I.J. Image, Graphics and Signal Processing, **2016, 9, 17-25**
Published Online September 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2016.09.03

Speech Feature Extraction for Gender Recognition

Anjali Pahwa
Department of CSE, THE NORTHCAP University, Gurgaon, Haryana, India
Email: anjalipahwa1912@gmail.com

Gaurav Aggarwal
Department of CSE, THE NORTHCAP University, Gurgaon, Haryana, India
Email: mtech.gaurav@gmail.com

***Abstract*—**Speech Recognition Technology can be embedded in various real time applications in order to increase the human-computer interaction. From robotics to health care and aerospace, from interactive voice response systems to mobile telephony and telematics, speech recognition technology have enhanced the human-machine interaction. Gender recognition is an important component for the application embedding speech recognition as it reduces the computational complexity for the further processing in these applications. The paper involves the extraction of one of the most dominant and most researched up on speech feature, Mel coefficients and its first and second order derivatives. We extracted 13 values for each of these from a data-set 46 speech samples containing the Hindi vowels (आ, इ, ई, उ, ऊ, ऋ, ए, ऐ, ओ, ऑ) and trained them using a combined model of SVM and neural network classification to determine their gender using stacking. The results obtained showed the accuracy of 93.48% after taking into consideration the first Mel coefficient. The purpose of this study was to extract the correct features and to compare the performance based on first Mel coefficient.

***Index Terms*—**Gender recognition, Hindi, mel-frequency, delta, delta-delta, neural network.

I. INTRODUCTION

Speech is the most convenient medium of communication among the human beings. Speech signals vary highly along with time and are the most random signals. Researches have been done in order to have such easy interaction between humans and machine [1]. As we know that technology is growing at a rapid rate which include enhancements in technologies like that of big data, innovation of various new algorithms, machine learning etc. Along with these the field of speech recognition has also developed. It was first introduced by Microsoft in China when it developed a real time application that translated English into Chinese. However, these advancements of technology have somewhere created a hindrance between the humans and development. Information Technology is increasing its contribution in India's development day by day, year by year. In such case to make IT relevant to the rural India, voice access to a variety of computer based applications is imperative [2]. But the major issue is the dearth of effective and efficient speech recognition systems in Hindi. Research in gender determination using speech recognition system has been the in the lime light from over a decade. Speech recognition involves real data and is used in combination to machine learning. Machine learning being the versatile field of artificial intelligence concentrates on learning from the data and is used in various other fields like finance, banking etc. Many practical applications exist that enhanced the human-computer interaction or information retrieval [3]. It not only improves system intelligibility but is also useful in surveillance systems. The speech production takes place naturally in humans and the lungs are the main source from where the air is discharged and is converted to speech passing through the wind pipe where the vocal cords are present. The main cause is the vibration of the vocal cords for the conversion of air to sound and the organs like mouth, lips, tongue, teeth etc., converts the sound produced further into speech, that is, the meaningful word.

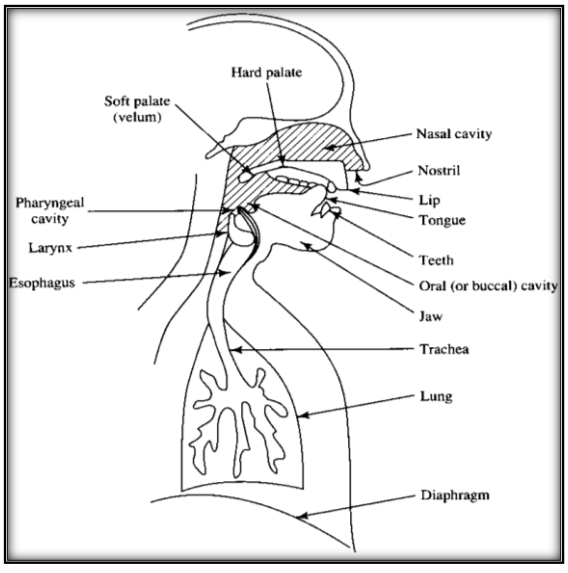


Fig.1. Human Voice Production Important Parts

This phenomenon produces speech that can be categorised as:

- Voiced Speech: When the air is discharged from lungs it moves through the windpipe to the vocal cords, if the vocal cords vibrate and breaks the air into periodic signals then the speech produced is termed as voiced speech.
- Unvoiced Speech: When there is vibration of the vocal cords and the air is broken into aperiodic signals then the speech produce is termed as unvoiced speech.
- Silenced: When there is no vibration of the vocal cords after the air is discharged from the lungs.

The **fig. 1** explains the human voice production system important parts. The detailed process of speech production is depicted in [3]. The **fig. 2** represents the three types of speech produced.

Speech processing field contains a lot of techniques that can be used to enrich the quality of the speech signal. But the real time used of the system proposed is very necessary as it gives the ground reality of the work done.

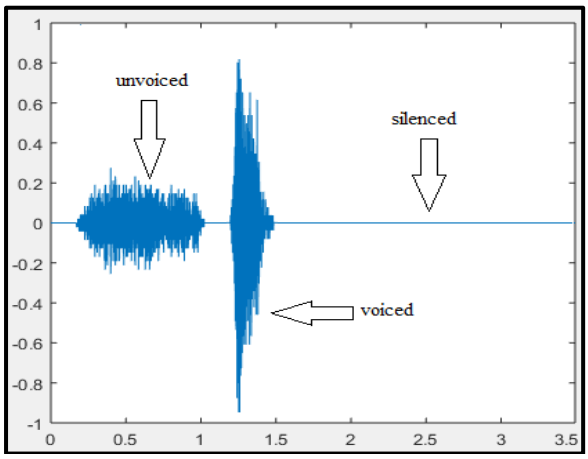


Fig.2. Types of Speech Produced

In this paper, we built a gender recognition system in Hindi to determine the gender using speech. We took speech samples of 46 speakers and pre-processed the signals and extracted there MFCC, first and second derivatives. We then trained the system using combined SVM and neural network classifier using the process of stacking of classifiers in a tool named RapidMiner Studio. the speech database prepared contains two datasets with or without the first value of the mel coefficients obtained. And as mentioned above the model has been trained and tested using the speech samples collected from the real time environment that has a considerable amount of background noise while recording. The work done has been explained in following sections.

