Chapter 5

Speaker Recognition - Verification and Identification

5.1 Introduction

This chapter deals with the identification of features that help in recognizing the speakers participated in the communication scheme. Feature vectors extracted in the feature extraction module are verified using three different classifiers such as ANN, k-NN and SVM. Best features identified in this module are utilized towards the generation of watermark for the proposed schemes. Speaker recognition itself is a challenging area of research and an effort has been taken to identify some signal dependent features that helps in this recognition task.
5.2 Speaker Recognition

The most accepted form of identification for human is his speech signal. The speaker recognition process based on a speech signal is treated as one of the most exciting technologies of human recognition [Orsag 2010]. As prescribed in Chapter 2, audio signal features can be classified either in the perceptual mode or in the physical mode. In the proposed work, we have mainly employed the physical features towards the speaker identification activities. Speaker identification and speaker verification are the two schemes that are part of a speaker recognition task. As discussed in [Orsag 2010], the process of the speaker identification answers a question “Who is speaking?” On the other hand, the speaker verification answers a question “Is the one, who is speaking, really the one, who he is claiming to be?” The speaker recognition can be done in several ways and the commonly used method is based on the hidden markov models with the gaussian mixtures (HMM-GM) [Baggenstoss 2008]. But the proposed scheme employs ANN, k-NN and SVM classifiers.

Following figures 5.1 & 5.2 illustrate signal representations of male and female voices for the utterance of same sentence.
5.2. Speaker Recognition

Figure 5.1: Male voiced speech

- Short-term waveform of a Male Voiced Speech
- Frequency domain with different window functions (a) Short-term representation of a Male voiced speech, (b) 25 ms Hamming Window, (c) 30 ms Hamming Window, (d) 25 ms Rectangular Window, (e) 50 ms Rectangular Window

Figure 5.1: Male voiced speech
Figure 5.2: Female voiced speech
5.3 A Brief Review of Literature

As discussed earlier, speaker recognition itself is a broad research area which comprises of speaker verification and speaker identification. Aim of this section is to brief out some of the works conducted in the field of speaker recognition.

In 2010 Filip Orság [Orsag 2010] created a new speaker dependent feature called speaker dependent frequency cepstrum coefficients (SDFCC) for the purpose of speaker recognition. This scheme employs the speaker recognition technology based on the HMM with the newly generated feature sets (SDFCC). Even if these coefficients aim at the speaker recognition activities, in some special cases these may be usable for the speech recognition. SDFCCs differ from the MFCCs is in the utilization of a separate filter bank. In the paper [Pathangay and Yegnanarayana 2002], a text dependent audio-visual biometric person authentication scheme is suggested using dynamic time warping (DTW). This scheme uses a combination of features extracted from video and audio such as the mid-face vertical intensity and the LPC coefficients respectively.

The work presented in [Kavitha 2013] introduces an automatic speaker recognition system which works with the Matlab toolboxes and the feature employed towards this scheme is the MFCC. Douglas in his paper [Reynolds 2002] provides a brief overview of the area of speaker recognition, which includes its applications, various techniques used to achieve this, performance evaluations, its strengths and weaknesses and finally some future scope in this field. In [Peacocke and Graf 1990], give a description on speech as well as speaker recognition schemes. They stated that, speaker recognition is related to speech recognition. When the task involves identifying the person talking rather than what is said, the speech signal must be processed
to extract measures of speaker variability instead of being analyzed by segments corresponding to phonemes or pieces of text one after the other. In his paper [Melin 1999], suggests an overview of the activities in this working group (WG). WG2 of the COST250 Action “Speaker Recognition in Telephony” has dealt with databases for speaker recognition. First results demonstrate an overview of 36 existing databases that has been used in speaker recognition research. Second result is the publicly available Polycost database, a telephony-speech multi-session database with 134 speakers from all around Europe.

