Text Independent Open-Set Cell phone Identification

Gyanendra Kumar Rahul
Department of Electronics Engineering
UIT University
Dehradun-248009, India
e-mail: gyanendrarahul@gmail.com

Devendra Singh
Department of Electronics Engineering
UIT University
Dehradun-248009, India
e-mail: eceuttaranchaluniversity@gmail.com

Abstract — This paper discusses the application of speech signals that convey various pieces of information such as the identity of its speaker, the language spoken, and the linguistic information about the text being spoken etc. The rapid developments in technologies related to cell-phones have resulted in their much broader usage than mere talking devices used for making and receiving phone calls. User-generated audio recordings from cell phones can be very helpful in a number of forensic applications. This thesis proposes a novel system for open-set cell-phone identification from speech samples recorded using the cell-phone. The proposed system uses different features based on original speech recordings and classifies them using sequential minimal optimization (SMO) based Support vector machine (SVM) and Vector Quantization (VQ). The performance of the proposed system is tested on a customised databases extracted from pre-recorded speech content of twenty-two cell phones of different manufacturers. Closed-set cell-phone recognition systems abound, and the overwhelming majority of research in cell-phone recognition in the past has been limited to this task. A realistically viable system must be capable of dealing with the open-set task. This effort attacks the open-set task, identifying the best features to use, and proposes the use of a fuzzy classifier followed by hypothesis testing as a model for text-independent, open-set cell-phone recognition.

Keywords*- Artificial neural network, cell-phone , WAV,AJFFt, Single line-to-ground fault, Double line to ground Fault, Three phase fault.*

*****

I. INTRODUCTION

Since we know that the primary source of communications between the people has been speech. This simple looking tasks carries much complex information within itself like gesture, emotions, expression, language etc. The speech science has been the most challenging areas of research. The most popular applications in the speech technology are speech recognition, speaker recognition, language recognition, speaker diarization, emotion recognition and gender recognition. With this large number of flexibility in the communication there arrives number of problems and issues relating to sense of understanding of listener and the capability of bluffing the person intended. Since the researches proves that the person almost lie at one time or another, it may create problem for others. Since speech is a natural signal voice can be characterize as biometric, the system and procedures have been devised to account for such forgeries. Forensic science is the area which deals with such type of the forgeries. If a person lies it can be found by biometric pattern recognition used by forensic science.

Since, now days in the current trend of latest and modern advancement in technology develops various gadgets for the human, which in wrong hands performs such forgeries at much larger space and with different purposes, which may create havoc among the society, leaked audio tapes, pirated content and digital speech content may be altered by malicious or amateur users are some examples in which users by using various low cost audio editing software can do forgeries and much more. To deal with such type of events other techniques to be developed. Therefore, multimedia forensic has been developed to take care of such type of forensic works. Although there are various number of system to capture devices from where specific image has been taken, but such devices are very less in case of the audio signals. The need however for such type of systems is more as lots of crime occurring these days involve significant audio evidences that can be analysed in many ways to obtain more information at larger scale for the purpose of compensating crime.

1.2 Classification of cell-phone recognition task

The cell-phone recognition is a generic term which can be used to refer two different tasks i.e. cell-phone identification and cell-phone verification. Presenting of ID cards wherever required is an act of verification. In the verification task, an identity claim is given to the system as an input and the system accepts or rejects the claim. However in cell-phone identification, only the input speech is provided and the system finds out the most resembling speaker with the help of code book. Identification encompasses two different types of applications known as ‘closed set identification’ and ‘open set identification’ which are both of academic interest as considerably different tasks. The closed set identification task aims to determine the identity of an input sample by comparing it with codebook/database made from the set of voice samples made from the known cell-phone from where input is assumed to come. However, open set identification system is aimed at detecting whether an unknown audio sample belong to a particular cell-phone but the search of it may or may not be confined to the present database. In open set identification it first decide whether the recording done from cell-phone or any other devices. In forensic applications it is common to first perform speaker identification process to create list of best matches and then perform a series of verification process to determine the conclusive results. In this thesis we are dealing with the open set identification problem, in which it is identified whether the whether the recorded voce sample even belongs to the known codebook/database or not.

