Effect of Gender on Improving Speech Recognition System

Amer Sallam
Faculty of Engineering & Information Technology,
TAIZ University, Taiz, Yeman

Sreedhar Bhukya
Speech and Vision Laboratory, IIIT- Hyderabad – India

ABSTRACT

Speech is the output of a time varying excitation excited by a time varying system. It generates pulses with fundamental frequency F0. This time varying impulse trained as one of the features, characterized by fundamental frequencyF0and its formant frequencies. These features vary from one speaker to another speaker and from gender to gender also. In this paper the effect of gender on improving speech recognition is considered. Variation in F0 and formant frequencies is the main features that characterize variation in a speaker. The variation becomes very less within speaker, medium within the same gender and very high among different genders. This variation in information can be exploited to recognize gender type and to improve performance of speech recognition system through modeling separate models based on gender type information. Five sentences are selected for training. Each of the sentences are spoken and recorded by 20 female's speakers and 20 male speakers. The speech corpus wills be preprocessed to identify the voiced and unvoiced region. The voiced region is the only region which carries information about F0. From each voiced segment, F0and the first three formant frequencies and also MFCC features are computed. Each forms the feature space labeled with the speaker identification: i.e., male or female. This information misused to parameterize the model for male and female. K-means algorithm is used during training as well as testing. Testing is conducted in two ways: speaker dependent testing and speaker independent testing. SPHINX-III software by Carnegie Mellon University has been used to measure the accuracy of speech recognition of data taking into account the case of gender separation which has been used in this research.

Keywords

Speech recognition (SR), Linear Prediction Coding (LPC), Accent, Speakers.

1. INTRODUCTION

As speech recognition is a Complex task it is still difficult to find the complete solution, because every human being has his/her own different characteristics of voice, this in itself has become one of the main problems in the field of speech recognition [9-16].

It is worth pointing out the fact that this research investigates an approach for identifying genders from spoken data and builds separated models for gender that can enhance the performance of speech recognition by reducing the search space in lexicon. Studies in gender classification using voice shall give insights to how humans process voice information. What may be important is that gender information is conveyed by the F0, type of sound, and the size of vocal tract. It can be assumed that modeling the vocal tract using Linear Prediction Coding (LPC) and Cestrum in formation [1,2]. Furthermore, small errors in gender classification can be allowed as sometimes it is even hard for a person to identify the gender of a speaker. Studies by Susan Schötz at Lund University [3] have showed that a feature based system with a trained decision tree can successfully classify male and female voices automatically. However, her studies have been most concentrated on age separation.

In the context of speech recognition, gender separation can improve the performance of speech recognition by limiting the search space to speakers from the same gender. The automatic speech recognition is aimed to extract the sequence of spoken words from a recorded speech signal and so it does not include the task of speech understanding, which can be seen as an even more elaborated problem. In the gender classification accuracy, it has been noticed that perhaps the most important variability across speakers besides gender is a role played by accent. Therefore, it is probably that any recognition system attempting to be robust to a wide variety of speakers and languages should make some effort to account for different accents that the system might encounter.

In this paper, two main approaches have been applied which can be used for gender separation. One approach is to use gender dependent features, such as the pitch and formants. The other approach is to use a general pattern recognition based on general speech features such as the Me1Frequency Cepstral Coefficients (MFCCs) [4]. The pitch information was used in [5] for the problem of gender separation. However, fundamental frequency and formant frequencies estimation relies considera -bly on the speech quality. Although the quality of speech used in this study was not free from noise, tried to improve the gender model by using K- means algorithm to get high accuracy. Current state of the art of speech recogn -ition systems (HMM based or Hybrid model) acoustic model performance is very low and without language model, current speech recognition system should not be applied to any purpose. Improvement of acoustic model has greater impact on speech recognition system. Moreover, designing gender dependent speech recognition system model improves the performance of speaker independent and large vocabulary size speech recognition system. Hence, this paper is intended to explore and test the performance of gender separation system. As mentioned above, gender recognition system will make use of the basic features like fundamental frequency and the first three formant frequencies and also MFCCs.

