Minor Project Report on

Comparison of Various Implementations of Automatic Speaker Recognition Systems

Shantanu Sampath (15EE245)

Under the Guidance of,

Dr. Krishnan CMC

Department of Electrical and Electronics Engineering, NITK Surathkal

Date of Submission: 2-MAY-2018

in partial fulfillment for the award of the degree

of

Bachelor of Technology

In

Electrical and Electronics Engineering

At

Department of Electrical and Electronics Engineering

National Institute of Technology Karnataka, Surathkal

National Institute of Technology Karnataka

Department of Electrical and Electronics Engineering

CERTIFICATE

This is certify that the thesis entitled **'Comparison of Various Implementations of Automatic Speaker Recognition Systems',** submitted by **Shantanu Sampath (15EE245)** is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of B.Tech in Electrical and Electronics Engineering at National Institute of Technology Karnataka, Surathkal.

Dr. Krishnan CMC

Assistant Professor
Department of EEE
NITK, Surathkal

ACKNOWLEDGEMENT

I would like to thank my guide, Dr. Krishnan CMC for his valuable insights into this project and his unwavering support and encouragement that helped me through this project. I would also like to extend my sincere gratitude to the authors of all the research papers mentioned below for inspiring this study. Special thanks to A. Nagrani, J. S. Chung and A. Zisserman for giving me access to VoxCeleb, which is a large-scale speaker identification dataset consisting of labelled recordings of 1,251 celebrities collected from sources like YouTube.

Abstract

The following is a study conducted in order to compare the accuracy and feasibility of various implementations of Automatic Speaker Recognition Systems. The system would identify a particular speaker's voice print when presented with a list of different audio samples. For this study, a dataset consisting of labelled recordings of 1,251 celebrities collected from sources like YouTube was used. Relevant features from particular samples were extracted and then processed. These systems are a staple of biometric authentication systems and voice recognition systems.

Contents

1 Literature Survey Overview

Paper [1] gives us an overall picture of how speaker identification systems work. It outlines the various blocks in the methodology, namely the steps involved in both the training, testing and final deployment of the model. The focus is on the features extracted from speech sampled, most notably the various types of cepstral features involved and their respective merits.

1.0.1 Usage of MFCC with Gaussian Mixture models

Paper [2] proved to be very insightful in its depiction of the overall system. It provided the following concise and very informative block diagram of how the system works. Virtually all the papers surveyed have followed a similar, if not exact match of the same block diagram. Short speech recordings of the subjects (30 seconds 2 minutes) are collected and are processed to extract useful features (pitch, Mel Frequency Cepstral Coefficients (MFCCs), etc.). A machine learning model is trained on the basis of these features. In the identification phase, the model compares the features extracted from the new input with the model and arrives at a decision on whether or not the new speech sample fed belongs to a particular person. The paper also gives valuable insights into the nature of the signals processed and the dataset used. For the presented implementation of the system 50 voices of speakers of various ages and both sexes have been gathered. The recordings were sampled at 22050 sps, mono, with a 16-bit amplitude resolution. The paper also describes vividly the method

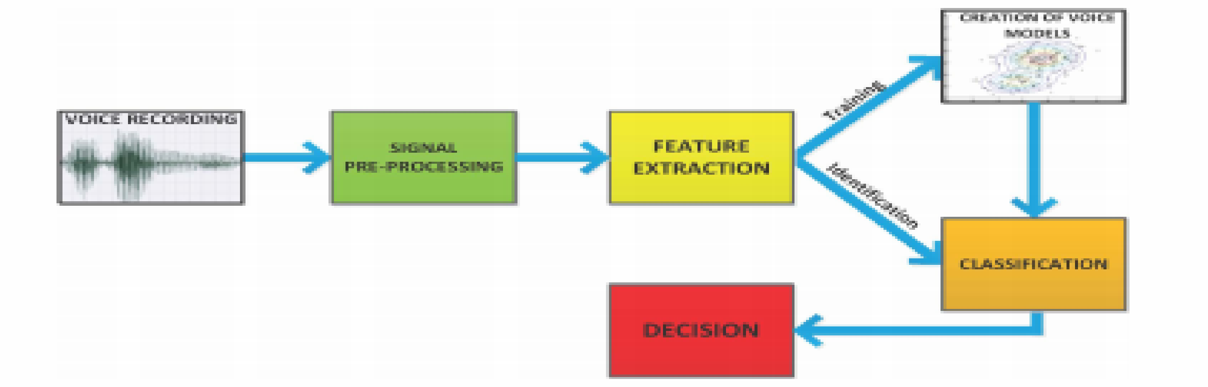


Figure 1. ASR system architecture

of extracting MFCCs from the speech signal. The paper also describes the use of a Gaussian Mixture Model as the classifier and also highlights the effectiveness of speaker recognition with training segment length and sampling frequency. The paper, The Research of speaker recognition based on GMM and SVM [3] describes the advantage of using a GMM with an SVM kernel. It also compares the accuracy of a GMM alone, and SVM alone and a GMM/SVM model. It is found that using a GMM along with an SVM kernel has improved accuracy.

diagram.png

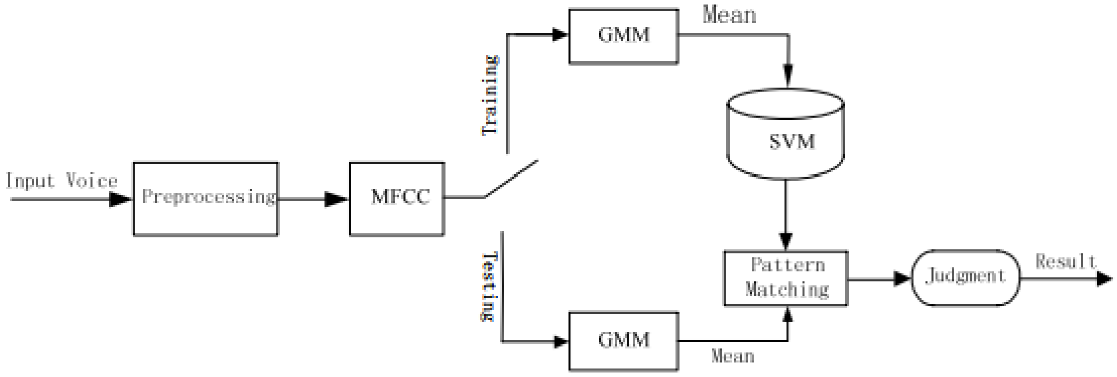


Fig. 4 The speaker recognition diagram based on GMM and SVM

The other papers [4] to [8] detail similar approaches on different datasets like telephone speech recordings and examine the robustness of said approaches. Performance Evaluation of an Automatic Forensic Speaker Recognition System based on GMM [7] presents a performance evaluation of a speech biometry system based on the statistical models GMM (Gaussian Mixture Models). In particular, the paper underlines the robustness to the degradation of various natural noises, and their impact on the system. Finally, the impact of the duration to both training and test sequences is highlighted.