Further, the brand and model of the cell-phone is identified, and then the system must be able to recognise the cell-phone irrespective of information available about recorded speech hence following a passive approach. Our goal in the thesis is to enhance the text independent open set cell-phone identification. The classification of cell-phone recognition task is shown in figure 1.1.

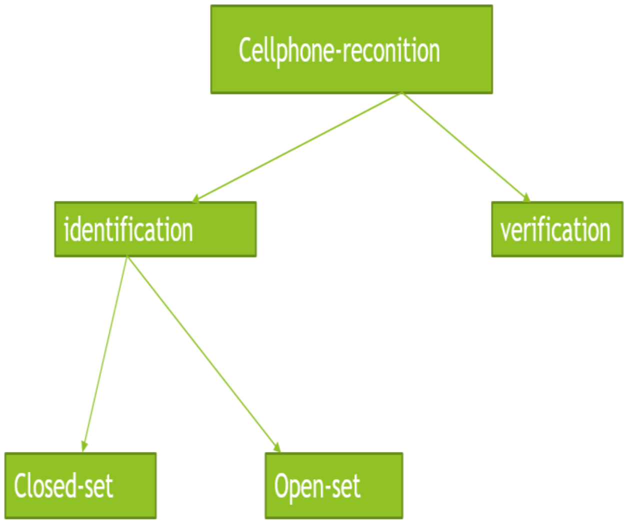


Figure 1.1.: Classification of cell-phone recognition tasks

1.3 The process of cell-phone identification

The process of cell-phone identification is divide into two main phase a) training stage followed by operational b) testing stage

.Training: In this stage features are extracted from each frame of cell-phone training utterance and the feature vectors are then further process to build a codebook/database. The same training feature vectors are subjected to fuzzy classifier and each frame is classified based on the maximum membership function value. A majority voting scheme of the frame is then used to classify the utterance, and those by frame membership function values corresponding to the winning cell-phone are stored at the cell-phone's reference membership value function (refU) . When this has been done then further we have testing stage.

.Testing: In this stage when test utterance is presented the features are extracted and fuzzy classifiers determine the most likely cell-phone based on frame by majority voting scheme. The membership function value corresponding to the winning cell-phone (testU) are then statistically compared to the cell-phone's reference membership function value (refU) for final decision whether to accept or reject the classification.

Based on the results obtained in the membership function values corresponding to the winning cell-phone in training stage and testing stage i.e. refU and testU, the hypothesis testing is done to determine the result about cell-phone identification. The identification process is shown in figure 1.2