[Ghosh et al. 2004] suggests a speaker recognition scheme which employs artificial neural network for the purpose of identifying the speakers participated in the communication by employing features such as MFCC and LPC. The work presented in the thesis [Feng 2004], introduces a new method to combine the pitch information with MFCC features for identifying the similarity in k-NN algorithm. This scheme improves the speaker pruning accuracy. Experiments were conducted on HMM modeling and recognition with different setups. A comparative study is conducted on the error rate obtained with and without speaker pruning. The work suggested in [Stolcke and Ferrer 2012], demonstrate that a speech recognizer trained on full-bandwidth, distant-microphone meeting speech data yields reduced speaker verification error for speaker models based on MLLR features and word-N-gram features. [Ferrer et al. 2011] introduces a new database created using data from NIST SREs from 2005 to 2010 for evaluation of speaker recognition systems. As the paper suggest, this database involves types of variability already seen in NIST speaker recognition evaluations (SREs) like language, channel, speech style and vocal effort and new types not yet available on any standard database like severe noise and reverberation.
The work presented by [Ellis 2001], entails the design of a speaker recognition code using Matlab. Speech signals are handled by analyzing its time and frequency domain and using a 3rd order Butterworth filter removes the background noise to a great extent. In his thesis [Jin 2007] focuses on improving the robustness of speaker recognition systems on far-field distant microphones. The work is performed in two dimensions. First, they have investigated approaches to improve robustness for traditional speaker recognition system which is based on low-level spectral information and in turn introduces a new reverberation compensation approach. Second, they have investigated approaches to use high-level speaker information to improve robustness which in turn introduces techniques to model speaker pronunciation idiosyncrasy from two dimensions: the cross-stream dimension and the time dimension. The documentation given in the paper [Biometrics.gov 2006], provides a better understanding in the field of speaker recognition - it discusses a brief introduction, some history behind it and the approach employed towards the speaker recognition tasks. In the paper [Hennebert et al. 2000], presents an overview of the POLYCOST database dedicated to speaker recognition applications over the telephone network. This paper describes the main characteristics of this database such as medium mixed speech corpus size (>100 speakers), English spoken by foreigners, mainly digits with some free speech, collected through international telephone lines, and minimum of nine sessions for 85% of the speakers. In their work presented in [Barlow, Booth, and Parr 1992], created two speech databases intended for the purpose of speaker recognition which includes a very large laboratory access database and small departmental database. The aim of this work was to overcome the drawbacks of existing speech databases towards the speaker recognition activities. In the work presented by [Kinnunen 2005], a better text-independent speaker recognition strategy has
been discussed which employs better features or better matching strategy or a combination of the two towards its recognition tasks.

[Liu, He, and Palm 1996] in their work reveals that among the various parameters such as pitch, LPCC, $\delta$LPCC, MFCC, $\delta$MFCC that are extracted from speech signals, LPCC and MFCC as well as a combination of these features with pitch are effective representations of a speaker and thus help in the recognition tasks to a great extent. In the paper [Carey et al. 1996] demonstrates the importance of prosodic feature in the recognition process and its performance is measured using HMM models. Gaussianization, another important process performed for the speaker verification is presented in [Xiang et al. 2002] by employing the gaussian mixture models (GMM). A viable alternative to GMM systems is proposed by [Zilca 2001], functions through two verification methods, namely, frame level scoring and utterance level scoring. Studies of [Yu, Mason, and Oglesby 1995] compares HMM, DTW and VQ for speaker recognition and identified that for text-independent speaker recognition, VQ performs better than HMMs and for text-dependent speaker recognition also DTW outperforms VQ and HMM based methods. Another technique proposed by [Inman et al. 1998] derives the segment boundary information from HMMs [Russell and Jackson 2005] which in turn provides a means of normalizing the formant patterns. An HMM based text-prompted speaker verification system was suggested in the work presented in paper [Che, Lin, and Yuk 1996]. Employing an adaptive vocal tract model that emulates the vocal tract of the speaker towards the speaker recognition task was presented by [Savic and Gupta 1990]. A New set of features termed as adaptive component weighting (ACW) cepstral coefficients introduced by Khaled T Assaleh are utilized for speaker recognition presented in [Assaleh and Mammone 1994]. The papers [Linghua, Zhen, and Baoyu 2004; Pelecanos et al. 2000] introduces
the work that are intended for improving the performance of the existing schemes that works through VQ based gaussian modelling and training VQ codebook for HMM-based speaker identification. In the work [Rao et al. 2007], introduces a text-dependent speaker recognition system for Indian languages by creating a password authentication system.