II. Related Work

There have been a lot of developments in the versatile field of speech Recognition technology since it is used in collaboration with the various machine learning algorithms. The speech recognition models have been trained by extracting various features, combining various kinds of features. In addition to this the different type and combination of classification algorithms used also contributed much to the progress. The authors of [4] created a combined model by using the methods GMM, SVM, GMM-Mean SVM. The accuracy obtained was the best in the third case, which was the combination of the first two. Another work related to this was the building of the home robot service for gender and age determination that extracted MFCC speech feature and the model was trained using the GMM, MLP classifiers and gave an accuracy of around 94% for the gender determination [5]. The related work is not confined to these boundaries.

In [6] the author presented the function of various features like pitch, formant, MFCC in the view of increasing the rate of identification. Pitch extraction was done using the converter technique that transforms the signal into cepstrum domain and used the homo-morphic analytical method for eliminating the channel impact. The information from the pumping part was then obtained and the fundamental frequency was gained. The paper adopted the linear forecasting method for formant estimation as the linear analysis is faster than the others. The LPC is computed first, which is then used to compute the spectrum and formant was computed according to the spectrum. For MFCC computation the series of predefined steps were used. And based on MFCC the first and second order difference was computed. These enhanced the accuracy of recognition. The conclusion drawn from the experiment was that pitch and formant were considered important features for speech as these are different in everyone. MFFCC is effective parameter as it works analogical to human ear. In [7] the author introduced the speech recognition system by giving a brief introduction about the speech and the human-computer interaction and also the speech recognition system's application in our daily lives. This paper aimed to represent the MFCC feature extraction process in detail and hence describe all the steps in detail. The steps were carried out in detail and also the delta coefficients were calculated based on these. The thirty-nine MFCC coefficients were extracted. The conclusion made was that the MFCC feature extraction technique is robust and can even help to normalize the features. The future work indicated was to design a speaker independent speech recognition system using the MFCC features.

The author in [8] presented the work for finding the gender of the users in order to be used for the telephone application. The main purpose was to find out the performance of the various classification algorithms [8]. The classification techniques used were – naïve baye's, KNN, MLP, SVM and the random forest. The telephonic conversation of the people was used as the database. The feature extracted was MFCC and the experiment was performed based on different on sizes of training data, different durations of the signal, and different parameters of the MFCC so that better performance can be obtained.

The results obtained showed Support Vector Machine gave the best performance in all the five classifiers used. The system also analysed the emotions of the users for the emergency calls.

The work proposed [9] extract MFCC using the typical steps and classified the user using the Artificial Neural Network classifier. The gender recognition and the speech recognition require same feature and so these two processes can be combined to form a system [9].The main advantage is that no assumptions in the voice creation process [9]. The system was implemented using the speech samples recorded in English and Polish language. The results obtained depicted that Artificial Neural Network classifier gives high accuracy when used for short utterances in the recognizing gender. And combining the gender recognition and speech recognition gave good performance. The system proposed worked on the fundamental frequency (f_0) of male and female and compared it with mel-frequency cepstral coefficients based model to give good results. And finally, the accuracy for gender recognition came out to be 98.4% when implemented using mel-frequency cepstral coefficients feature extraction and 95.1% when implemented using the f_0.

III. Methodology

In the research work, we proposed a system that provides information about the gender of the speaker. A wav file containing the speech sample of the speaker is given as input and how these techniques can be used to make the speech recognition systems robust and efficient. For gender identification we evaluate a speech database recorded from fifty speakers, which contains their speech samples. Then we calculate MFCC, delta and delta-delta speech features (each having thirteen values) for each of the audio file. After extracting these features from all the audio files we stored them into a database. Then we used cross validation technique for training and testing the dataset and recognition of the gender (male, female).

Speech is the most convenient and fundamental way of communication. The speech signals are naturally occurring signals and random signals. The following steps describe the major processes involved in the development of the system.

A. Data Selection

We collected the speech samples and form the dataset of audio files. All the speech samples collected had the time duration in seconds. It is important that the data to be re3corded should be noise proof and there should be no background noise present during the time of recording. The data was collected from twenty males and twenty six females and the speakers were made to speak all the vowel phonemes of Hindi (i.e., आ, इ, ई, उ, ऊ, ऋ, ए, ऐ, ओ, ऑ). The speech was recorded using the software Audacity.

B. Pre-processing

After the creation of the speech database, we closely examined the database. It was found that the speech signals obtained from the audio dataset cannot be directly used as input in the feature extraction code. There were long silences, weak signals present in the speech samples and need refinement. The weaker signals were amplified in during the feature extraction process and long silences were removed by performing pre-processing of each signal. **Fig. 3(a) and (b)** below represents the original signal and the signal after pre-processing respectively. We also used the VAD technique for detecting the voiced part of the speech signal and hence further processed the signal by only considering the voiced part and removing the unvoiced of noisy parts.

C. Feature Extraction

This step transformed the random speech signals into compact representations that were far more stable than original signal. There exist a number of speech features that can be extracted from a speech signal. These include pitch, energy, mel-frequency cepstral coefficient (mfcc), first order derivative (delta), second order derivative (delta-delta), formant etc. In this research we concentrated on the extraction of three features that are mfcc, delta and delta-delta features. We extracted the first thirteen values each of mfcc, delta and delta-delta. And obtained 39 numerical representations for each audio signal and prepared an extracted feature database.