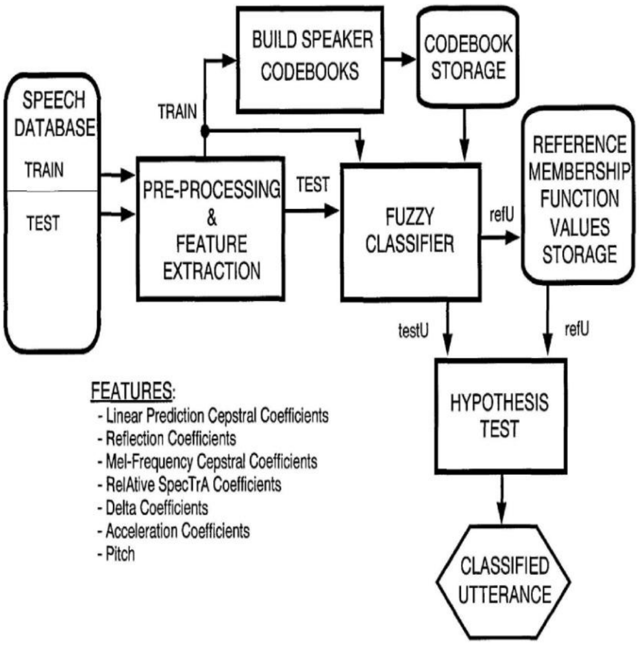
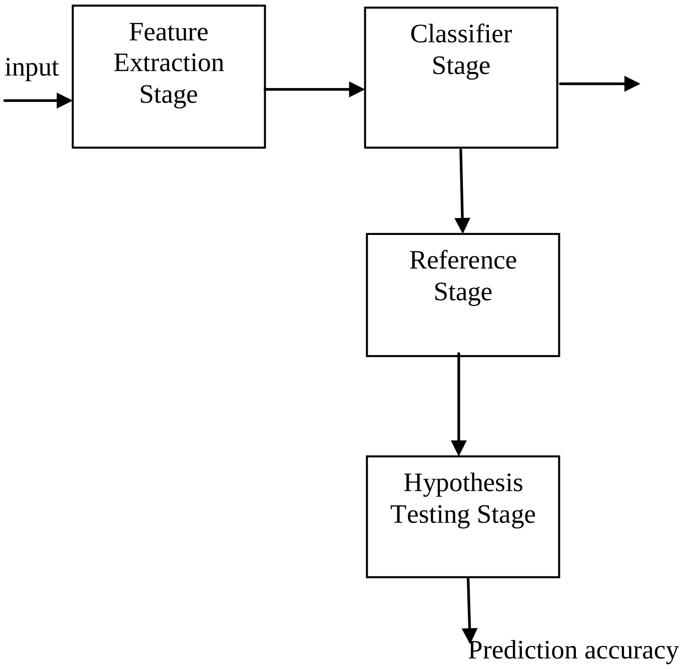

BUILD SPEAKER CODEBOOKS
CODEBOOK STORAGE
SPEECH DATABASE
TRAIN
TEST
TRAIN
PRE-PROCESSING & FEATURE EXTRACTION
TEST
FUZZY CLASSIFIER
refU
REFERENCE MEMBERSHIP FUNCTION VALUES STORAGE
testU
refU
FEATURES:
- Linear Prediction Cepstral Coefficients
- Reflection Coefficients
- Mel-Frequency Cepstral Coefficients
- RelAtive SpecTrA Coefficients
- Delta Coefficients
- Acceleration Coefficients
- Pitch
HYPOTHESIS TEST
CLASSIFIED UTTERANCE

Figure 1.2.: Two stage cell-phone identification

II. SYSTEM CONFIGURATION

Figure 2.1 presents a block diagram enlisting various stages of the system. Each stage highlights work done at each level. The complete system is divided into four stages: feature extraction stage, classifier stage, reference stage and hypothesis testing stage as shown in figure 4.1. At each level we are implementing the procedure to achieve a particular aim. The proposed work aims to design a proper hypothesis testing stage for better accuracy in results. We describe this stage in this chapter theoretically as well as its implementation. Further the implementation details of complete block diagram are discussed here.



Block diagram shows the implementation stages of the proposed work for designing open-set cell-phone recognition system. In actual implementation, these stages may be repeated more than once to obtain the results. The block diagram giving actual implementation is shown in figure 4.2 where noise estimation stage, feature extraction stage and clustering stage is used twice. In some cases of implementation, clustering appears at only model creation stage and not for test feature stage.