In the paper [Kinnunen and Li 2010], provides a brief overview of the speaker recognition technology which comprises of the classical and state-of-the-art methods, recent developments and the evaluation methodologies adopted. [Garcia-Romero and Espy-Wilson 2011] introduces a method that helps to boost the performance of probabilistic generative models that work with i-vector representations in speaker recognition process. In the article [Sahidullah and Saha 2012], presents a novel feature extraction scheme that captures complementary information to wide band information. These features are tested on NIST SRE databases that work with GMM and obtained good performance results.

Many works have been reported in the field of speaker recognition. Out of these, few papers are collected and details of the works presented are reported over here.

5.4 Verification Process

General outline of the existing speaker verification schemes can be represented as in the following figure 5.3 [Reynolds 2002]:

The hypothesis that needs to be tested is: whether the test speech comes from the claimed speaker or from an imposter. In this process, features extracted from the speech signal are compared to a model representing the claimed speaker to a previous enrollment and also to some models of the imposter. Likelihood ratio statistic decides whether to accept or reject the speaker where the ratio represents the difference in log domain of speakers.

General techniques as outlined in figure 5.3 use three main components such as front-end processing, speaker models and imposter models in speaker recognition tasks. Front-end processing also termed as feature extraction is described in detail in the former chapter 4. Next is the speaker modeling that incorporates a theoretical foundation that helps in understanding the model behavior and the mathematical evaluations. Then generalize already enrolled data which helps to fit the new data appropriately and finally gives a mean representation of data in both size and computation.

Speaker modeling techniques that are employed in the proposed scheme include the nearest neighbor, neural networks and support vector machine models. In ANN technique all feature vectors from the designated speech are retained to characterize the speaker. The ANN methods used have
different methods like multi-layer perception or radial basis functions. The key difference in this model is that it utilizes explicit training to differentiate the speaker being modeled and other speakers. The drawback to this model is that training is computationally costly and cannot be generalized. In the case of k-NN classifier, with verification a match score is derived as the cumulated distance of respective feature vector to its k nearest neighbors in the speaker’s training vectors. Support vector machine doesn’t need an explicit model instead all feature vectors from the designated speech are retained to represent the speaker.

Third step is the imposter modeling which is critical for good performance results. It basically acts as a normalization to help minimize non-speaker related variability (e.g., text, microphone, noise) in the likelihood ratio score as presented in [Reynolds 2002]. First approach in this process is to calculate the imposter match score which is a function of maximum or average of the scores for a set of imposter speaker models. Second approach uses a general speaker-independent model that is to be compared with a speaker-dependent model. General model provides better performance than the first cohort scheme, as it allows the use of maximum A-posteriori (MAP) training to adapt the claimant model from the background model, which can increase performance and decrease computation and model storage requirements.

5.5 Speaker Recognition with Artificial Neural network

Artificial neural networks (ANN) are computational models inspired by animal central nervous systems that are capable of machine learning and
pattern recognition. They are usually presented as systems of interconnected “neurons” that can compute values from inputs by feeding information through the network [Encyclopedia 2013; Science and Engineering Laboratories 2010]. Many works have been reported in the field of speaker recognition using ANN models [Von Kriegstein et al. 2005; Syrdal and Gopal 1986; Von Kriegstein and Giraud 2006; Neto et al. 1995; Hambley 2008].

Neural network classification phase in this scheme needs a large set of feature data that are extracted in the feature extraction module from a diverse collection of recorded speech signals. ANN speaker recognition module has been conducted with the use of signal processing toolbox in Matlab.