D. Classification

This step involves the training, testing and the validation of the data. There are various classification algorithms that can be applied in order to get the accurate detection in the system. In fact, the different classifiers are combined together to give better results. The classification method used was stacking, where SVM and NN were stacked together and the combined prediction was made by Naïve Baye's.

Fig. 4 shows the system training and testing model.

IV. Feature Extraction

It involves the extraction of apt information from the speech signal. The feature extraction process basically represents the speech signal in a numerical representation. The pre-processed signal is given as input to this step and output obtained id the numerical representation with respect to different features. There exist various speech features of speech that can be extracted to get the gender information of the user but the most dominating is the mel-frequency cepstral coefficients along with its first and second derivatives. These three features are explained in detail as follows.

A. Mel Frequency Cepstral Coefficient Measurement

MFCC is one of the most dominating features among all the features of speech. MFCC makes a close analogy with the human ear by considering those parameters extracted by human ear.

In the production of speech the air expelled from lungs passes through the oesophagus resulting in the vibration of the vocal cords, turning the air into the quasi-periodic signals. After this, there are certain modulations in the frequency that highly depend on the shape of the vocal tract that includes oral cavity (includes the mouth, tongue and lips), the pharyngeal cavity (including the throat) and the nasal cavity [12]. This shape is represented by an envelope. If the human ear accurately determines this shape then it can make out exact phoneme being produced. MFCC determines this shape accurately and represents it's envelop [11].

This section describes how each step is carried out during the MFCC feature extraction. The following block diagram in **fig. 5** represents the MFCC feature extraction process.

The steps of implementation carried out for MFCC feature extraction and the significance of each step are explained as follow:

- **Pre-emphasis**

At this step the signal passed through a filter, which emphasizes higher frequencies in the signal. That is, it increases the energy of the signal wherever it is low and also compensates high frequency parts of the speech signal.

$$Z`(n) = Z(n) - \alpha * Z(n-1) \quad (1)$$

the value for α is called as pre-emphasis coefficient and ranges between 0.9 to 1.0.

- **Framing**

In this step the speech signal is segmented into blocks that are called frames. We segmented the signal into 25 ms frames (standard size). Assuming the sample rate of 8 kHz, frame length for the signal was 0.025*8000 = 200 samples. Frame shift taken was of 10ms (80 samples) duration, which allowed some overlap to the frames. The first 200 sample frame started at sample 0, the next 200 sample frame start at sample 80 etc. until the end of the speech file is reached.

- **Windowing**

The next step in the processing is the windowing of each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. Each frame produced at previous step is multiplied with a hamming window, that is, S(n) * h(n), where w(n) is the hamming window,

$$h(n) = 0.54 - 0.46 * \cos(2\prod n/N-1) \quad (2)$$

the value of n ranges from 0 to N-1. The generalised hamming window is shown in fig. 6.

- **Fast Fourier Transform**

This step converts each frame of N samples from the time domain into the frequency domain using the following formula:

$$z_i(k) = \sum_{n=1}^{N} z_i(n)h(n)e^{-2\prod ke-2\prod ke/N} \quad (3)$$

where $S_i(n)$ is the signal in time domain, $S_i(k)$ is the signal in frequency domain ranging from 1 to K, h(n) is an N sample long window and K is the length of the FFT.

After the conversion to from time to frequency domain the next step involves the estimation of different frequencies present in the signal. It is called periodogram estimate which involves the calculation of the power spectrum for each frame.

This step is analogical to the organ present human ear known as cochlea. It is present in the middle ear and contains a viscous liquid that is sensitive to the vibration, that is, depending on the different frequencies present in the speech signal the liquid in cochlea vibrates at different places. Depending on which different auditory nerves carry the signal to the brain informing about the presence of the certain frequency [11].

For representing the periodogram estimate we take the absolute value of the FFT and square the result. The periodogram estimate for a speech frame is represented as:

$$P_i = \frac{1}{N}|S_i(k)| \quad (4)$$

Fig. 7 represents the power spectrum for a speech signal.

- **Mel Filter Bank**

As mentioned above, MFCC takes into consideration the human perception sensitivity in context to frequencies, the frequency domain representation is mapped to Mel Frequency using:-

$$M(f) = 1125\ln\left(1+\frac{f}{700}\right) \quad (5)$$

$$M^{-1}(m) = 700\left(\exp\left(\frac{m}{1125}\right)-1\right) \quad (6)$$

We used 20 triangular filter banks. The filters were narrow initially and increased as there was increase in the frequency. The **fig. 8** below shows the band of filters in Mel Scale as in [10]. The energy from each band is extracted and converted to hertz and the log of energy is calculated.

- **Discrete Cosine Transform**

This step involves the conversion of the signal back to time domain. The formula used for the conversion is:

$$C_m = \sum_{K=1 \text{ to } N} \cos[m*(k-0.5)*\prod/N]*E_k \quad (7)$$

The value of m ranges from 1, 2, 3……L (L is the number of cepstral coefficients), E_k is the energy obtained from previous step.

B. Delta and Delta-delta Coefficient Measurement

MFCC represents the shape of the vocal tract in the form of an envelope, but the speech signals varying highly with time are also subjected to information that is dynamic in nature. In order to enhance the accuracy of the system, the first and second derivatives of MFCC are calculated, that adds the trajectories of the MFCC coefficients obtained [11].

The first derivative is calculated from the thirteen MFCC coefficients. These coefficients are passed to the following equation and are called delta coefficients.

$$\text{delta}_t = \frac{\sum_{n=1}^{N} n(c_{t+N} - c_{t-N})}{2\sum_{n=1}^{N} n^2} \quad (8)$$

where delta_t is the delta coefficient, for frame t computed from c_{t+N} to c_{t-N} from MFCC coefficients (static coefficients). Standard value of N is taken 2. Similarly, the second derivatives are obtained from the thirteen delta coefficients being passed to the (8) and are called delta-delta coefficients.

