banking, database services, shopping or voice mail, and access to secure equipment. Both technologies require users to "enroll" in the system, that is, to give examples of their speech to a system so that it can characterize (or learn) their voice patterns.

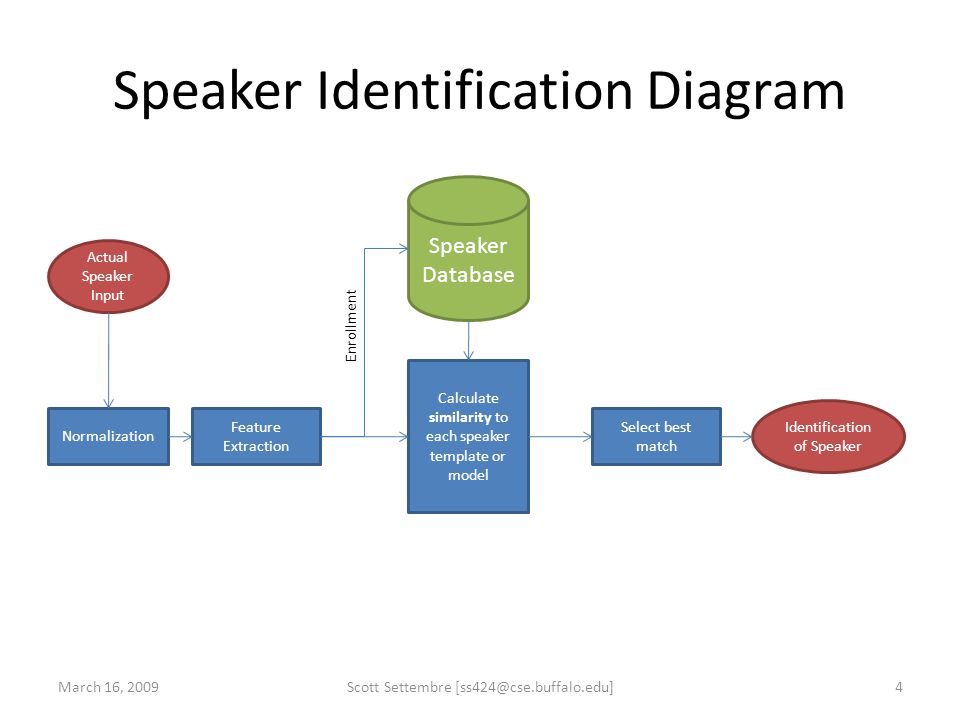
**Speaker Recognition:** The speaker recognition task can be performed in two ways: Text Dependant (Speaker Verification) or Text Independent (Speaker Identification).



**Speaker Verification:** To perform speaker verification, one must use the same data for training and testing purpose. Viz. train data and test data are same. It is basically **1: 1** recognition system; in which identity of 1 speaker is matched and verified with the claimed identity of 1 speaker in audio sample.



**Speaker Identification:** To perform speaker identification, different train and test data is used. It is basically **1: N** recognition system; in which identity of 1 speaker is matched with N speaker’s voice samples stored in the training model.



1. **Technology**

Speaker recognition is a “[**pattern recognition**](https://en.wikipedia.org/wiki/Pattern_recognition)” problem. The various technologies used to process and store voice prints include [*frequency estimation*](https://en.wikipedia.org/wiki/Frequency_estimation)*,*[*hidden Markov models*](https://en.wikipedia.org/wiki/Hidden_Markov_model) *(HMM),*[*Gaussian mixture models*](https://en.wikipedia.org/wiki/Gaussian_mixture_model) *(GMM),*[*pattern matching*](https://en.wikipedia.org/wiki/Pattern_matching)*algorithms,*[*neural networks*](https://en.wikipedia.org/wiki/Neural_networks) *(ANN or MLP),*[*matrix representation*](https://en.wikipedia.org/wiki/Matrix_representation)*, Vector Quantization* and [*decision trees*](https://en.wikipedia.org/wiki/Decision_tree_learning). Some systems also use "**anti-speaker**" techniques, such as [*cohort models*](https://en.wikipedia.org/wiki/Cohort_model), and *world models*. Spectral features are predominantly used in representing speaker characteristics.

[*Ambient noise levels*](https://en.wikipedia.org/wiki/Ambient_noise_level) can impede both collections of the initial and subsequent voice samples. **Noise reduction algorithms** can be employed to improve accuracy, but incorrect application can have the opposite effect. Performance degradation can result from changes in behavioral attributes of the voice and from enrollment using one telephone and verification on another telephone ("cross channel"). Integration with [two-factor authentication](https://en.wikipedia.org/wiki/Two-factor_authentication) products is expected to increase. Voice changes due to ageing may impact system performance over time. Some systems adapt the speaker models after each of the successful verification to capture such long-term changes in the voice, though there is debate regarding the overall security impact imposed by automated adaptation.

Capture of the biometric is seen as non-invasive. The technology traditionally uses existing microphones and voice transmission technology allowing recognition over long distances via ordinary telephones (wired or wireless).

1. **Applications**

The first international patent was filed in 1983, coming from the telecommunication research in [CSELT](https://en.wikipedia.org/wiki/CSELT) (Italy) by *Michele Cavazza* and [*Alberto Ciaramella*](https://en.wikipedia.org/wiki/Alberto_Ciaramella) as a basis for both future *Telco* services to final customers and to improve the noise-reduction techniques across the network.