2. IMPLEMENTATION

Gender separation is the process of separating the speaker's gender type directly from the acoustic information. This is possible using gender invariant feature separation and learning from variations within gender. Furthermore, Speech is produced as a result of convolution of excitation of the vocal cord with the vocal tract system coupled with the nasal tract. It is decided to make a practical study of effect of gender on improving speech recognition system, trying to find out the most correct results by implementing different methods, techniques and procedures. This paper depends on the use of several ways of techniques to extract the features that can be more useful to identify gender and build separate models for gender that can enhance the performance of speech recognition through those models by limiting the search space to speakers from the same gender. This practical study was carried out though several stages that will be described as follows.

Stage 1: Data Collection

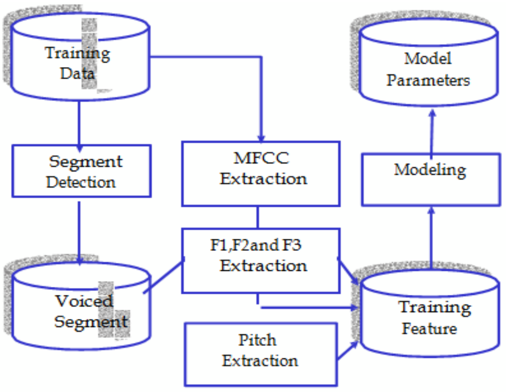
Training Data:
20 female and 20 male speaker subjects are selected to record 5 properly selected sets of sentences. The selected recorded sentences for the training purpose are: welcome, where Mike is, I believe you are fine, Have fun with him, Thanks to God.

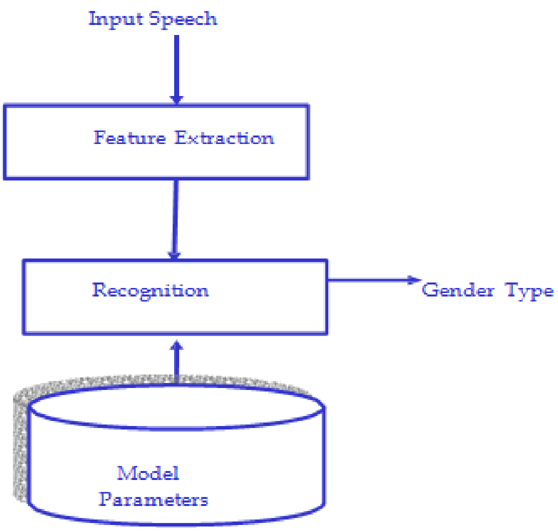
Testing Data:
For the testing purpose, 10 females and 10 males have been selected to speak one sentence and each has been recorded. In fact, every speaker as given a sentence different from other speaker. It has to be observed that the group of testing is actually independent from that of the training data.

Stage 2: Feature Extraction

In this stage, we are going to build the process that can identify the gender type on the basis of individual information included in speech waves through extracting the features, viz, fundamental frequency F0, formant frequencies F1, F2, and F3 and MFCCs from the collected training data making separate models depending on the type of features and optimal parameters for each gender. These formant frequencies and the fundamental frequency show high variation from one speaker to another. The variation becomes higher when the comparison is among speakers of different gender.

This process is represented in Figure 2.1. Below technically, when the voice characteristics of utterance are checked, there will be a wide range of probabilities that exist for parametrically representing the speech signal for the gender separation task such as Linear Prediction Coding (LPC), Mel-Frequency Cestrum Coefficients MFCCs).





Stage3: Feature Matching and Decision Making

In this stage, extracted the same feature type for testing data as done during training stage. Then applied the concept to pattern recognition to classify objects of interest in to one of the desired gender types. The objects of interest are called sequences of vectors that are extracted from an inputs speech. The type of gender here refers to individual gender. Since the classification procedure in our case is applied on extracted features, it can be also referred to as feature matching.

The K-means approach [6, 7, 8] is used because of its flexibility and ability to give high recognition accuracy. K-means can be simply defined as a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a code word. The collection of all code words is called a code book. The gender separation System will compare the code books of the tested speaker with the code books of the trained speaker. In other words, during recognition, among the models, the best model that maximizes the joint probability of the given observation will be selected as recognized model. The best matching result will be the desired gender type and this can be verified as the decision making logic taking into account. That the system decision may misclassify the gender type unless the two following criteria are satisfied.

□□The system has found the lowest Euclidean Distance between the code book tested and the various trained codebooks.

□□The distance calculated falls below are defined threshold of acceptance.

Finally, the gender type which is modeled by their cognized model will be given as an output of our system.