The thirteen values were obtained for each feature of MFCC, delta and the delta-delta making a feature vector of 39 dimensions for one sample, as shown in the **fig. 9**. The speech database was prepared using the Microsoft Excel. The speech database obtained is shown in **fig. 10.**

V. Classification

The classification step involved the training and testing of the model to recognize gender was carried out in the tool named RapidMiner. This tool is developed by a company named "RapidMiner" and provides an integrated environment for the various tasks like data mining, business and predictive analytics. Rapid Miner allows the users to import any kind of database to examine and analyse it and is the only tool that is independent of any language limitation and is considered appropriate for advanced users as compared in the study in [14]. Rapid Miner version 7 was used in the research work.

In order to perform training and testing on the same dataset the cross validation technique was used and the results were obtained. Rapid Miner provides X-Validation operator that is a nested operator, that is, it further has two sub-processes namely, training sub-process and testing sub-process. The former sub-process trains the model which is them applied to the later sub-process [15]. Inside the X-Validation operator we used the stacking operator. This operator combines the different classifiers to give the final result. The two classifiers used were SVM and NN. The Naïve Baye's classifier was used as a learner model, means it obtained the prediction made by SVM and NN and combined those predictions to give the final decision resultant.

The stacked model was used to train the model and the same was applied to test on unseen data. The apply model operator is used for testing the dataset and the performance was measured using the performance operator [15].

The **fig. 11 & 12** below shows the classification model for both training and testing. The main reason for choosing support vector and neural network combination involves the robustness and accuracy of support vector machine and theoretical aspects of neural networks. SVM functions well on small as well as large datasets and are insensitive to the dimension of the feature vector [16]. Since the research work involves two class learning task. SVM finds the classification function that best distinguishes between the two classes.

Whereas Neural Network is the most researched among various machine learning algorithms and has also evolved a promising alternative to other classifiers. NN is a self-adaptive method that is data driven, that is, they are adjustable to the data without any specified function or model [16].

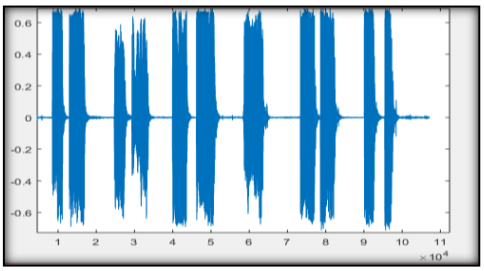


Fig.3(a). Original Signal

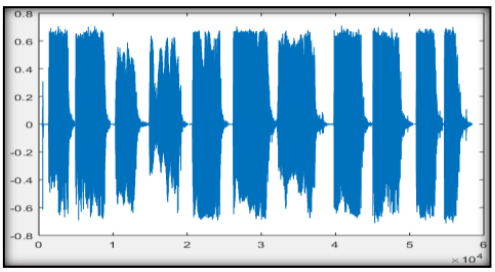


Fig.3(b). Signal after Pre-processing (right)

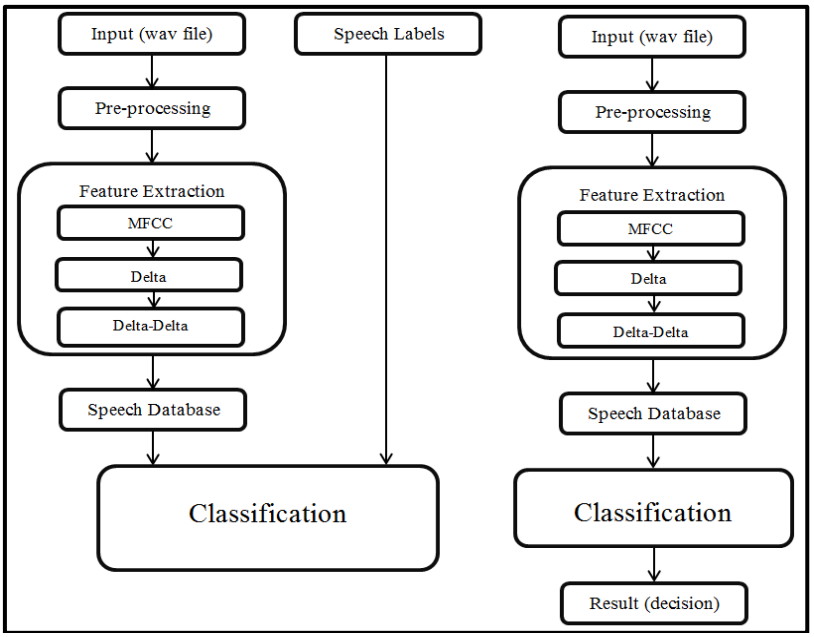


Fig.4. Proposed System (training and testing)