5.5.1 Training Data

In this work, a total of $6 \times 7 \times 10$ Malayalam speech signals are selected for the use as training data for ANN network. 10 speakers, 5 males and 5 females volunteered to individually record the signals at different times. These signals include words as well as sentences with varying signal duration. These include ‘keralam’, ‘poojyam’, ‘shadpadam’, ‘vaazhappazham’, ‘keralathilae mughya bhasha aanu malayalam’, ‘malayalam polae thannae pradhanappetta bodhana maadhyamam aanu english’ and ‘tamizhu kannada bhashakalkku neunapaksha bhasha padavi undu’. Classification is done with the data from five male and five female speakers with 6 samples of each on the 7 words and/or sentences uttered. The ANN network trained is more robust and be able to differentiate between speakers of both sexes with related precision. Recording signals at different time guarantees that speech trials are pronounced autonomously of the preceding trials which
make it more lenient to the variations in a person’s voice that occur over short time duration.

Thus we have collected a total of 420 initial sound samples using the music editor sound recorder. Randomly selected utterances of each word from every person (109 training samples) were used as training data. The other two samples (59 testing samples) were used for validation and test purposes in the networks. Speaker modeling techniques performed confirms the signal characteristics that aid in speaker recognition. In perspective of the proposed watermarking schemes the samples collected holds good though more number of samples improves the accuracy of classification.

On obtaining the pre-processed sound samples for training purpose, nprtool in Matlab had been executed. Command nprtool open the neural network pattern recognition GUI. It is characterized by a hidden layer, logistic activation functions and back propagation algorithms. In order to test different combinations of data and features, the data sets are also grouped in different combinations and saved as different files. A total of 28 different data sets have been created with different combinations of speakers and features extracted from each of these samples. Values in the matrix or file is then arranged such that each speaker had the same number of training samples which helps to avoid the biasing of network towards any speaker. Rests of the values were utilized towards the validation and testing purpose. Also, the training data were arranged randomly to ensure that the network would not be trained on a long sequence that corresponds to any one speaker’s signal consecutively which in turn guarantees the training of the network more evenly among speakers.

Initially, four different types of inputs were used, as shown in Table 5.1, with six speaker combinations for each, as presented in Table 5.2.
Table 5.1: Types of inputs (420 inputs signals of 10 members)

<table>
<thead>
<tr>
<th>Type of Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 MFC coefficients (for a signal)</td>
</tr>
<tr>
<td>20 centroid values (for a signal)</td>
</tr>
<tr>
<td>20 short-time energy values (for a signal)</td>
</tr>
<tr>
<td>20 energy-entropy values (for a signal)</td>
</tr>
<tr>
<td>20 spectral flux (for a signal)</td>
</tr>
<tr>
<td>20 spectral roll-off (for a signal)</td>
</tr>
<tr>
<td>20 zero-cross rate (for a signal)</td>
</tr>
</tbody>
</table>

Table 5.2: Types of speech files (5 male and 5 female speakers)

<table>
<thead>
<tr>
<th>Type of Speech File</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isolated malayalam words</td>
</tr>
<tr>
<td>Continuous malayalam speeches</td>
</tr>
</tbody>
</table>

Likewise, total of $6 \times 7 \times 10$ data files have been employed, each used to train a separate network. This diversity should give an accurate assessment of the proposed algorithm’s ability to perform under different situations.

### 5.5.2 Testing Data

After completion of proper network training, next phase is testing the network with two types of data that falls under text-dependent and text-independent speaker recognition schemes. This is to compare the ability of the network to differentiate the speakers who have participated in the communication for these two types of data items.

As the first step, network is tested with randomly selected data sets for each of the words or sentences used in training. In an ideal case, the network should be able to recognize the speaker with maximum probability if he/she speaks a word or sentence used in training.
5.5.3 Experimental Results

Experiments were conducted based on a voice dataset and performance of the retrieved features were tested based on this. Voice dataset used for all the experiments that were conducted as part of this scheme was gathered from fellow research team members and colleagues through face to face interaction. This exercise gave a wide set of sample speech signals that are more or less similar to each other.

Voice database consists of 10 speakers, 5 males and 5 females. All speakers were in the age group 28 - 38 and collected a total of 6 samples for each 7 words/sentences they uttered. The speaker age group by virtue got restricted to 28 - 38 as the voice dataset gathered was from fellow research team members and colleagues through face to face interaction. Speakers age or the spoken language does not impact accuracy of the result set due to the fact that proposed schemes evaluate the physical signal characteristics of the respective speakers which is independent of the speaker age or spoken language. Recorded samples were used to test the quality of derived features towards speaker recognition. Speech samples in ‘.wav’ file format were recorded using the music editor sound recorder with a low signal-to-noise ratio. Sampling frequency for the recordings was 44,100 Hz with a precision of 16-bits per sample.