2.1 Intuition behind the proposed work

Different systems proposed in existing literature varied in terms of feature extraction and different classifier schemes. Some of them uses noise as the input signal for the better results. However, most of them extracted features either directly from the recorded speech signals or from the trimmed version of signal that is extracted based on comparing the speech signal with some pre-selected threshold. But in every situation all of them has performed their research on the closed set basis, even in open set only the speaker recognition task has been done but not cell-phone recognition task has been performed on open se basis. The proposed work gets its intuition from the work presented in **[1],** where the open-set task is performed for the text independent speaker recognition, in which hypothesis testing has been used to get result accuracy. Thus that experiment reveals that speaker identification can be achieved with better accuracy even in open-set tasks. We try to use same concept by using the defined procedure of feature extraction from the recorded speech signal with the purpose that better accuracy in feature extraction will reflect better traces of the cell-phones and the traces will be more pronounced than when considering the recorded speech. The proposed system is segregated into training and testing stages with number of stages of each as shown in figure 4.2. These sub-stages will now be discussed in detail.

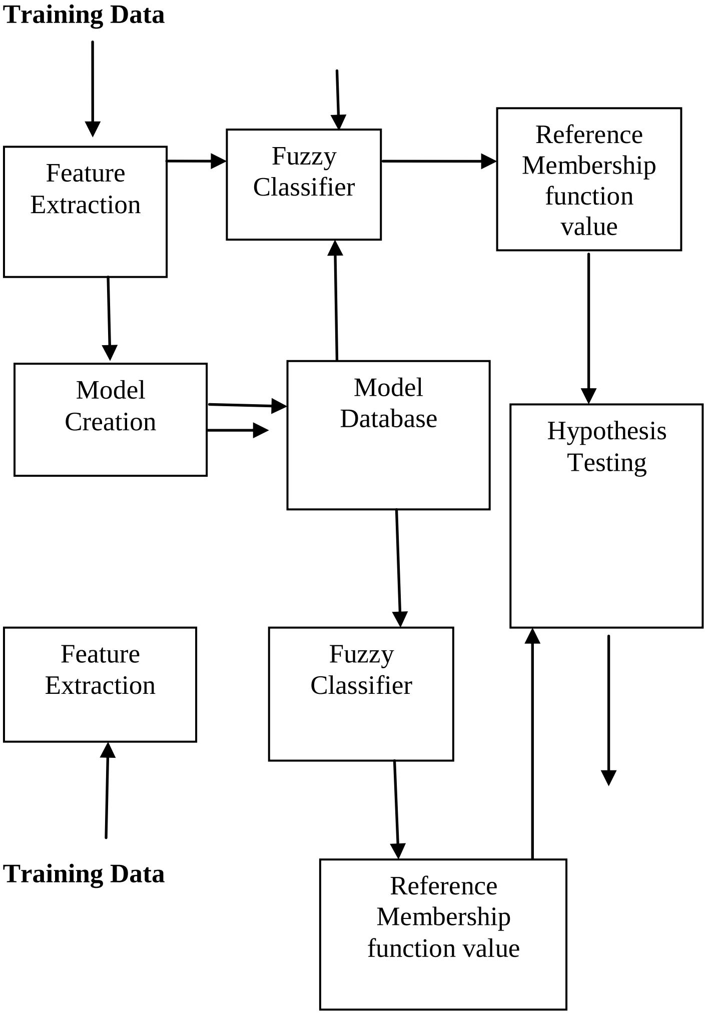


Figure 2.2: Overview of proposed system for speech recorder identification

2.2 Feature Extraction stage

In section 3 of the thesis, we have discussed different features used to model our system. The purpose of each scheme is to model the system so as to create a unique representation of it hence improving system's classification. In this section we discuss the implementation details of each feature extraction procedure as implemented in our experiment.

2.2.1 Mel-Frequency Cepstrum Coefficients (MFCC)

Initially, speech signal is framed into 30 ms frames with 15 ms overlap between subsequent frames. If the speech signal is not divided into exact number of frames, we discard last frames to avoid unnecessary interference of zeros. The framed signal is passed through a hamming window.

In the next step periodogram estimate of power spectrum is computed by taking DFT of each frame. We take the absolute value of complex Fourier transform and square the result.