1.0.2 Usage of Support Vector Machines (SVMs)

The paper titled Support Vector Machines, Mel-Frequency Cepstral Coefficients and the Discrete Cosine Transform Applied on Voice Based Biometric Authentication [11] gives us a good picture of how SVMs are used as robust classifiers as an alternative to GMMs. The features extracted from the voice recordings remain the same; the only difference being the classifier used. The voice samples used are in the Brazilian Portuguese language and had its features extracted through the Discrete Cosine Transform. Extracted features are applied on the Mel-frequency Cepstral Coefficients to create a two-dimensional matrix used as input to the SVM algorithm. This algorithm generates the pattern to be recognized, leading to a reliable speaker identification using few parameters and a small dataset.

The paper titled Support Vector Machines Using GMM super vectors for Speaker Verification [9]proposes a method of stacking the means of the GMM model to form a GMM mean super vector and examines the idea of using the GMM super vector in a support vector machine (SVM) classifier. Two new SVM kernels based on distance metrics between GMM models are proposed. It can be shown that these SVM kernels produce excellent classification accuracy in a NIST speaker recognition evaluation task. The paper titled A new study of GMM-SVM system for text-dependent speaker recognition [10] has also sought to do the same.

1.0.3 Various pre-processing methods and feature extraction methods

The paper titled Mel-Frequency Cepstral Coefficients as Features for Automatic Speaker Recognition [12] details the methodology of extraction of MFCCs from a voice sample. Recognition accuracy when exponential auditory critical bands are applied outperforms recognition accuracy of automatic speaker recognizer when triangular or rectangular auditory critical bands are applied. Application of transformation on elements of speaker model, which target decreasing of difference between testing and training models of the same speaker, can increase recognition accuracy. Mel Frequency Cepstral Coefficients Based Text Independent Automatic Speaker Recognition Using Matlab [13] talks about doing the same, but with using k-means clustering to classify the speech signal. Text Indpendent Speaker Recognition Using Wavelet Cepstral Coefficient and Butter Worth Filter [14] presents a method to carry out wavelet transform and to find wavelet cepstral coefficients to reduce the noise and also to improve accuracy. In Two-Step Noise Reduction Based on Soft Mask for Robust Speaker Identification [17], a two-step noise reduction algorithm based on soft mask and minimum mean square error short time spectral amplitude estimator was proposed. It is used in the signal pre-processing stage for more robust speaker identification. The proposed algorithm was tested and compared with the existing noise reduction algorithms in the problem of speaker identification. The advantage of the new noise reduction algorithm for some noise samples and signal-to-noise ratios was shown. The other papers, [15] and [16] propose similar methods in order to marginally improve accuracy and reduce noise.

1.0.4 Special Mentions

Paper [18] proposes the use of a neural network, trained with speech samples of the same person under various physiological conditions such as coughing, shouting, during chewing, mouth covered etc. A dictionary is created to store the signature features of each user's voice. A neural network is then trained using back propagation and accordingly weights are obtained to recognize voice in the testing phase.

Paper [19] explores the effect of varying number of weights and hidden layers in a neural network on the classification accuracy.

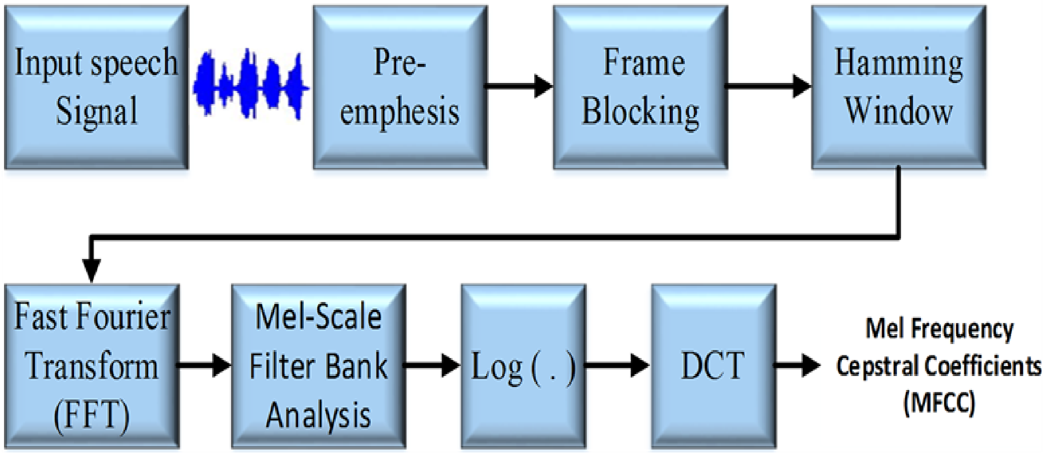
Paper [20] proposes a recognition strategy based on the use of Multicondition Gaussian Probabilistic Linear Discriminate Analysis (MGPLDA) to compensate the effects of this type of perturbation sources. Experiments using an Algerian Speaker Recognition Database collected in different recording conditions have shown that the proposed approach successfully improves the system performances in terms of discrimination and calibration quality.

Paper [21] analyzes the effects popular audio compression algorithms have on the performance of a speaker recognition system. Popular audio compression algorithms were used to compress both clean and noisy speech before being passed to a speaker recognition system. The experiments show that compression will have a negative effect on recognition rates if the compressed speech is clean. However, if small amounts of white Gaussian noise are added before the speech is compressed, recognition rates can be increased by as much as 7% with certain compression algorithms.

2 Feature Extraction

2.1 MFCC

The features used almost ubiquitously in speech / speaker recognition problems are the Mel Frequency Cepstral Coefficients. In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. They are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum").
MFCCs are commonly derived as follows:
1) Take the Fourier transform of (a windowed excerpt of) a signal.
2) Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3) Take the logs of the powers at each of the mel frequencies.
4) Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
The MFCCs are the amplitudes of the resulting spectrum. For the current project, audio samples of five celebrities were accumulated, namely: Quentin Tarantino,Billie Joe Armstrong,Carrie Fisher,David Attenborough and Eva Green. Quentin Tarantino was chosen as the speaker in question. The audio clips were divided in the ratio 70:30 (approx), the former amounting to 884 samples and the latter to 341. The MFCCs were ex-


Input speech Signal
Pre-emphasis
Frame Blocking
Hamming Window
Fast Fourier Transform (FFT)
Mel-Scale Filter Bank Analysis
Log (.)
DCT
Mel Frequency Cepstral Coefficients (MFCC)

tracted from each clip and compiled into two CSV files.