Feature Sets

Testing is conducted independently with individual feature sets for all selected samples and with a selected combination of 2 or more feature sets. First feature sets consists of 20 MFCC values or other feature values and the second type of feature sets includes different combinations of features
that are presented in table 5.1 for each speech samples. Feature sets selected for the training as well as the testing schemes varies depending upon the type or types of features selected in the process of classification.

**Evaluation Using Feature Sets**

In an experiment that have been conducted, evaluation of speaker recognition based on MFCCs are performed by employing 109 samples in training data set each with 20 MFCC values, 25 samples in validation data set and the rest 34 samples in test data set. In the four members, 2 of them are male speakers and the other two are female speakers. On multiple training, different results have been obtained; the reason behind this is that, its initial conditions and sampling for each iteration.

From this set, a particular iteration takes 27 samples of the first member, 29 samples of the second member, 26 samples of the third member and 27 samples of the fourth member in the training data set. That is, these are the data used in training algorithm to compute the adjustments for weights of connections. Then the validation data sets are used to confirm the networks generalization ability. Validation data sets in this case are 7 samples from the first member, 5 samples from the second member, 7 samples from the third member and 6 samples from the fourth member. Quality of the classifier is evaluated using the test data set. The test data set used in this case includes 8 samples from the first member, 8 samples from the second member, 9 samples from the third member and 9 samples from the fourth member.

Mean squared error (MSE) and percentage of errors (%E) obtained for a particular iteration is indicated in the following figure 5.4:
As we all know, mean squared error (MSE) indicates the average squared difference between outputs and targets. A low value of MSE indicates better results and zero means there are no errors.

Performance of this test can be understood from the following graph 5.5:

With MFCC feature 100% accuracy have been obtained in the speaker
recognition process. Then the same experiments were conducted for all other features and a combination of different feature sets mentioned in the tables - table 5.1 and table 5.2.

5.6 Speaker Recognition with k-Nearest Neighbor and Support Vector Machine Classifiers

5.6.1 k-NN Classifier

The k-NN classifier is a non-parametric method for classification and regression. This method predicts objects values or class memberships based on 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. k-NN algorithm is amongst the simplest of all machine learning algorithms. Using this classifier an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor [Encyclopedia 2013]. Different methods that employ k-NN for its speaker recognition activity are available in the literature [Lu, Zhang, and Jiang 2002; Lu, Jiang, and Zhang 2001].

5.6.2 SVM Classifier

In machine learning, SVMs (also known as support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Basic SVM takes a set of input data and predicts for each given input,
5.6. Speaker Recognition with k-Nearest Neighbor and Support Vector Machine Classifiers

which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that 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 on. SVMs are able to perform both linear and non-linear classification [Encyclopedia 2013]. Different algorithms that are proposed towards SVM speaker recognition are studied from the papers [Hatch, Kajarekar, and Stolcke 2006; Campbell et al. 2006; Solomonoff, Campbell, and Boardman 2005; Solomonoff, Quillen, and Campbell 2004; Campbell, Sturim, and Reynolds 2006].

As with the ANN classification scheme, both k-NN and SVM classifiers need a large set of feature data that are extracted in the feature extraction module from a diverse collection of recorded speech signals. Classification is conducted on different groups of data sets to test different combinations of features extracted from the speech signals using Matlab.

5.6.3 Training Data

Speaker recognition modeling with k-NN/SVM also uses features from the same speech database that are collected and thus includes $6 \times 7 \times 10$ speech samples. With the data from five males and five females speakers and with 6 speech samples of each 7 speech signals the k-NN/SVM classifier is trained so as to obtain good results with better robustness and might be able to differentiate between speakers of both sexes with similar precision.
A total of 420 initial sound samples were taken using music editor free sound recorder. First five utterances of each word from every person (350 training samples) were used as training data. Rest of the samples (70 testing samples) were used for validation and test purposes in these classification schemes.