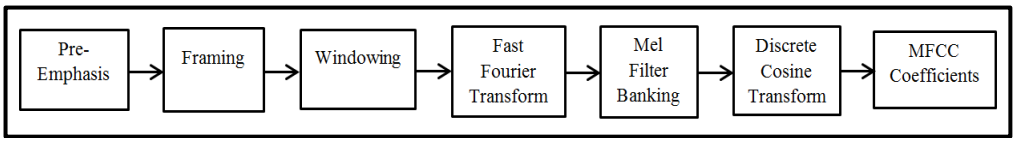


Fig.5. MFCC Steps

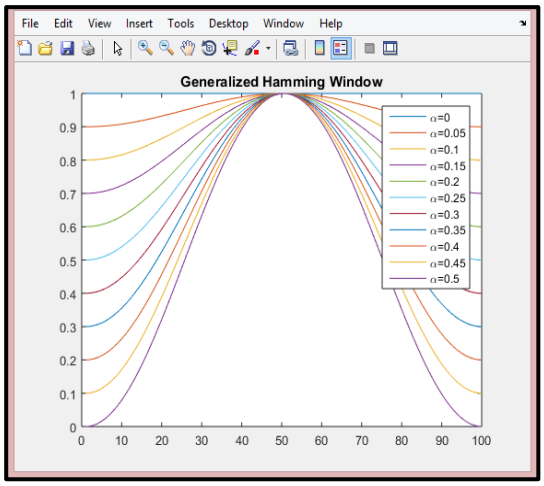


Fig.6. Generalized Hamming Window

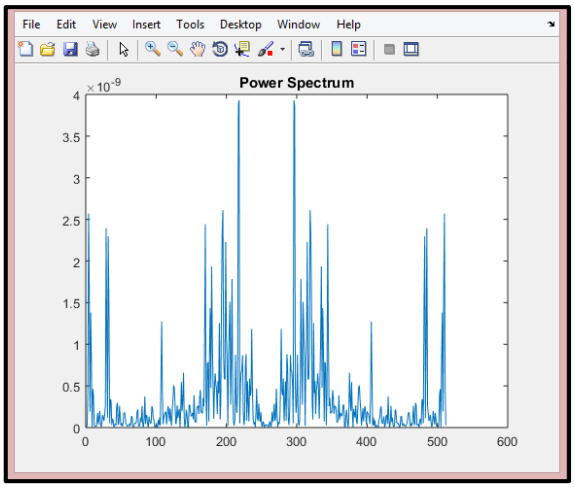


Fig.7. Power Spectrum

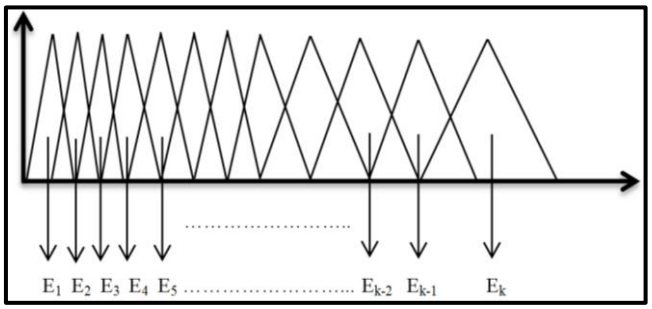
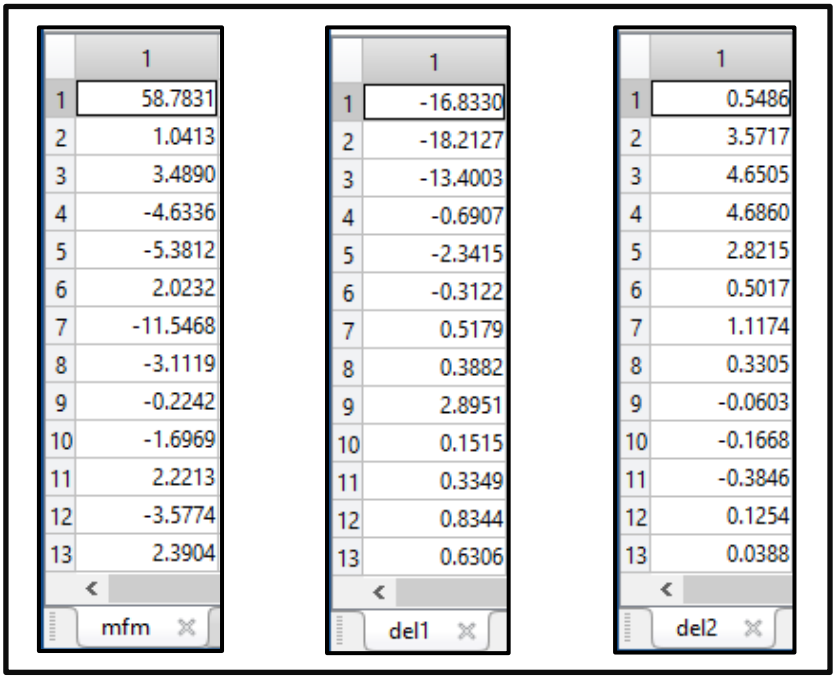


Fig.8. Mel Scale Filter Bank from [10]



	1
1	58.7831
2	1.0413
3	3.4890
4	-4.6336
5	-5.3812
6	2.0232
7	-11.5468
8	-3.1119
9	-0.2242
10	-1.6969
11	2.2213
12	-3.5774
13	2.3904

mfm

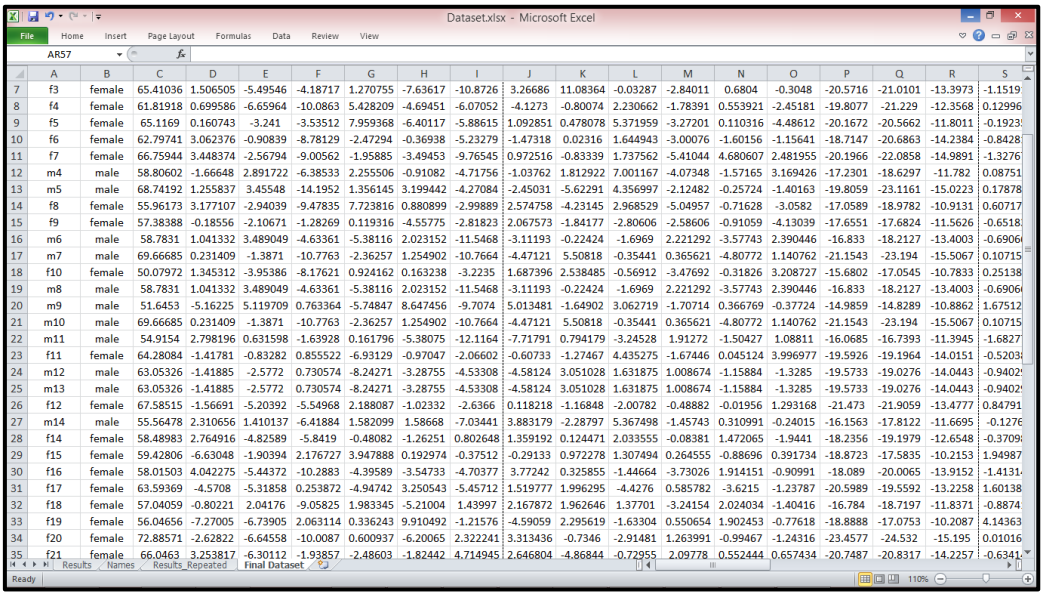
	1
1	-16.8330
2	-18.2127
3	-13.4003
4	-0.6907
5	-2.3415
6	-0.3122
7	0.5179
8	0.3882
9	2.8951
10	0.1515
11	0.3349
12	0.8344
13	0.6306