Delta Cepstral Features:

Delta-cepstral features were proposed to add dynamic information to the static cepstral features. They also improve recognition accuracy by adding a characterization of temporal dependencies to the hidden-markov models (HMM) frames, which are nominally assumed to be statistically independent of one another. For a short-time cepstral sequence C[n], the delta-cepstral features are typically defined as

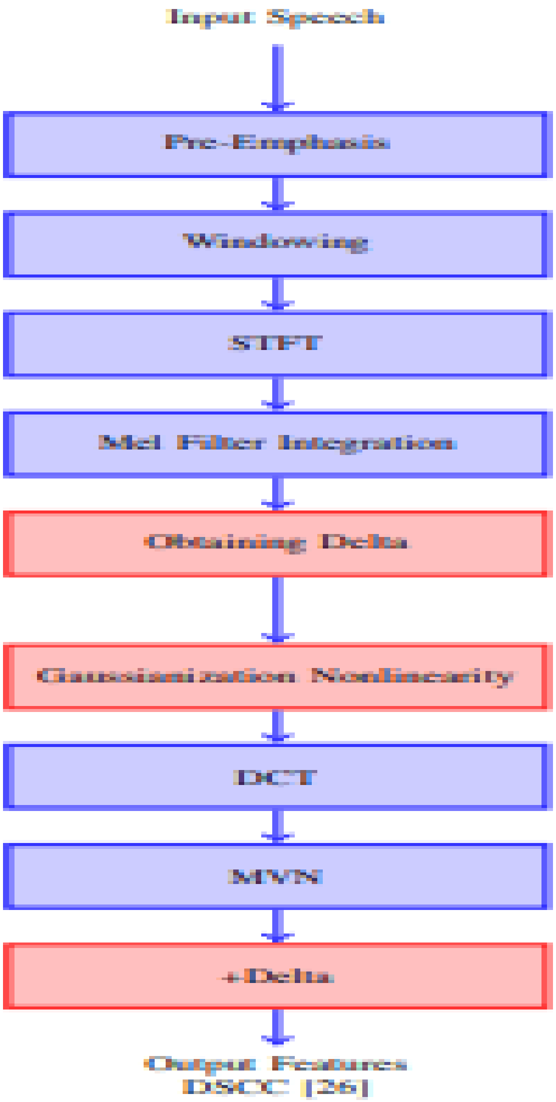
D[n] = C[n + m] - C[n - m]

where n is the index of the analysis frames and in practice m is approximately 2 or 3. Similarly, double-delta cepstral features are defined in terms of a subsequent delta-operation on the deltacepstral features.

We note that the addition of delta-cepstral features to the static 13-dimensional MFCC features strongly improves speech recognition accuracy, and a further (smaller) improvement is provided by the addition of double-delta cepstral. For these reasons some form of delta and double-delta cepstral features are part of nearly all speech recognition systems. It can be seen that the improvement provided by delta features gradually diminishes with lower SNR

Delta Spectral Cepstral Coefficients:

We now discuss the delta-spectral cepstral coefficients for ASR. These features are motivated by the non-stationarity of speech signals where it is easily observed in that figure that the short-time power of speech varies much more rapidly than the short-time power of noise. The vast differences between the rate of change of power for of speech and noise are likely to be one of the many cues that human ears can use to ignore the relatively stationary noise signals and focus on the rapidly-changing power of speech signals.


Input Speech
Pre-Emphasis
Windowing
STFT
Mel Filter Integration
Obtaining Delta
Gaussianization Nonlinearity
DCT
MVN
+Delta
Output Features
DSCC [26]

2.2 Pitch

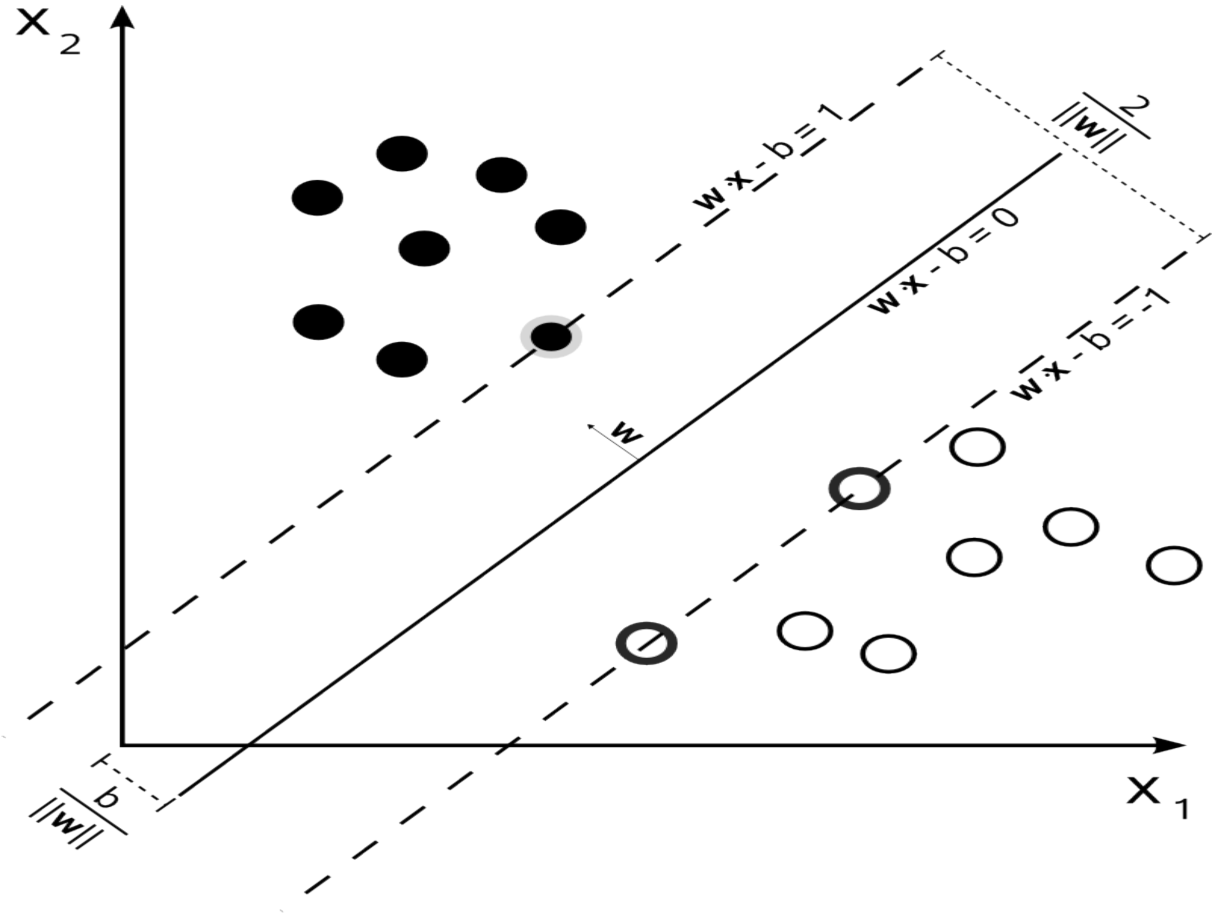
As an important voice signal parameter, pitch is widely used in voice compression coding, analysis synthesis, speech recognition and so on. Gender classifier is designed and implemented by pitch and MFCC, improves the recognition performance in SR . Taking into account the feature extraction from fixed-length windowed frame cannot reflect the nature of speech signal, the pitch is used to determine and analyze the length of frame in and a perfect performance for speaker verification under different noisy conditions is achieved. Speech can be broadly categorized as voiced and unvoiced. In the case of voiced speech, air from the lungs is modulated by vocal cords and results in a quasi-periodic excitation. The resulting sound is dominated by a relatively low-frequency oscillation, referred to as pitch. In the case of unvoiced speech, air from the lungs passes through a constriction in the vocal tract and becomes a turbulent, noise-like excitation. In the source-filter model of speech, the excitation is referred to as the source, and the vocal tract is referred to as the filter. Characterizing the source is an important part of characterizing the speech system.