3. DATA ANALYSIS & OBSERVATION

In this section, the data analysis of the work is discussed. Figure2.3.shows feature extraction (F0, F1, F2 and F3) in which the utterance welcome of both genders is analyzed showing that female on the right and male on the left. It shows that female speakers have higher pitch and formant frequency than male. The other utterances can be considered by means of analogy. Regarding the techniques of feature extraction, each sub figure has its description as below in Figure2.3.

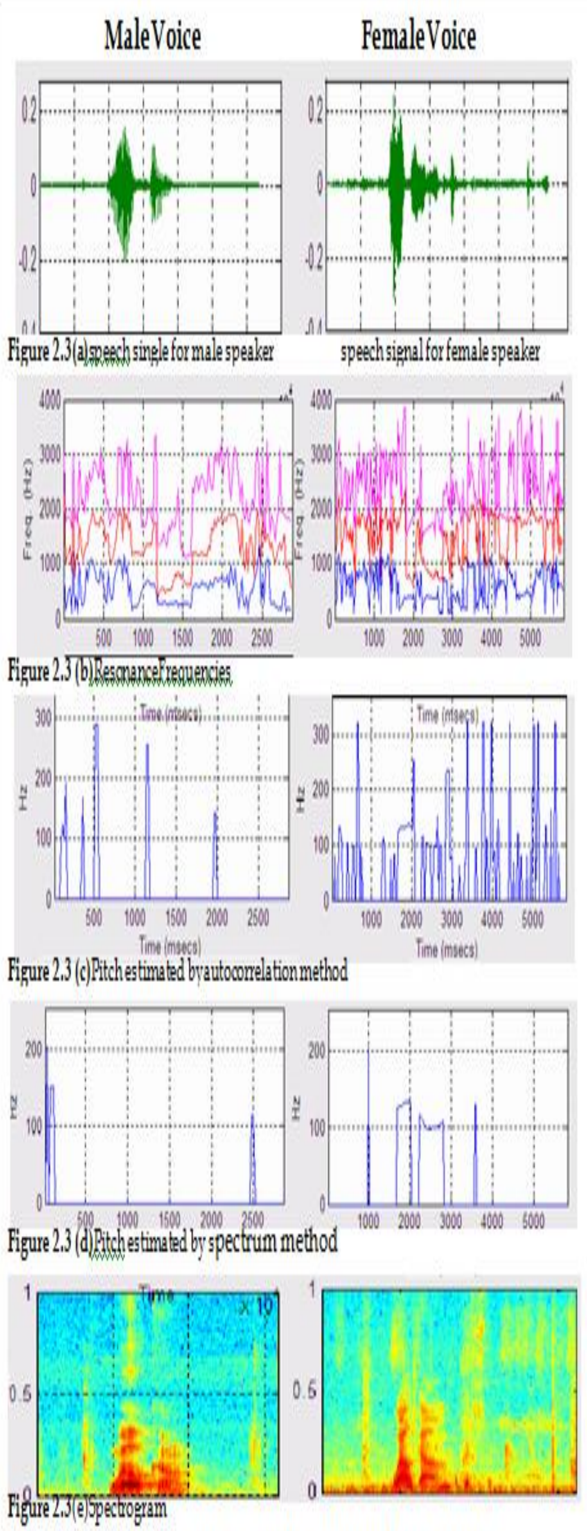


Figure 2.3(a)speech single for male speaker speech signal for female speaker

Figure 2.3 (b)ResonanceFrequencies

Figure 2.3 (c)Pitch estimated by autocorrelation method

Figure 2.3 (d)Pitch estimated by spectrum method

Figure 2.3(e)Spectrogram

Note: X axe represents the time in msec and Y axe represents Frequency in Hz

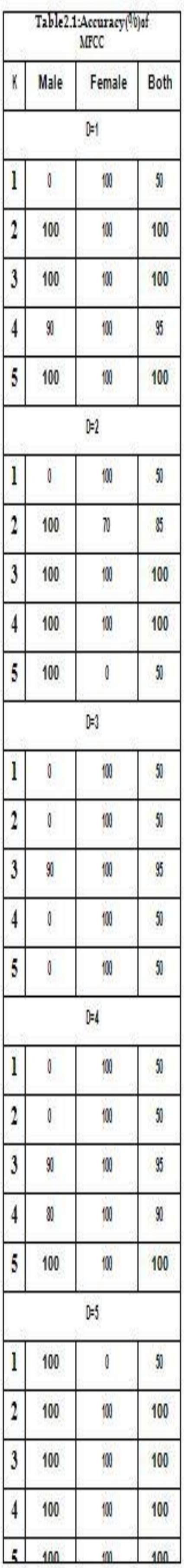


Table2.1:Accuracy(%)of MFCC

K	Male	Female	Both
D=1			
1	0	100	50
2	100	100	100
3	100	100	100
4	90	100	95
5	100	100	100
D=2			
1	0	100	50
2	100	70	85
3	100	100	100
4	100	100	100
5	100	0	50
D=3			
1	0	100	50