del1

	1
1	0.5486
2	3.5717
3	4.6505
4	4.6860
5	2.8215
6	0.5017
7	1.1174
8	0.3305
9	-0.0603
10	-0.1668
11	-0.3846
12	0.1254
13	0.0388

del2

Fig.9. MFCC, Delta and Delta-delta Coefficients



Dataset.xlsx - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
7	f3	female	65.41036	1.506505	-5.49546	-4.18717	1.270755	-7.63617	-10.8726	3.26686	11.08364	-0.03287	-2.84011	0.6804	-0.3048	-20.5716	-21.0101	-13.3973	-1.1519
8	f4	female	61.81918	0.699586	-6.65964	-10.0863	5.428209	-4.69451	-6.07052	-4.1273	-0.80074	2.230662	-1.78391	0.553921	-2.45181	-19.8077	-21.229	-12.3568	0.12996
9	f5	female	65.1169	0.160743	-3.241	-3.53512	7.959368	-6.40117	-5.88615	1.092851	0.478078	5.371959	-3.27201	0.110316	-4.48612	-20.1672	-20.5662	-11.8011	-0.1923
10	f6	female	62.79741	3.062376	-0.90839	-8.78129	-2.47294	-0.36938	-5.23279	-1.47318	0.02316	1.644943	-3.00076	-1.60156	-1.15641	-18.7147	-20.6863	-14.2384	-0.8428
11	f7	female	66.75944	3.448374	-2.56794	-9.00562	-1.95885	-3.49453	-9.76545	0.972516	-0.83339	1.737562	-5.41044	4.680607	2.481955	-20.1966	-22.0858	-14.9891	-1.3276
12	m4	male	58.80602	-1.66648	2.891722	-6.38533	2.255506	-0.91082	-4.71756	-1.03762	1.812922	7.001167	-4.07348	-1.57165	3.169426	-17.2301	-18.6297	-11.782	0.08751
13	m5	male	68.74192	1.255837	3.45548	-14.1952	1.356145	3.199442	-4.27084	-2.45031	-5.62291	4.356997	-2.12482	-0.25724	-1.40163	-19.8059	-23.1161	-15.0223	0.17878
14	f8	female	55.96173	3.177107	-2.94039	-9.47835	7.723816	0.880899	-2.99889	2.574758	-4.23145	2.968529	-5.04957	-0.71628	-3.0582	-17.0589	-18.9782	-10.9131	0.60717
15	f9	female	57.38388	-0.18556	-2.10671	-1.28269	0.119316	-4.55775	-2.81823	2.067573	-1.84177	-2.80606	-2.58606	-0.91059	-4.13039	-17.6551	-17.6824	-11.5626	-0.6518
16	m6	male	58.7831	1.041332	3.489049	-4.63361	-5.38116	2.023152	-11.5468	-3.11193	-0.22424	-1.6969	2.221292	-3.57743	2.390446	-16.833	-18.2127	-13.4003	-0.6906
17	m7	male	69.66685	0.231409	-1.3871	-10.7763	-2.36257	1.254902	-10.7664	-4.47121	5.50818	-0.35441	0.365621	-4.80772	1.140762	-21.1543	-23.194	-15.5067	0.10715
18	f10	female	50.07972	1.345312	-3.95386	-8.17621	0.924162	0.163238	-3.2235	1.687396	2.538485	-0.56912	-3.47692	-0.31826	3.208727	-15.6802	-17.0545	-10.7833	0.25138
19	m8	male	58.7831	1.041332	3.489049	-4.63361	-5.38116	2.023152	-11.5468	-3.11193	-0.22424	-1.6969	2.221292	-3.57743	2.390446	-16.833	-18.2127	-13.4003	-0.6906
20	m9	male	51.6453	-5.16225	5.119709	0.763364	-5.74847	8.647456	-9.7074	5.013481	-1.64902	3.062719	-1.70714	0.366769	-0.37724	-14.9859	-14.8289	-10.8862	1.67512
21	m10	male	69.66685	0.231409	-1.3871	-10.7763	-2.36257	1.254902	-10.7664	-4.47121	5.50818	-0.35441	0.365621	-4.80772	1.140762	-21.1543	-23.194	-15.5067	0.10715
22	m11	male	54.9154	2.798196	0.631598	-1.63928	0.161796	-5.38075	-12.1164	-7.71791	0.794179	-3.24528	1.91272	-1.50427	1.08811	-16.0685	-16.7393	-11.3945	-1.6827
23	f11	female	64.28084	-1.41781	-0.83282	0.855522	-6.93129	-0.97047	-2.06602	-0.60733	-1.27467	4.435275	-1.67446	0.045124	3.996977	-19.5926	-19.1964	-14.0151	-0.5203
24	m12	male	63.05326	-1.41885	-2.5772	0.730574	-8.24271	-3.28755	-4.53308	-4.58124	3.051028	1.631875	1.008674	-1.15884	-1.3285	-19.5733	-19.0276	-14.0443	-0.9402
25	m13	male	63.05326	-1.41885	-2.5772	0.730574	-8.24271	-3.28755	-4.53308	-4.58124	3.051028	1.631875	1.008674	-1.15884	-1.3285	-19.5733	-19.0276	-14.0443	-0.9402
26	f12	female	67.58515	-1.56691	-5.20392	-5.54968	2.188087	-1.02332	-2.6366	0.118218	-1.16848	-2.00782	-0.48882	-0.01956	1.293168	-21.473	-21.9059	-13.4777	0.84791
27	m14	male	55.56478	2.310656	1.410137	-6.41884	1.582099	1.58668	-7.03441	3.883179	-2.28797	5.367498	-1.45743	0.310991	-0.24015	-16.1563	-17.8122	-11.6695	-0.1276
28	f14	female	58.48983	2.764916	-4.82589	-5.8419	-0.48082	-1.26251	0.802648	1.359192	0.124471	2.033555	-0.08381	1.472065	-1.9441	-18.2356	-19.1979	-12.6548	-0.3709
29	f15	female	59.42806	-6.63048	-1.90394	2.176727	3.947888	0.192974	-0.37512	-0.29133	0.972278	1.307494	0.264555	-0.88696	0.391734	-18.8723	-17.5835	-10.2153	1.94987
30	f16	female	58.01503	4.042275	-5.44372	-10.2883	-4.39589	-3.54733	-4.70377	3.77242	0.325855	-1.44664	-3.73026	1.914151	-0.90991	-18.089	-20.0065	-13.9152	-1.4131
31	f17	female	63.59369	-4.5708	-5.31858	0.253872	-4.94742	3.250543	-5.45712	1.519777	1.996295	-4.4276	0.585782	-3.6215	-1.23787	-20.5989	-19.5592	-13.2258	1.60138
32	f18	female	57.04059	-0.80221	2.04176	-9.05825	1.983345	-5.21004	1.43997	2.167872	1.962646	1.37701	-3.24154	2.024034	-1.40416	-16.784	-18.7197	-11.8371	-0.8874
33	f19	female	56.04656	-7.27005	-6.73905	2.063114	0.336243	9.910492	-1.21576	-4.59059	2.295619	-1.63304	0.550654	1.902453	-0.77618	-18.8888	-17.0753	-10.2087	4.14363
34	f20	female	72.88571	-2.62822	-6.64558	-10.0087	0.600937	-6.20065	2.322241	3.313436	-0.7346	-2.91481	1.263991	-0.99467	-1.24316	-23.4577	-24.532	-15.195	0.01016
35	f21	female	66.0463	3.253817	-6.30112	-1.93857	-2.48603	-1.82442	4.714945	2.646804	-4.86844	-0.72955	2.09778	0.552444	0.657434	-20.7487	-20.8317	-14.2257	-0.6341