As an example of voiced and unvoiced speech, consider a time-domain representation of the word ”two” (/T UW/). The consonant /T/ (unvoiced speech) looks like noise, while the vowel /UW/ (voiced speech) is characterized by a strong fundamental frequency.

3 Classifiers Used

3.1 Support Vector Machines

Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.



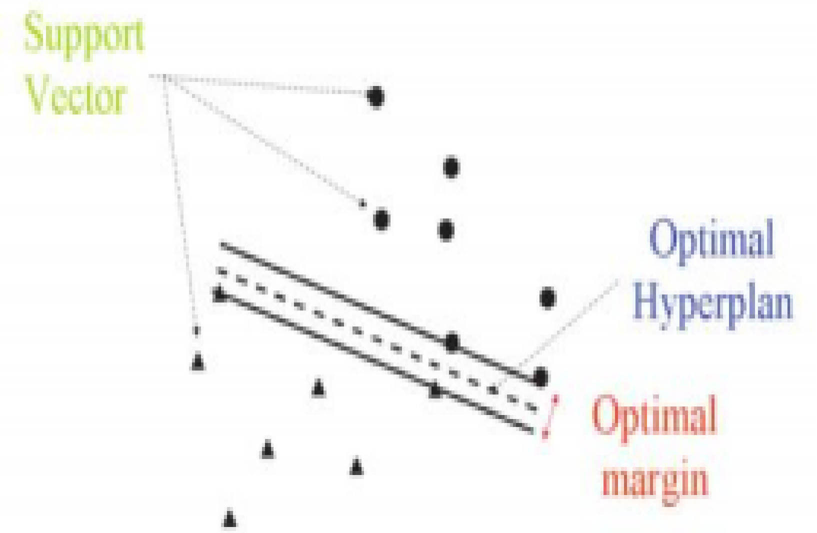
x_2
x_1
$\mathbf{w}\cdot\mathbf{x}-b=1$
$\mathbf{w}\cdot\mathbf{x}-b=0$
$\mathbf{w}\cdot\mathbf{x}-b=-1$
$\frac{2}{\|\mathbf{w}\|}$
$\frac{b}{\|\mathbf{w}\|}$
$\mathbf{w}$

SVM Kernels:

The nature of the classification of support vector machines depends on the type of kernel used. There are primarily two types:

1) Linear Kernel

We should now consider the case of two classes problem with N training samples. Each samples are described by a Support Vector (SV) Xi composed by the different band with n dimensions. The label of a sample is Yi. For a two classes case we consider the label - 1 for the first class and +1 for the other. The SVM classifier consists in defining the function
$f(x) = \text{sign}(\langle \omega, X \rangle + b)$
which finds the optimum separating hyperplane as presented in Figure 1, where is normal to the hyperplane, and —b— ω is the perpendicular distance from hyperplane to the origin. The sign of f(x) gives the label of the sample. The goal of the SVM is to maximize the margin between the optimal hyperplane and the support vector. So we search the min ω / 2. To do this, it is easier to use the Lagrange multiplier. The problem comes to solve:
$f(x) = \text{sign}(\Sigma$ yi. ai $\langle$ x.xi $\rangle$+ b) where i is the Lagrange multiplier.