2	0	100	50
3	90	100	95
4	0	100	50
5	0	100	50
D=4			
1	0	100	50
2	0	100	50
3	90	100	95
4	80	100	90
5	100	100	100
D=5			
1	100	0	50
2	100	100	100
3	100	100	100
4	100	100	100
5	100	100	100

Figure2.4: Recognition Accuracy (RA)of MFCC Vs. No.of clustering K and varying of Dimension (D) between1to5

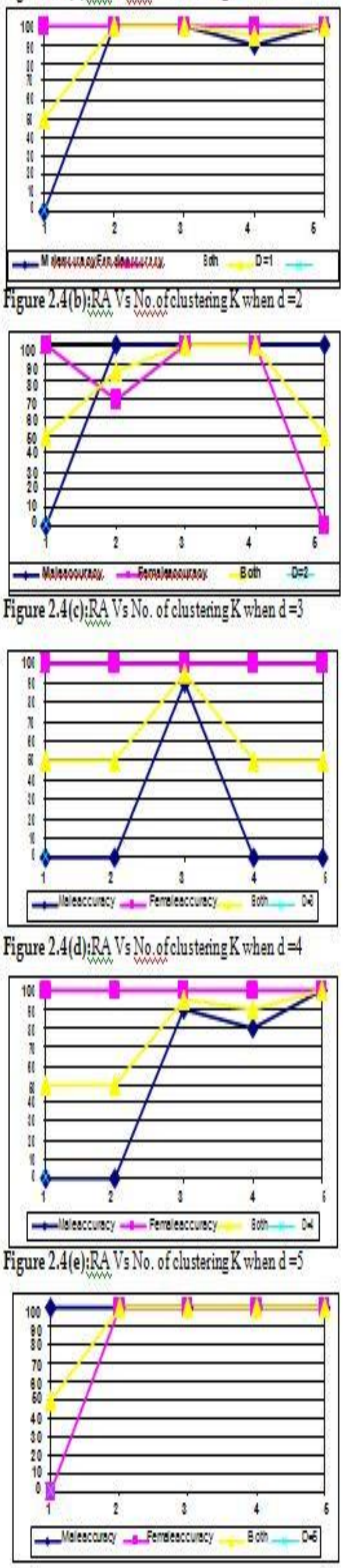


Figure 2.4(a):RA V sNo. of clustering K when d =1

Figure 2.4(b):RA Vs No.of clustering K when d =2

Figure 2.4(c):RA Vs No. of clustering K when d =3

Figure 2.4(d):RA Vs No.of clustering K when d =4

Figure 2.4(e):RA Vs No. of clustering K when d =5

.

Table2.2:Accuracy(%)of AutocorrelationMethod			
K	Male	Female	Both
D=1			
1	100	20	60
2	90	100	90
3	60	100	80
4	80	80	80
5	10	100	55
D=2			
1	100	0	50
2	100	100	100
3	90	100	95
4	90	100	95
5	80	70	75
D=3			
1	100	10	55
2	100	100	100
3	90	100	95
4	90	100	95
5	90	100	95
D=4			
1	100	30	65
2	100	100	100
3	90	100	95
4	90	100	95
5	90	70	80
D=5			
1	100	0	50
2	100	100	100
3	100	90	95
4	90	100	95
5	90	80	85

Figure2.5: Recognition Accuracy (RA) of Auto correlation Vs. No.of clustering Kand varying of Dimension (D) between1to5