Results | Names | Results_Repeated | Final Dataset

Ready

Fig.10. Speech Database

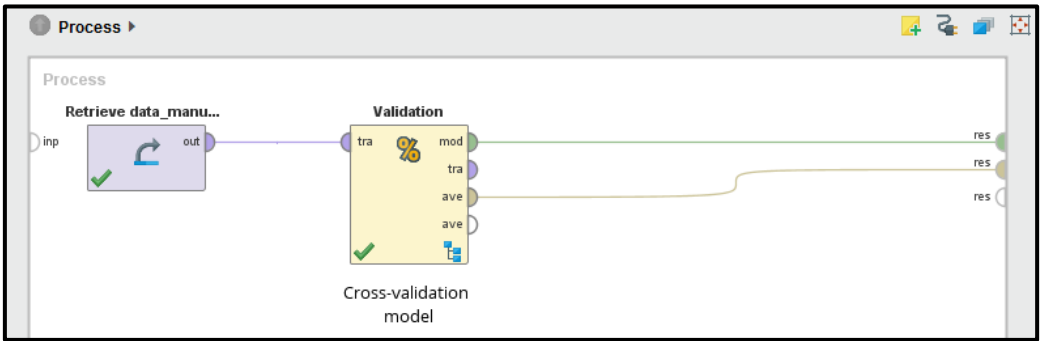


Fig.11. Classification Using Cross Validation

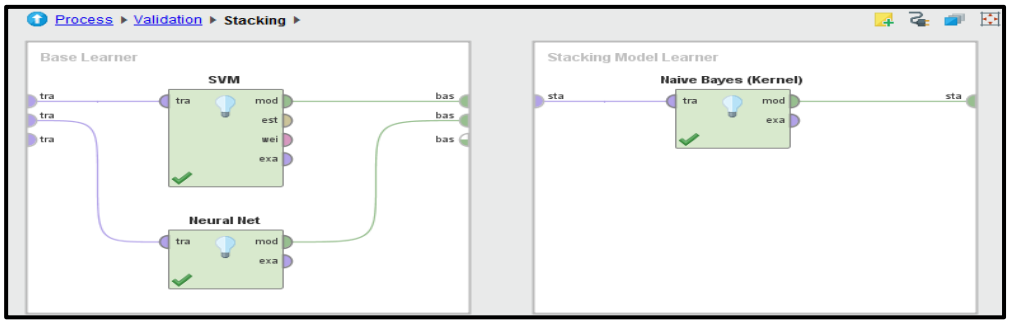


Fig.12. Stacking Model (Training and Testing)

VI. Result and Conclusion

The feature extraction process was carried out using MATLAB code and a total of 39 parameters were extracted from three features. From the extracted features the speech database (excel file) was prepared, which contained the columns named, the unique id for each sample, the thirty-nine parameters and the gender column. The further process was carried out on Rapid Miner Studio where the speech data base was loaded and connected to the classification model prepared.

A. Considering the first Mel coefficient

The accuracy obtained was 93.48% with the recall of 92.33% for female and 95.22% for male. The precision obtained was 96.00% for female and 90.48% for male.