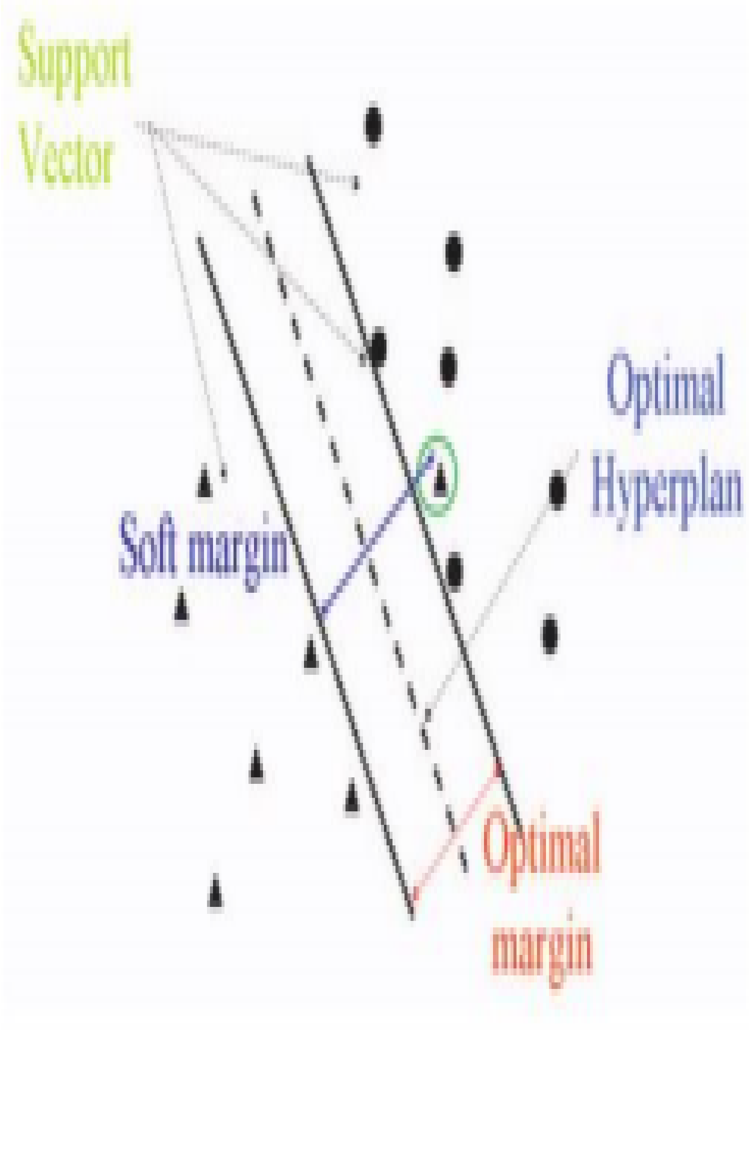
2) Non-Linear Kernels:

The kernel is a function that simulates the projection of the initial data in a feature space with higher dimension Φ: n → H. In this new space the data are considered as linearly separable. To apply this, the dot product ⟨xi, xj⟩ is replaced by the function, K(x, xi) = ⟨ ϕ(x), ϕ(xi)⟩ Then the new function to classify the data are:

f(x) = sign(Σ yi. ai K⟨ x.xi ⟩+ b)

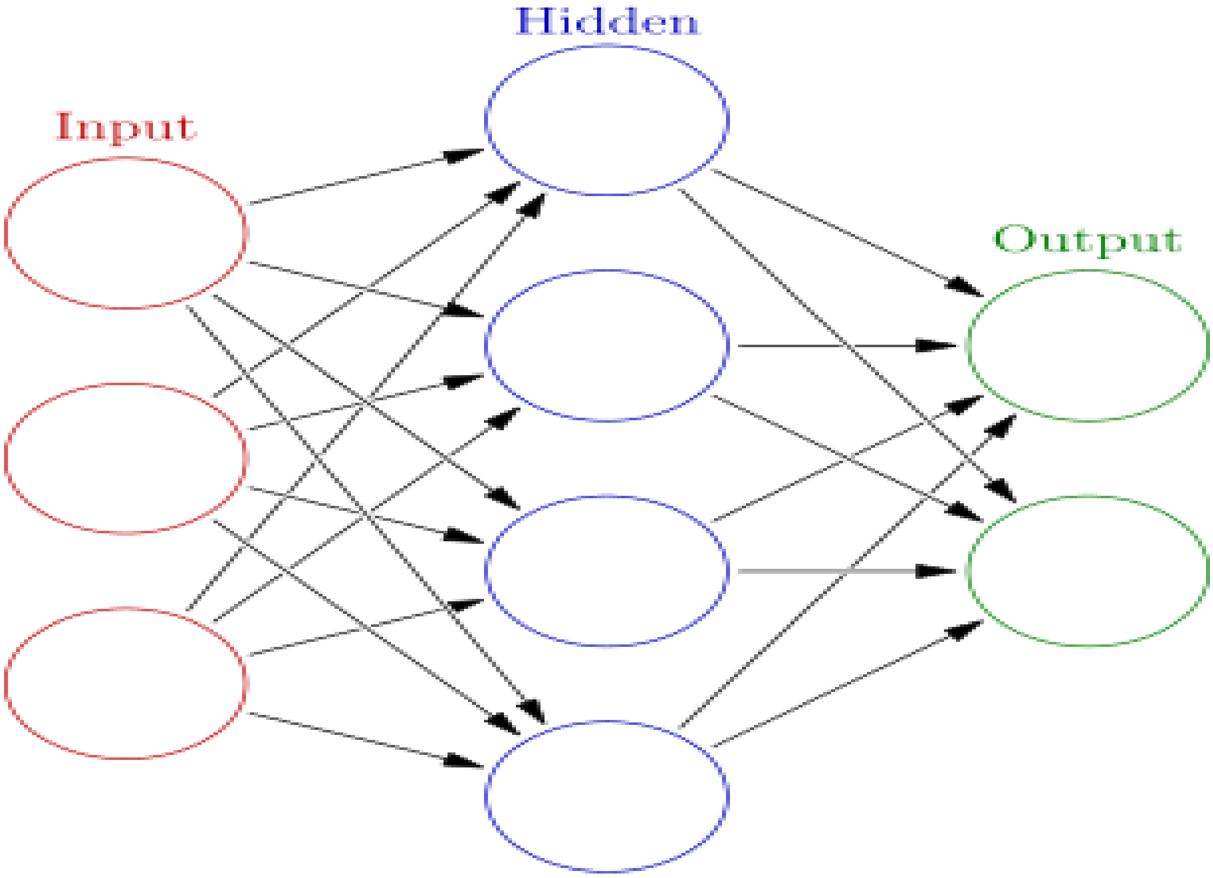
Most commonly used kernel is the RBF kernel or Gaussian Kernel:

$K(x, xi) = e^{-(x-xi)^2/2\sigma^2}$



3.2 Artificial Neural Networks

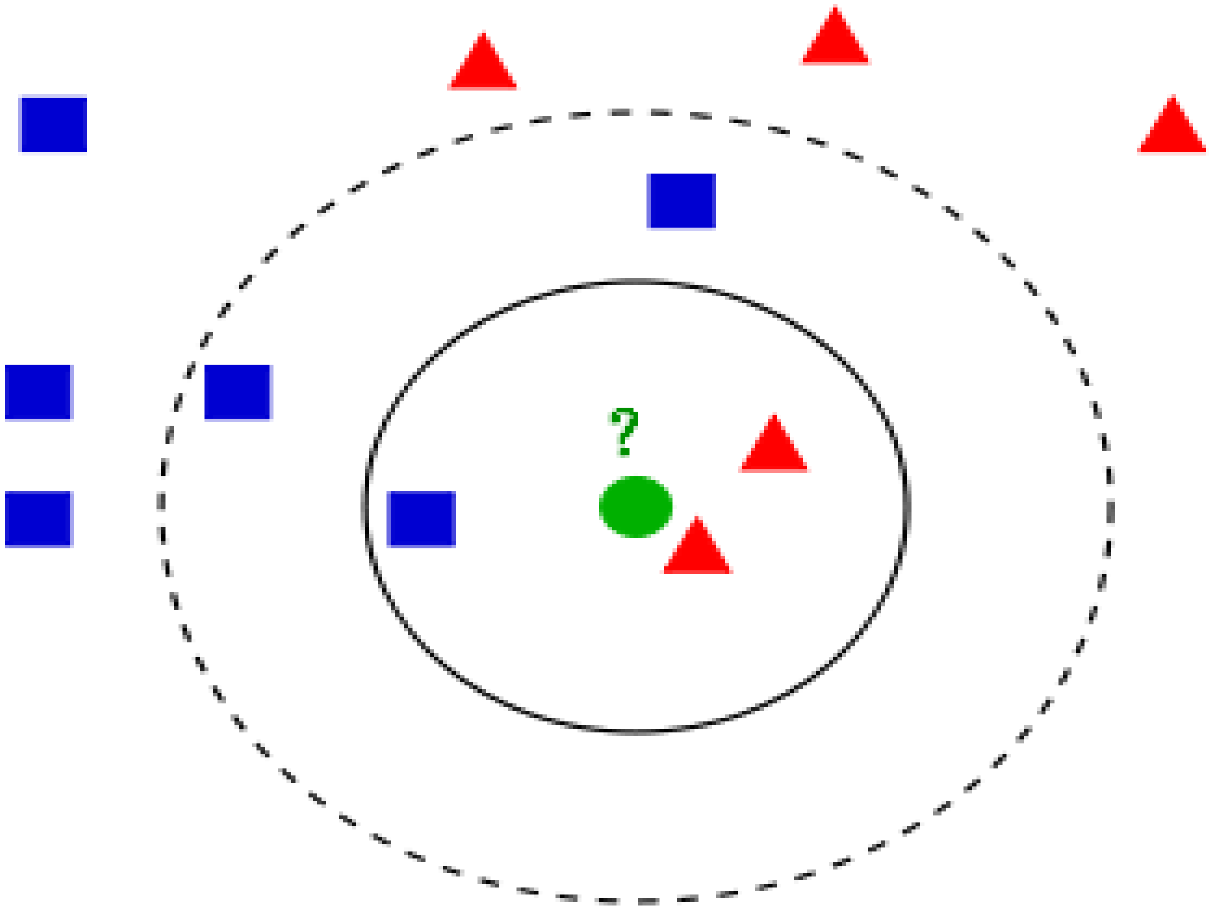
Artificial neural networks (ANNs) are computing systems vaguely inspired by the biological neural networks that constitute animal brains. Such systems ”learn” (i.e. progressively improve performance on) tasks by considering examples, generally without task-specific programming. In common ANN implementations, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is calculated by a non-linear function of the sum of its inputs. Artificial neurons and connections typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that only if the aggregate signal crosses that threshold is the signal sent. Typically, artificial neurons are organized in layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first (input), to the last (output) layer, possibly after traversing the layers multiple times.



3.3 k- Nearest Neighbours Algorithm

The k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space.k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

In the classification phase, k is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point.



Example of k-NN classification. The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles. If k = 3 (solid line circle) it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle. If k = 5 (dashed line circle) it is assigned to the first class (3 squares vs. 2 triangles inside the outer circle).

4 Results

4.1 Results after averaging features

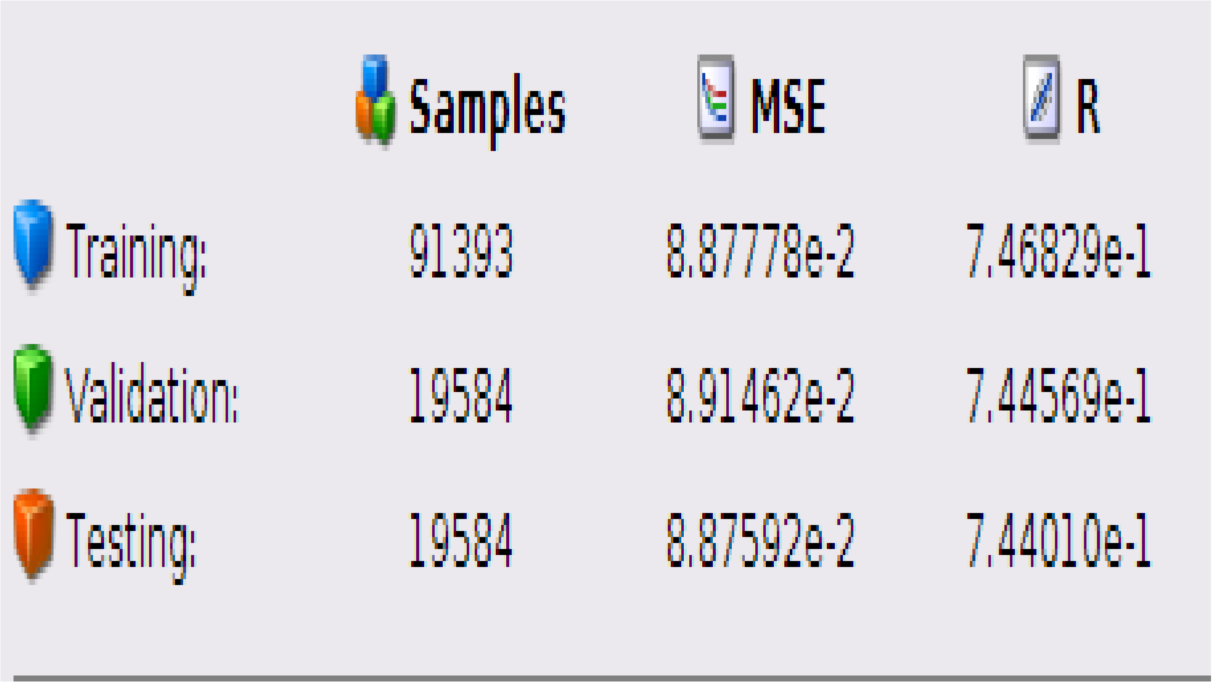
1) Features included 13 MFCC values
2) MFCC values were averaged over the entire track
Results as follows:
Artificial Neural Network:
Number of Layers = 3
Number of hidden layer neurons = 10
Number of samples in training and testing data combined = 1225
Training to Validation to Testing data ratio = 70:15:15
Using Bayesian Regularisation training algorithm, mean square errors are as follows:
Training: 2.10802e-1
Validation:0.00000e-0
Testing: 2.15709e-1
Support Vector Machines
Using polynomial Kernel, accuracy = 27.57%
Using Gaussian kernel, accuracy = 73.02%