Figure 2.5(a):RA VsNo. of clustering K when d =1

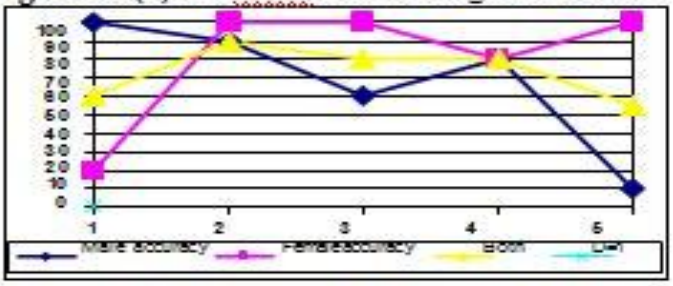


Figure 2.5(b):RAVsNo.of clustering K when d =2

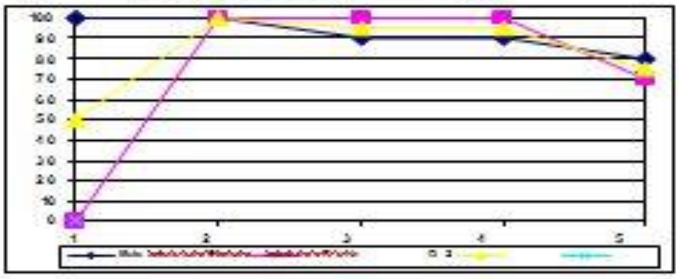


Figure 2.5(c):RA VsNo. of clustering K when d =3

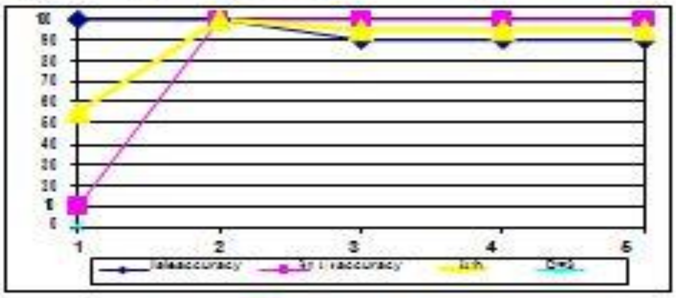


Figure 2.5(d):RAVsNo. ofclusteringKwhen d =4

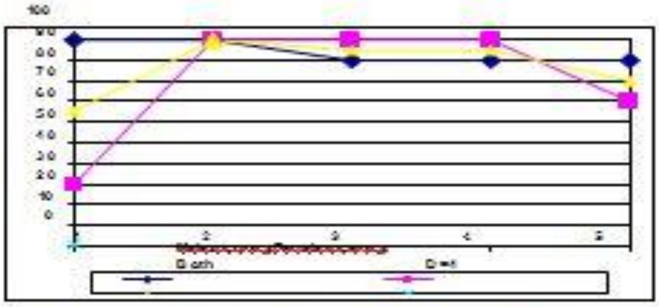
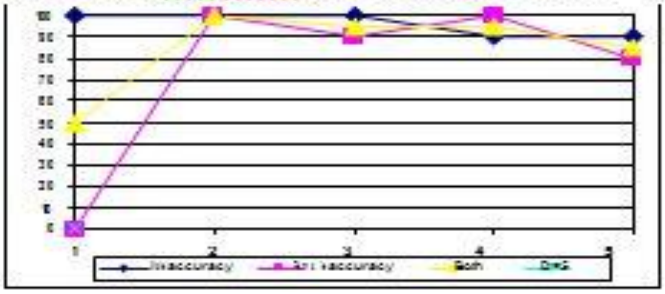