Table 1. Confusion Matrix obtained after considering MFCC1 value with an overall accuracy of 93.48%

	true female	true male	precision
predicted female	24	1	96.00%
predicted male	2	19	90.48%
recall	92.33%	95.22%	

The recall in **Table 1** depicts recognition rate, it says that 24 out of 26 females were predicted correctly and the two were predicted as males (92.33%) whereas 19 out of 20 males were predicted correctly and one was predicted as female (95.22%). The precision depicts that out of all the predicted females, that is, 25, 24 females were correctly predicted (96.00%) whereas out of all those predicted to be males (that is 21), only 19 were correctly predicted (90.48%). The accuracy of the model was calculated as percentage of the sum of correct predictions made in both cases divided by total number of cases, that is, (24+19) ÷ (46) which comes out to be 93.48%.

B. Ignoring the first Mel coefficient

The accuracy obtained was 91.3% and a recall of 88.46% for female and 95% for male was measured. The precision obtained was 95.83% for female and 86.36% for female.

Table 2. Confusion Matrix obtained after ignoring MFCC1 value with an overall accuracy of 91.3%

	true female	true male	precision
predicted female	23	1	95.36%
predicted male	3	19	86.36%
recall	88.46%	95.00%	

The recognition rate in this case (Table 2) is depicted as, 24 out of 26 females were correctly recognised whereas 3 were recognised as males. And 19 out of 20 males were recognised as males and only one was recognised incorrectly as female, giving a recall of 88.46% and 95% respectively. The precision obtained was out of all predicted females (that is, 24) 23 were correctly predicted and out of all predicted males (that is, 22).

VII. Conclusion

The system was said to perform well when the MFCC 1 values were taken into consideration than when they were ignored giving an improvement of 2.18 in overall performance. The recognition rate was also higher when the value was considered (that is, 92.33 for females and 95.22% for male) as compared to when the value was ignored (that is, 88.46% for females and 95.00% for males). The final conclusion derived is that since the system gives considerable good performance the features have been correctly extracted and the MFCC 1 value for each sample in the dataset tends to increase the system performance and the recognition rate. So it was better to consider the first value in our case.

Acknowledgment

Sincere thanks to Mr Gaurav Aggarwal, Assistant Professor, The NorthCap University for his valuable comments, advice, continuous support and encouragement during the tenure of this work. Also the authors would like to thanks all the people who gave their samples for the completion of this project work.

REFERENCES

[1] Parwinder Pal Singh and Pushpa Rani, An Approach to Extract Feature using MFCC, IOSR Journal of Engineering, Vol. 04, August. 2014.

[2] D.Shakina Deiv, Gaurav, Mahua Bhattacharya, Automatic Gender Identification for Hindi Speech Recognition, International Journal of Computer Applications (0975 – 8887) Volume 31– No.5, October 2011.

[3] M.A Anusuya and S.K. Katti, Front End Analysis of Speech Recognition- A review, Int J Speech Technol, Springer, DOI 10.1007/s10772-010-9088-7.

[4] M.Li, K. Han and S. Narayanan, automatic Speaker Age and Gender Recognition Using Acoustic and Prosodic Level Information Fusion, Computer Speech and Language, Jan 2013.

[5] H. Kim, K. Bae, H. Yoon, Age and Gender Classification for a Home-Robot Service, Proc. 16th IEEE International Symposium on Robot and Human Interactive Communication.

[6] Qiyue Liu, Mingqiu Yao, Han Xu, Fang Wang, Different Feature Parameters in Speaker Recognition, Journal of Signal and Information Processing 2013.

[7] Parwinder Pal Singh and Pushpa Rani, An Approach to Extract Feature using MFC, IOSR Journal of Engineering (IOSRJEN) Vol. 4, Issue 08(August 2014).

[8] Jamil Ahma, Mustansar, Fiaz, Soon-il Kwon, Maleerat Soanil, Bay Vo and Sung Wook Baik, Gender Identification using the MFCC for Telephone Applications- A Comparitive Study, International Journal of Computer Science and Electronics Engineering (IJCSEE) 2015.

[9] Jerzy SAS, Aleksaner SAS, Gender Recognition Using Neural Networks and ASR Techniques, Journal of Medical Informatics and Technology 2013.

[10] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques, Journal Of Computing, Volume 2, Issue 3, March 2010.

[11] Mfcc Tutorial, practical cryptography.

[12] M.A. Anusuya, Frontend Analysis of Speech Recognition-a review, Int J Speech Technol, Springer.

[13] Kalpana Rangra and Dr. K. L. Bansal, Comparative Study of Data Mining Tools, International Journal of Advanced Research in Computer Science and Software Engineering, June 2014.

[14] Rapid Miner Documentation, Operator reference Manual.pdf.

[15] Nidhi H. Ruparel and Nitin M. Shahane, Learning from Small Data Set to Build Classification Model: A Survey, International Journal of Computer Applications, International Conference on Recent Trends in engineering & Technology-2013(ICRTET'2013)

[16] Han and Kamber, Data Mining Concepts and Techniques, Second Edition.

Authors' Profiles

Anjali Pahwa was born in India in 1992. She received her bachelor's degree in Information and Technology from Maharishi Dayanand University, India class of June 2014 and master degree in Computer Science and Engineering from The NorthCap University, India class of June 2016.

Her major field of interest include Speech Signal Processing, Java, Web Designing and Pattern Recognition.

Gaurav Aggarwal was born in India. He received his bachelor's as well as master's degree from Kurukshetra University, India. He prepares his PhD in Speech Signal Processing from The NorthCap University, India and is an expert in the field of Computer Graphics and Java. He teaches as an Assistant Professor in the same University.

He attended the IT&T International Conference, 2012, held at Cork, Ireland, and presented two research papers and one poster, doing NCU proud.

How to cite this paper: Anjali Pahwa, Gaurav Aggarwal,"Speech Feature Extraction for Gender Recognition", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.8, No.9, pp.17-25, 2016.DOI: 10.5815/ijigsp.2016.09.03