Third step involves computing a mel-scale filterbank that is applying triangular filters to periodogram power spectral estimate. To calculate filterbank energy we multiply each filterbank with power-spectrum and ad up the results. The result gives us an indication of amount of energy in each filterbank. The filters are linearly spaced in mel-domain and triangular filters are implemented.

To the filterbank outputs, Log function is applied for scaling the spectrum envelope and multiplying by 20 to obtain the spectral envelope in decibels. MFCC is finally obtained by taking the Discrete Cosine Transform (DCT) of spectral envelope. It is the sum of all cosine functions oscillating at different frequency. The 0th coefficient contains only the dc component of energy for each frame and hence discarded. This coefficient if included would dominate the classification and hence discarded.

2.2.2 Linear Predictive Cepstrum Coefficient (LPCC)

The implementation of LPCC coefficients starts with dividing the signal into frames of 30 ms with 50% overlap between the frames. The framed signal is passed through a hamming windowed sequence. After windowing each frame LPC command is used to extract coefficients from one speech frame. These coefficients extracted for one frame is acted upon by a built-in function in lpcc2cep from rastamat. This function results in conversion of LPC coefficients to LPCC coefficients. The number of cepstrum coefficient is also taken equal to the number of LPC coefficients.

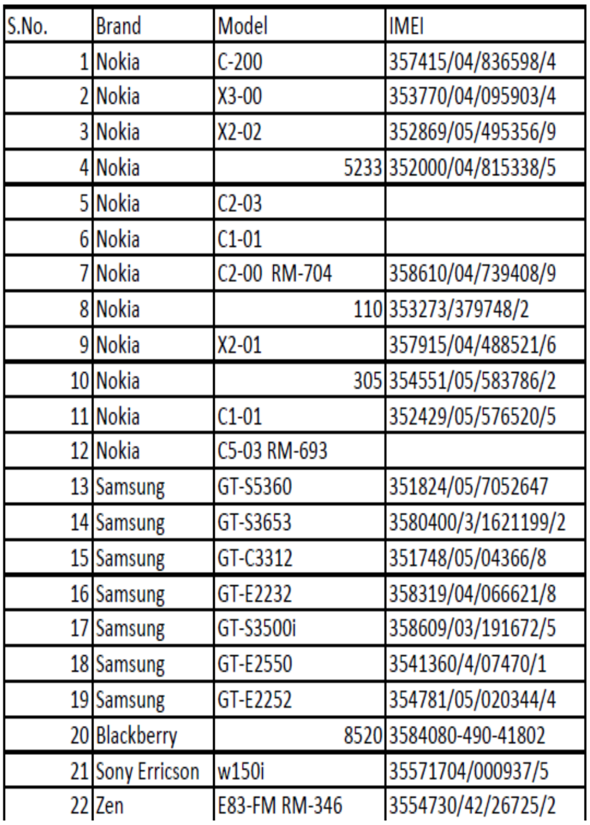
3.1 Speech database

Data collection is one of the basic requirements for any experiment. The results of the experiment are based on the protocol followed for the recordings. The speech database description is as follows. We have taken a database of 22 cell-phones with brand, model number and imei mentioned in the database. Each phone has one file named domain_specific of 5 min duration. Each speaker has to speak these three pre-defined sets of items:

- Domain-Specific Vocabulary speech which is continuously reading for 5 minute. Here we selected the paragraph of Indian Constitution.
- The last one is to record the silence of the recording room from every cell-phones.
- The calling party has continuously read a document related to Indian constitution up to 5 minute.

After the recording is completed, cell-phone specification chart used during the recording is prepared where the name of the cell-phones, brands, models, and lastly IMEI number of each cell-phone is quoted. It helps to discriminate one cell-phone to other which is of same brand. Here the tabular form of the cell-phones used in the recording is listed.

The table 5.2 shows the 22 cell-phones used in the recording with brand, model number and imei mentioned in the database. Each phone has one file named domain_specific of 5 min duration.