In May 2013 it was announced that [*Barclays Wealth*](https://en.wikipedia.org/wiki/Barclays_Wealth) was to use passive speaker recognition to verify the identity of telephone customers within 30 seconds of normal conversation The system used had been developed by voice recognition company [*Nuance*](https://en.wikipedia.org/wiki/Nuance_Communications) (that in 2011 acquired the company [*Loquendo*](https://en.wikipedia.org/wiki/Loquendo), the spin-off from CSELT itself for speech technology), the company behind [*Apple's*](https://en.wikipedia.org/wiki/Apple_Inc.)[*Siri*](https://en.wikipedia.org/wiki/Siri_(software)) technology. A verified voiceprint was to be used to identify callers to the system and the system would in the future be rolled out across the company.

The private banking division of Barclays was the first financial services firm to deploy voice biometrics as the primary means to authenticate customers to their [call centers](https://en.wikipedia.org/wiki/Call_center). 93% of customer users had rated the system at "9 out of 10" for speed, ease of use and security.

Since then, *Nuance Voice Biometrics solutions* have been deployed across several financial institutions, including *Banco Santander, Royal Bank of Canada, Tangerine Bank*, and *Manulife*.

In August 2014 *GoVivace Inc.* deployed a speaker identification system that allowed its telecom industry client to positively search for an individual among millions of speakers by using just a single example recording of their voice.

Speaker recognition may also be used in criminal investigations, such as those of the 2014 executions of, amongst others, [*James Foley*](https://en.wikipedia.org/wiki/James_Foley_(journalist)) and [*Steven Sotloff*](https://en.wikipedia.org/wiki/Steven_Sotloff).

In February 2016 UK high-street bank [HSBC](https://en.wikipedia.org/wiki/HSBC) and its internet-based retail bank [First Direct](https://en.wikipedia.org/wiki/First_Direct) announced that it would offer 15 million customers its biometric banking software to access online and phone accounts using their fingerprint or voice.

***Source:*** *Wikipedia*

1. **Concepts**

Imagine a scenario. A criminal is being tried in court. He denies saying something. The prosecution brings a recording, saying they have his confession on tape. As the accused vigorously denies the voice being his, an expert shows just why the voice could be no one else’s.

A fictional scene! Perhaps, but it is a reality that **no two persons** in the world have **exactly the same voices**. Do you know why this is so?

Our vocal cords can be in any of about 170 different positions. If the vocal cords are slack, they may vibrate about 80 times per second and the result is deep tones. If they are tense, they vibrate rapidly, perhaps a 1000 times a second, and produce short sound waves or high tones. The pitch of voice depends upon the length of the vocal cords. Each voice has a certain range of frequencies. It is this range that determines the kind of voice a person has. The movement of the tongue against the palate, the shaping of the lips and arrangement of teeth also bring about changes in the voice. Since the structures and movements of all these organs are different in different persons, the **voices of no two persons** in the world **can be identical**.

This is the basic idea behind this huge application of audio analysis and speaker recognition or speaker identification/ verification. The key features that differentiate two human voices can be extracted from a voice sample through feature extraction. These can be further used for model reference training; and finally, the identification of the speaker.

**Conclusion:**

The performances of speaker identification system using various feature extraction and matching techniques. MFCC algorithm is used in our system because it has least false acceptance ratio. I n order to improve system performance and also to achieve high accuracy GMM model can be used in feature matching technique. The speakers were trained and tested by using MFCC and GMM model. They give better identification rate for speaker features. In future work, this technique integrated the pitch information with MFCC and also to analyze the speaker identification system performance in the presence of noise.

**Future Scope:**

Speaker Recognition field is very interesting and exploring. But, we have got a very limited research in this field as per current scenario. So, there is a lot to achieve in this field. We deal with a huge chunk of audio data daily. Plus, in this advancing world, we could never leave behind synthesizing speeches and recognizing speakers.

My project can be further extended with a **GUI interface** for ease in user experience. If we talk about its applicability, it has enormous future scope. The procedure followed and outputs from my project would be **helpful for other researchers**. They can find and suggest new methods to **improve accuracy and consistency** of the model.

This particular project can be used as a **voice biometric system** to identify and catch criminals. I have a vision of extending this project further by adding Speech Recognition into this project. **Speech Recognition + Speaker Recognition** would serve perfect for my purpose. I would like to make an intelligent system that can start **tapping phone calls** when getting any **vulnerable keywords**, mostly used by terrorists; then the speaker recognition system would **recognize that terrorist** or other criminal. This would help a lot in reducing crime and terror attacks in our country.

**Bibliography/References**

**1.** Dalei Wu, Andrew Morris, Jacques Koreman; MLP Internal Representation as Discriminative Features for Improved Speaker Recognition, 2005.

**2.** [Mitchell McLaren](https://www.sri.com/about/people/mitchell-mclaren); Advances in deep neural network approaches to speaker recognition, **ICASSP** – April, 2015.

**3.** Getting Started with Audio Data Analysis using Deep Learning (with case study) **- AnalyticsVidhya.com**

**4.** David Moffat, David Ronan, Joshua D. Reiss ;AN EVALUATION OF AUDIO FEATURE EXTRACTION TOOLBOXES; Nov 30 - Dec 3, 2015

**5.** B.Mathieu, S.Essid, T.Fillon, J.Prado, G.Richard; YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software, proceedings of the 11th **ISMIR conference, Utrecht,** Netherlands, 2010.