Figure 2.5(e):RA VsNo. of clustering K when d =5



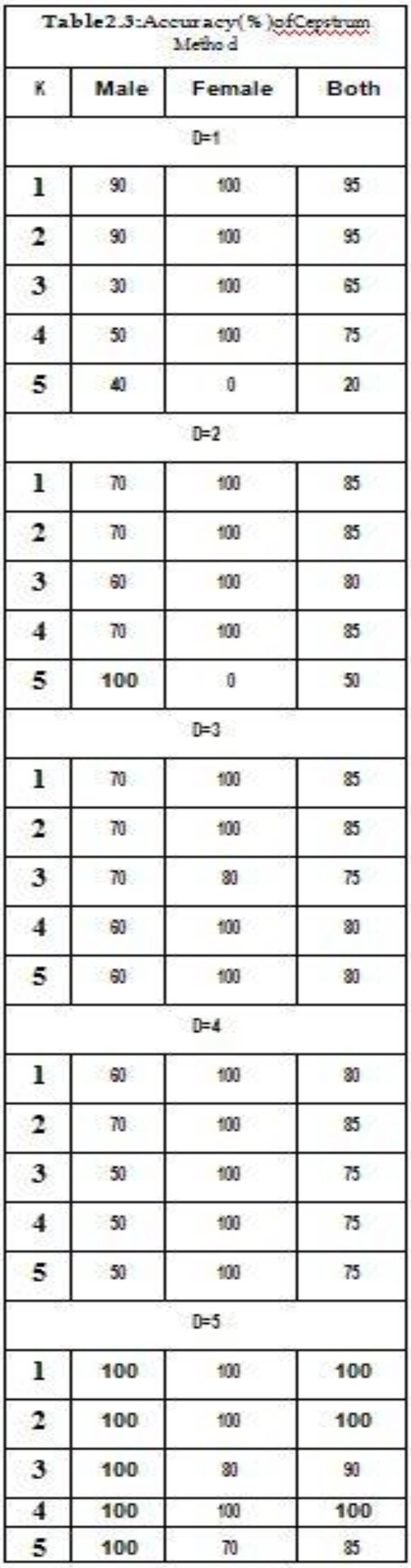


Table2.3:Accuracy(%)ofCepstrum Metho d

K	Male	Female	Both
D=1			
1	90	100	95
2	90	100	95
3	30	100	65
4	50	100	75
5	40	0	20
D=2			
1	70	100	85
2	70	100	85
3	60	100	80
4	70	100	85
5	100	0	50
D=3			
1	70	100	85
2	70	100	85
3	70	80	75
4	60	100	80
5	60	100	80
D=4			
1	60	100	80
2	70	100	85
3	50	100	75
4	50	100	75
5	50	100	75
D=5			
1	100	100	100
2	100	100	100
3	100	80	90
4	100	100	100
5	100	70	85

Figure2.6:RecognitionAccuracy(RA)of Cepstrum Vs. No.ofclusteringKandvaryingofDimension(D)between 1to5