In order to test different combinations of data and features, the data sets are also grouped in different combinations and saved as different files. A total of 28 different data sets have been created with different combinations of speakers and features are extracted from each of these samples. In order to avoid biasing of the classifier towards any speaker, the number of samples selected for training should be same for all speakers. Also an attempt is given to create the training set in a rather random manner that helps the classification scheme to be not trained towards any speaker specifically and also to train the network evenly among the speakers.

As in ANN, four different types of inputs, as shown in table 5.1, with ten speaker combinations for each, as presented in table 5.2 are used.

\section{5.6.4 Testing Data}

Once the training is done, next phase involves testing the extracted feature values with two types of data for text dependent and text independent speech recognition activities.

As the first step, network is tested with 1 or 2 utterances for each of the words or sentences used in training. Testing is done so that the classifier should be able to recognize the speaker with extreme probability for text-dependent data sets. Performance of the same scheme is also tested with text-independent data sets. Here what we have done is to select a set of test data that are not present in the train data sets. Comparison is performed
5.6. Speaker Recognition with k-Nearest Neighbor and Support Vector Machine Classifiers

on the text-dependent and text-independent data for single or different combinations of features.

5.6.5 Experimental Results

In this case, the same data sets as in the ANN classification scheme are used. As in the above case, the voice database consists of 10 speakers with 5 male and 5 female speakers in the age group of 28 - 38. Speech samples were recorded with a sampling frequency of 44,100 Hz with 16-bits per sample with the help of music editor sound recorder.

Feature Sets

Feature sets are created by first grouping the same feature values from different speech samples and then consider different combinations of these feature sets. The first feature set includes MFCC coefficients of 420 speech samples. Likewise testing is conducted independently with individual feature sets for all the selected samples, then consider different combinations of these feature sets towards the training and testing data sets. Thus the first group of feature sets consisted of 20 MFCC coefficients or other feature values listed in table 5.1. The second group of feature sets includes a combination of 2 different data sets or a combination of 3 different data sets and so on of the same table 5.1. The feature sets selected for training as well as testing schemes varies depending upon the type or types of features selected in the process of classification.

The classification scheme performed in this work employs two different speaker recognition schemes such as speaker verification and speaker identification. Speaker verification approach tests unknown samples with the train data sets. On the other hand, the speaker identification approach
takes only known samples with the train data sets. Another scheme employed here includes a combination of unknown and known samples to test with a set of train data. Of these three schemes, the third one matches more closely to the real world scenario.

**Speaker Verification Approach**

Following figures 5.6, 5.8 & 5.10 show the distribution of test data sets on the training data sets for k-NN and figures 5.7, 5.9 & 5.11 for SVM classifiers respectively. In speaker verification unknown samples are taken as the test data sets and the result obtained reveals the distribution of these data sets on the data sets available for the same speaker in the train data sets. In below figures, blue circles corresponds to the test data set, green circles represents the correctly classified items and red circles represents the incorrectly classified items on each of the classes and it reveals that using the MFCC values, speaker verification can be achieved to a good extent.

![Figure 5.6: k-NN speaker verification](image-url)
5.6. Speaker Recognition with k-Nearest Neighbor and Support Vector Machine Classifiers

Speaker Identification Approach

Speaker identification is the second approach in speaker recognition task. Here a known set of data have been taken for in classification and obtained good results on execution of the code. In these figures also blue circles corresponds to the test data set, green circles represents the correctly classified items and red circles represents the incorrectly classified items on each of the classes. Speaker identification can be achieved to a good extent by using the MFCC values.

Figure 5.7: SVM speaker verification
Figure 5.8: k-NN speaker identification

Figure 5.9: SVM speaker identification
Speaker Recognition Approach

In the third scheme a known set as well as an unknown set of samples have been employed against the train data sets. Intention behind this task is to make the work more realistic with data set that matches more closely to the real world scenario. Here also, blue circles corresponds to the test data sets, green circles represents the correctly classified items and red circles represents the incorrectly classified items on each of the classes and speaker recognition can be achieved to a good extent.