S.No.	Brand	Model	IMEI
1	Nokia	C-200	357415/04/836598/4
2	Nokia	X3-00	353770/04/095903/4
3	Nokia	X2-02	352869/05/495356/9
4	Nokia	5233	352000/04/815338/5
5	Nokia	C2-03	
6	Nokia	C1-01	
7	Nokia	C2-00 RM-704	358610/04/739408/9
8	Nokia	110	353273/379748/2
9	Nokia	X2-01	357915/04/488521/6
10	Nokia	305	354551/05/583786/2
11	Nokia	C1-01	352429/05/576520/5
12	Nokia	C5-03 RM-693	
13	Samsung	GT-S5360	351824/05/7052647
14	Samsung	GT-S3653	3580400/3/1621199/2
15	Samsung	GT-C3312	351748/05/04366/8
16	Samsung	GT-E2232	358319/04/066621/8
17	Samsung	GT-S3500i	358609/03/191672/5
18	Samsung	GT-E2550	3541360/4/07470/1
19	Samsung	GT-E2252	354781/05/020344/4
20	Blackberry	8520	3584080-490-41802
21	Sony Erricson	w150i	35571704/000937/5
22	Zen	E83-FM RM-346	3554730/42/26725/2

Table 3.1: Table shows the brands and models of cell-phones

WAV file: Waveform Audio File Format (WAVE, or more commonly known as WAV due to its filename extension). It is an application of the Resource Interchange File Format (RIFF) bit stream format method for storing data in "chunks", and thus is also close to the 8SVX and the AIFF format used on Amiga and Macintosh computers. It is the main format used on Windows systems for raw and typically uncompressed audio. The usual bit stream encoding is the linear pulse-code modulation (LPCM) format. WAV files are uncompressed, loss-less files in a linear code modulation (LPCM) format. Audio files in the WAV format have the maximum audio quality. Uncompressed WAV files are large so file sharing of WAV files over the internet is uncommon. Due to the un-compression of these types of files, it occupies more space than any other file format.

AMR file: The Adaptive Multi-Rate (AMR or AMR-NB or GSM-AMR) audio codec is an audio compression format optimized for speech coding. AMR speech codec consists of a multi-rate narrowband speech codec that encodes narrowband (200–3400 Hz) signals at variable bit rates ranging from 4.75 to 12.2 kbit/s with toll quality speech starting at 7.4 kbit/s. AMR was adopted as the standard speech codec by 3GPP in October 1999 and is now widely used in GSM and UMTS. It uses link adaptation to select from one of eight different bit rates based on link conditions. AMR is also a file format for storing spoken audio using the AMR codec. Many modern mobile telephone handsets can store short audio recordings in the AMR format, and both free and proprietary programs exist to convert between this and other formats, although AMR is a speech format and is unlikely to give ideal results for other audio. The common filename extension is .amr. AMR is a hybrid speech coder, and as such transmits both speech parameters and a waveform signal. It is a lossy, compressed and encoded file format. So the speech which is used for dataset is first recorded in AMR files then after convert it into the WAV file with the help of Audacity software. Audio data compression allows for more storage on voice files. ".amr" is a popular filename extension for AMR.

Since our aim is to evaluate systems for cell-phone identification and not for speaker or speech content identification. Therefore, the primary goal of the gathered database is to capture variations due to cell-phones keeping other factors fixed. The recorded speech in AMR format is converted into WAV format with the same specifications such as same sampling rate. In every experiment conducted in literature, speech is converted to WAV format and then used. The reason for using this format is that it has a set basis to work with. The format is in uncompressed form. Therefore, it is easily understandable and easily configured. Another format i.e. mp3 are compressed formats. Another reason for not using any compressed format is that there are many such formats and if we devise a system that works with one compressed method then there will be different implementation for each such compressed format that results from recorded speech. The use of uncompressed format leads to an implementation which is common to all the recorded speech signals. Any compressed format is initially converted to uncompressed WAV format and later processed using the methodology described to obtain necessary results.