4.2 Results without averaging features

1) Features included 20 MFCC values and 20 Delta-Delta values computed
2) The features were extracted for each 20 ms frame and not averaged over the entire track
Results are as follows:

Artificial Neural Network

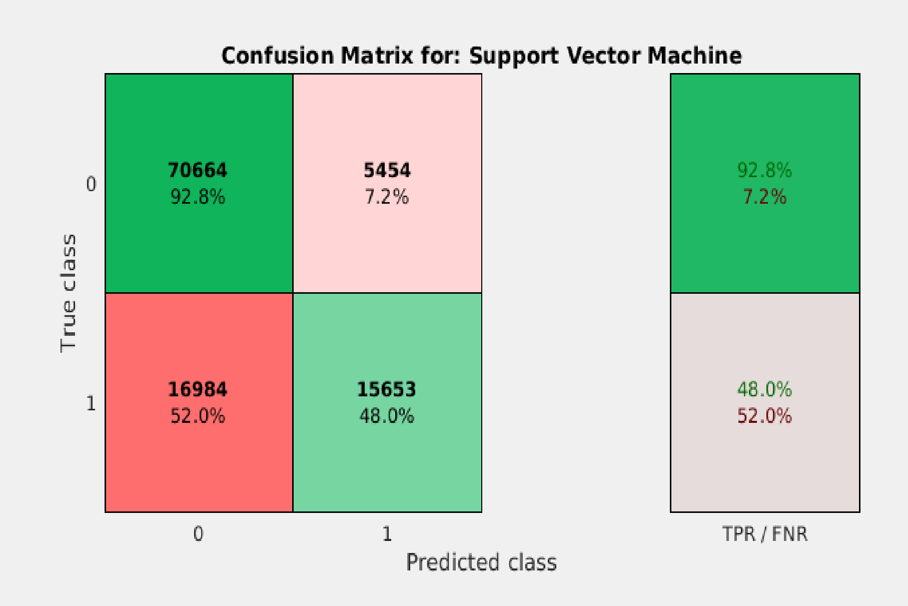
Number of Layers = 3
Number of hidden layer neurons = 10
Training to Validation to Testing data ratio = 70:15:15
Results are as follows:



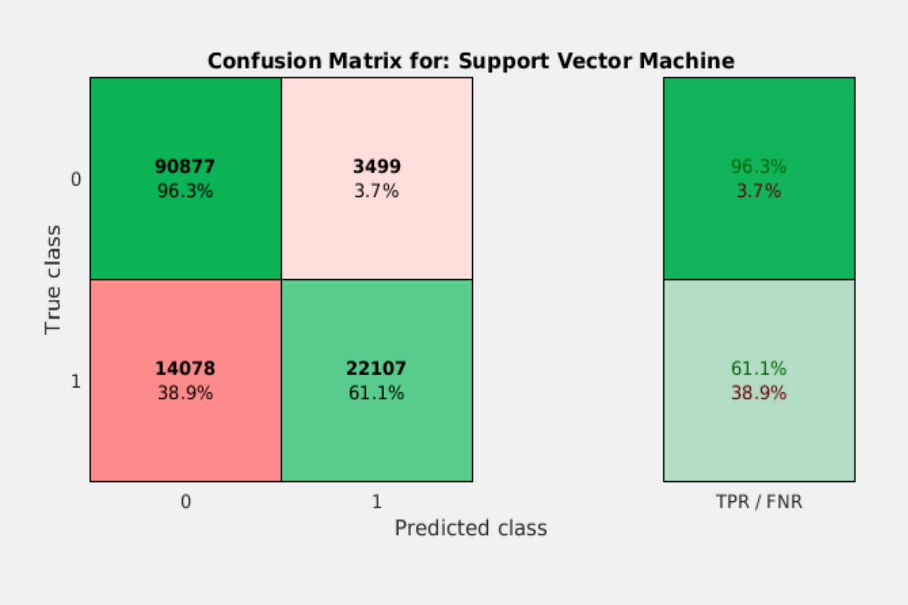
	Samples	MSE	R
Training:	91393	8.87778e-2	7.46829e-1
Validation:	19584	8.91462e-2	7.44569e-1
Testing:	19584	8.87592e-2	7.44010e-1

Support Vector Machine

Linear Kernel:

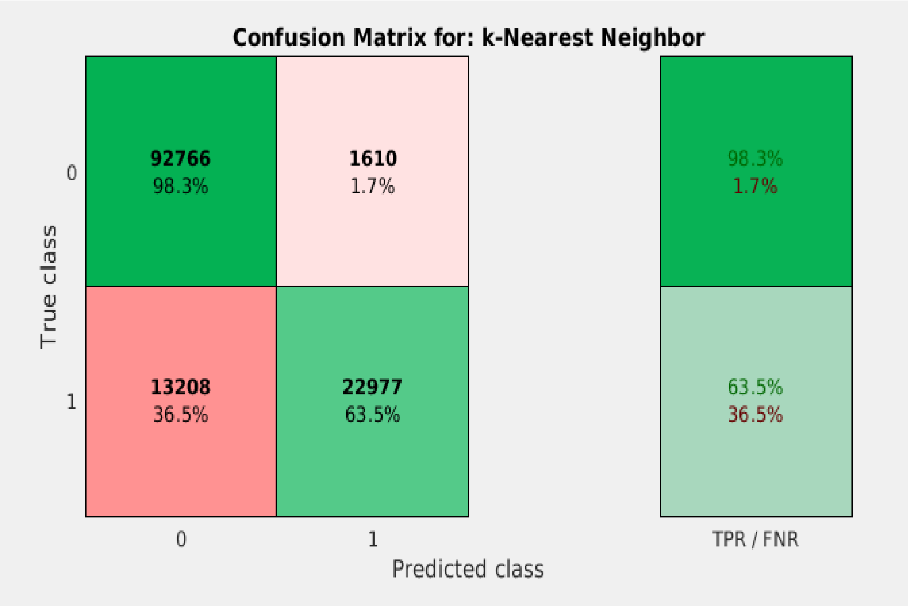


Gaussian Kernel:

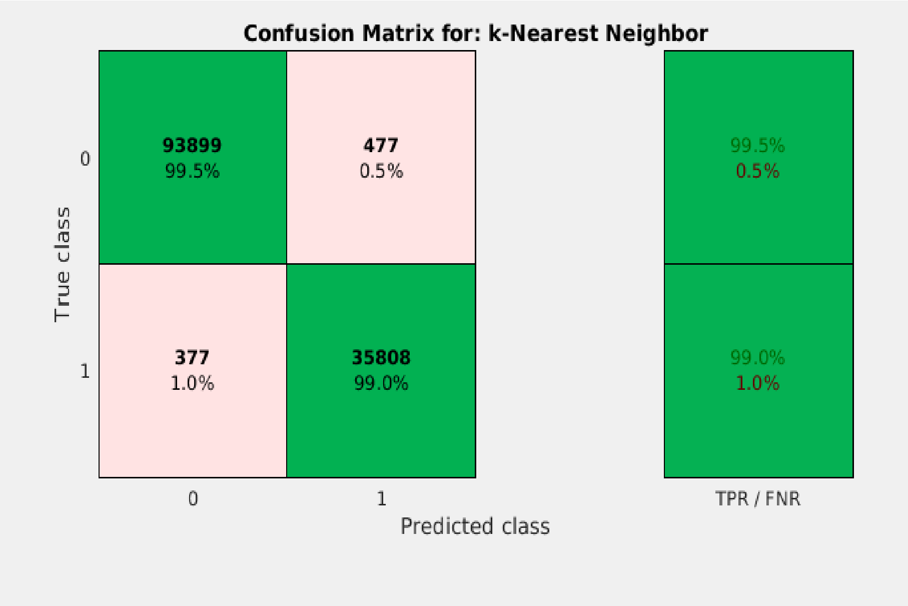


kNN Algorithm

Coarse KNN yielded:



Fine KNN yielded:



5 Concluding Remarks

The problem of speaker recognition, in its wide range applications, is complex and with a wide variety of solutions. The most common feature extracted from speech signals for this purpose are the MFCCs, which yield some promising results. However, as shown above, the accuracies could use a lot of improvement. Thus, the delta-cepstral and delta spectral cepstral coefficients, when appended to the existing MFCCs, and also the pitch provided more information about the voice print of the speaker in question.

Various machine learning algorithms are used to classify speech signals. The most popular ones include support vector machines, artificial neural networks and k- nearest neighbour algorithms. All three algorithms performed much better in the later case with additional features added. Further steps would involve multiple classes and tuning of the features of the various algorithms. Unsupervised machine learning algorithms are also to be tested.

References

[1] 1) Robust Speaker Recognition: A feature based approach by RICHARD J. MAMMONE, Clayton J. Pricerhe, XIAOYU ZHANG and RAVI P. RAMACHANDRAN

[2] Automatic speaker recognition using a unique personal feature vector and Gaussian Mixture Models by Kamil Kaminski and Ewelina Majda, Andrzej P. Dobrowolski

[3] The Research of speaker recognition based on GMM and SVM by Huo Chun bao and Zhang Cai juan

[4] Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models by Douglas A. Reynolds

[5] Speaker Recognition Using Weighted Dynamic MFCC Based on GMM by Zufeng WengLin Li* Donghui Guo

[6] Analysis of Compressed Speech Signals in an Automatic Speaker Recognition System by Richard A. Metzger and John F. Doherty and David M. Jenkins

[7] Performance Evaluation of an Automatic Forensic Speaker Recognition System based on GMM by Francesco Beritelli and Andrea Spadaccini

[8] Improving Text-independent Speaker Recognition with GMM by Rania Chakroun, Leila Beltafa Zouari, Mondher Frikha, and Ahmed Ben Hamida

[9] Support Vector Machines Using GMM Super vectors for Speaker Verification by W. M. Campbell, D. E. Sturim and D. A. Reynolds.

[10] A NEW STUDY OF GMM-SVM SYSTEM FOR TEXT-DEPENDENT SPEAKER RECOGNITION Hanwu SUN, Kong Aik LEE and Bin MA

[11] Support Vector Machines, Mel-Frequency Cepstral Coefficients and the Discrete Cosine Transform Applied on Voice Based Biometric Authentication by Felipe Gomes Barbosa and Washington Lus Santos Silva

[12] Mel-Frequency Cepstral Coefficients as Features for Automatic Speaker Recognition by Ivan D. Joki, Stevan D. Joki, Vlado D. Deli and Zoran H. Peri

[13] Mel Frequency Cepstral Coefficients Based Text Independent Automatic Speaker Recognition Using Matlab by Amit Kumar Singh, Rohit Singh and Ashutosh Dwivedi

[14] Text Indpendent Speaker Recognition Using Wavelet Cepstral Coefficient and Butter Worth Filter by Sandeep Rathor and R. S. Jadon

[15] Pseudo-pitch-synchronized phase information extraction and its application for robust speaker recognition by Longbiao Wang, Seiichi Nakagawa, Jianwu Dang, Jianguo Wei, Tongtong Shen, Lantian Li and Thomas Fang Zheng

[16] Speaker Recognition Based on the Improved Double-Threshold Endpoint Algorithm and Multistage Vector Quantization by Jun Zhu, Jingjing Zhang, Qiang Chen and Peipei Tu

[17] Two-Step Noise Reduction Based on Soft Mask for Robust Speaker Identification by Gennadiy Tupitsin, Artem Topnikov and Andrey Priorov

[18] Design and Implementation of an Automatic Speaker Recognition System using neural and fuzzy logic in Matlab by Shivam Jain, Preeti Jha and Suresh. R

[19] Study of speaker recognition system based on Feed Forward Deep Neural Networks exploring text dependent mode by Ben Jdira Makrem, Jema Imen and Ouni Kas

[20] Multi-condition Gaussian Probabilistic Linear Discriminate Analysis in Automatic Speaker Recognition by Mourad DJELLAB, Abderrahmane AMROUCHE, Noureddine MEHALLEGUE and Ahmed BOURIDANE.

[21] Analysis of Compressed Speech Signals in an Automatic Speaker Recognition System by Richard A. Metzger and John F. Doherty and David M. Jenkins

[22] 1) A. Nagrani, J. S. Chung, A. Zisserman VoxCeleb: a large-scale speaker identification dataset INTERSPEECH, 2017

Dataset consisting of labelled recordings of 1,251 celebrities collected from sources like YouTube.