Figure 5.10: k-NN speaker recognition
Thus it can be concluded that both k-NN and SVM classifiers also give better results in the speaker recognition task for a known and unknown data set or for the combination of these two data sets.

5.7 Comparative Study of ANN, k-NN and SVM

This section deals with a comparative analysis of the speaker recognition results obtained on different features as well as on its different combinations. Following table 5.3 demonstrates the features and its combinations that are employed in the classification schemes.
Table 5.3: Feature selection

<table>
<thead>
<tr>
<th>Signal Features</th>
<th>MFCC</th>
<th>SF</th>
<th>SR</th>
<th>SC</th>
<th>ZCR</th>
<th>EE</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SF</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SR</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ZCR</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>EE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Below tables 5.4, 5.5 and 5.6 indicate the percentage of correctly classified speakers for the ANN, k-NN and SVM classifiers.

Table 5.4: Classification accuracy for single features

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>MFCC</th>
<th>SC</th>
<th>SR</th>
<th>SF</th>
<th>ZCR</th>
<th>EE</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>100%</td>
<td>95.1%</td>
<td>96.33%</td>
<td>97.5%</td>
<td>97.5%</td>
<td>89.5%</td>
<td>90.5%</td>
</tr>
<tr>
<td>kNN</td>
<td>100%</td>
<td>90.11%</td>
<td>85%</td>
<td>78.25%</td>
<td>77.25%</td>
<td>60.75%</td>
<td>64.25%</td>
</tr>
<tr>
<td>SVM</td>
<td>100%</td>
<td>91.23%</td>
<td>88%</td>
<td>79.33%</td>
<td>78.33%</td>
<td>61.23%</td>
<td>63.23%</td>
</tr>
</tbody>
</table>

Speaker recognition module is performed for individual signal features as well as for a combination of these features. This module enables us in identifying the signal features that help in speaker recognition tasks to a great extent. Features or feature combination identified in the process are employed towards the development of data codes which are treated as watermark in the proposed watermarking schemes.
Table 5.5: Classification accuracy for a combination of features

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>All Features (excluding MFCC)</th>
<th>SR &amp; SF</th>
<th>SR &amp; ZCR</th>
<th>SR &amp; SC</th>
<th>SR &amp; EE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>98.1%</td>
<td>88.50%</td>
<td>90.90%</td>
<td>90.90%</td>
<td>89.50%</td>
</tr>
<tr>
<td>kNN</td>
<td>89%</td>
<td>85.00%</td>
<td>82.50%</td>
<td>79.25%</td>
<td>76.50%</td>
</tr>
<tr>
<td>SVM</td>
<td>91.54%</td>
<td>80.00%</td>
<td>73.50%</td>
<td>76.92%</td>
<td>76.15%</td>
</tr>
</tbody>
</table>

Table 5.6: Classification accuracy for a combination of features

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>SF &amp; EE</th>
<th>SF &amp; SR &amp; ZCR</th>
<th>SC &amp; SR &amp; ZCR</th>
<th>SC &amp; SE &amp; SR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>87.50%</td>
<td>98.10%</td>
<td>90.90%</td>
<td></td>
</tr>
<tr>
<td>kNN</td>
<td>83.25%</td>
<td>98.10%</td>
<td>90.90%</td>
<td>86.25%</td>
</tr>
<tr>
<td>SVM</td>
<td>83.84%</td>
<td>88.54%</td>
<td>81.54%</td>
<td>83.85%</td>
</tr>
</tbody>
</table>

From the tables 5.4, 5.5 and 5.6, it can be concluded that the MFCC feature itself can be utilized towards the speaker recognition activity.

5.8 Summary

The speaker recognition or the classification module determines the best features that are employed towards the generation of signal dependent watermark. Features are selected in such a way that, they can by itself or with a combination of other feature sets have the ability to distinguish the speakers reliably. Initially, individual features are considered and then a combination of these features with different classifiers such as ANN, k-NN
and SVM. Thus the experimental results reveals that the classification module helps in the selection of ideal and most reliable characteristics (from the features that are extracted) for speaker recognition towards the preparation of the FeatureMark.