3.2 Database for experimental purpose

Since the initial aim of experiment was to gather capture variations due to cell-phones, keeping other factors fixed. Hence, all the files from folder are taken. For fixed speech content, three audio files from same folder, either days, digits or months, is taken and MFCC features are extracted. Two audio files out of three were used for training and last one was used for testing purposes. Two files from both the folders are selected for training purposes and remaining one file from each folder is used for testing

From the extracted feature vectors training and test sets are configured as per the discussion for each of the selected databases and performance results are obtained for different train and test setups.

3.3 Experimental results of the set-up

Initially experimental setup is using only one type of feature i.e. MFCC coefficients computed for original speech and spectral subtraction output. The feature vectors are clustered using k-means and classified using SVM with Sequential Maximal Optimization algorithm that appears as default SVM classifier in WEKA **[4]**.

Each file from each folder was taken and MFCC features corresponding to it is extracted. In this scheme, clustering was performed on each audio file corresponding to each folder and hence resulting in comparatively large number of features for classification purposes. The obtained clusters per audio file per phone were used to train and test the classifier.

Initially, speech was hamming windowed into 30 ms frames with 15 ms overlap and STFT of frame was taken to predict the noise spectrum that is later used to extract MFCC coefficients from each frame along with their delta coefficients. K-means clustering with different codebook sizes k= 8, 16, 32, 64 and 128 is performed on MFCC feature vectors of each audio file to obtain a uniform and compact feature matrix. The number of features after clustering equals:
k*number_of_audio_per_folder*number_of_folder_per_phone * number_of_phones

For eg: for k=8 it results to: 8*3*3*26 = 1872

The dataset is randomly split into two non-overlapping subsets to be used for training and testing of the proposed system.

The performance results are represented in terms of confusion matrix. Confusion matrix depicts actual instances along the rows and predicted class instances along the columns. The diagonal elements of such matrix correspond to percentage of correctly classified instances for a particular class. Confusion matrix provides an easy way to conclude whether (and how much) the system is confusing different classes with each other.

The results presented are obtained for cluster size k . This cluster size is selected by observing the fact that when results in terms of average classification accuracy is computed, then the results obtained with k cluster size.

After that the lpcc is extracted with the help of the same database and clustering size used for mfcc.

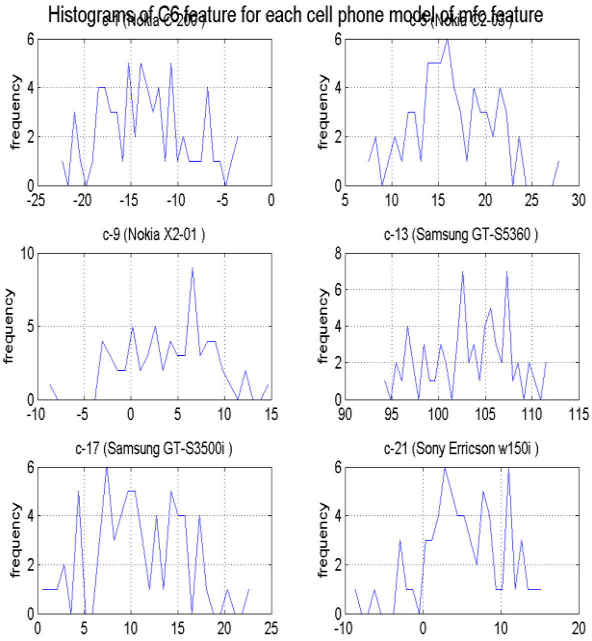


Figure 3.3 : Histogram of mfc feature of cellphones

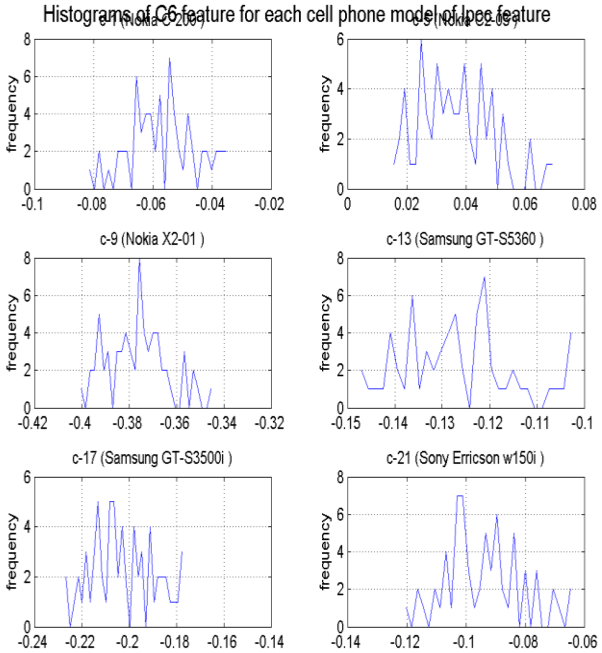


Figure 3.4: Histogram of lpcc feature of cell-phones

Similarly, the results of the onwards processes has been taken out to complete the experiment. After the completion of the hypothesis testing the required results output gives the classified utterances which will finally show the required results.

Conclusion

The aim of this paper is to present a novel approach for classifying the audio recordings from cell phones can be very helpful in a number of forensic applications such as securing the information left behind at a crime scene. The thesis presented a system for identification of open-set cell-phone from audio recordings initially using MFCC features and LPCC features corresponding to each recording. Database is one the important part of the any experiments. Result of any experiment is depends upon what type of database is being used. Work with short in time and length of database type will generate different results than lengthy database and hence, may not be a generalised or robust in open-set, since the recording may or may not belongs to the database.

Each utterance is classified as belonging to the speaker of cell-phone to whom the majority of the frames are classified, and the membership function values associated with those winning frames are used in hypothesis testing.

Another important result is with use of different clustering schemes where we found that k-means outperforms fuzzy c-means in terms of classification results obtained. In each case, the classification achieved using k-means is high compared to fuzzy c-means. For some feature vectors, the accuracy is similar using either approach while for others the k-means achieves higher accuracy.

References

[1] Text-Independent, Open-Set Speaker Recognition, THESIS Presented to the Faculty of the School of Engineering of the Air Force Institute of Technology, Air University In Partial Fulfilment of the Requirements for the Degree of Master of Science in Electrical Engineering.

[2] Cemal Hanilçi, Figen Ertaş, Tuncay Ertaş, and Ömer Eskidere, "Recognition of Brand and Models of Cell-phones from Recorded Speech Signals," IEEE Transaction on information forensic and security, vol. 7, No. 2, pp. 625-634, April 2012.

[3] Cellphone Identifications Using Noise Estimates from recorded Audio, Rachit Aggarwal, Shivam Singh, Amulya Kumar Roul, Nitin Khanna.

[4] Mark Hall et al., "The WEKA Data Mining Software: An Update," SIGKDD Explorations, vol. 11, no. 1, 2009.

[5] Lawrence Rabiner and Biing Hwang Juang, Fundamentals of speech recognition. New Jersey: PTR Prentice Hall Inc., 1993.

[6] Lawrence R. Rabiner and Ronald W. Schafer, Digital processing of speech signals. Englewood Cliffs: Prentice Hall, 1978.

[7] Alan V. Oppenheim, W. Schafer, and John R. Buck, , Ronald W. Schafer, and John R. Buck. Discrete-time signal processing. Vol. 2. Englewood Cliffs: Prentice-hall, 1989.

[8] Fuzzy Classification System by Self Generated Membership Function Using Clustering Technique, Shruti S. Jamsandekar1 and Ravindra R. Mudholkar2, Submitted in August, 2013; Accepted in February, 2014.