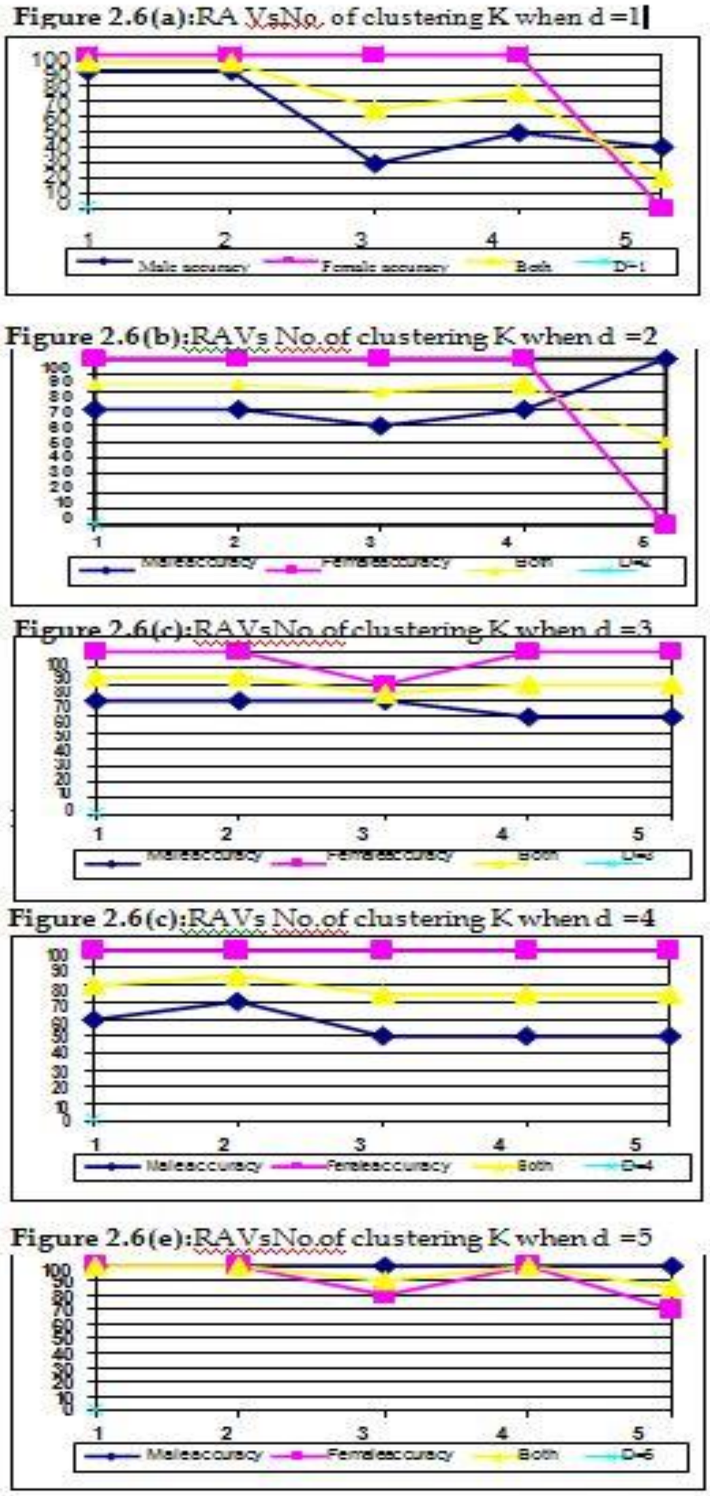


Figure 2.6(a):RA VsNo. of clustering K when d =1

Figure 2.6(b):RAVs No.of clustering K when d =2

Figure 2.6(c):RAVsNo.of clustering K when d =3

Figure 2.6(c):RAVs No.of clustering K when d =4

Figure 2.6(e):RAVsNo.of clustering K when d =5

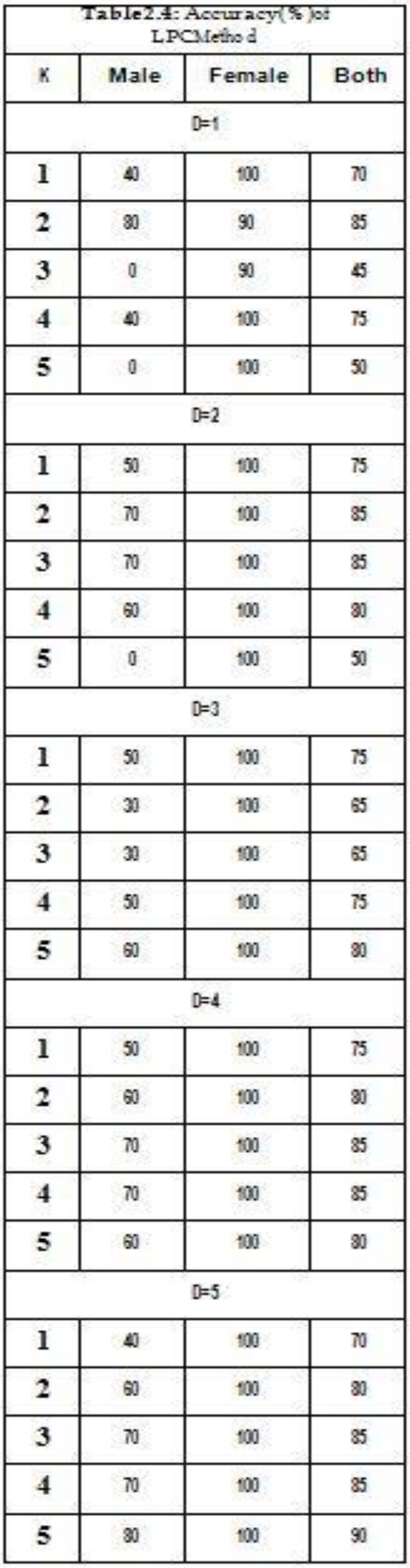


Table2.4: Accuracy(%)of LPCMetho d

K	Male	Female	Both
D=1			
1	40	100	70
2	80	90	85
3	0	90	45
4	40	100	75
5	0	100	50
D=2			
1	50	100	75
2	70	100	85
3	70	100	85
4	60	100	80
5	0	100	50
D=3			
1	50	100	75
2	30	100	65
3	30	100	65
4	50	100	75
5	60	100	80
D=4			
1	50	100	75
2	60	100	80
3	70	100	85
4	70	100	85
5	60	100	80
D=5			
1	40	100	70
2	60	100	80
3	70	100	85
4	70	100	85
5	80	100	90

Figure2.7: RecognitionAccuracy(RA)ofLPCVs.No.of clusteringKandvaryingofDimension(D)between1to5

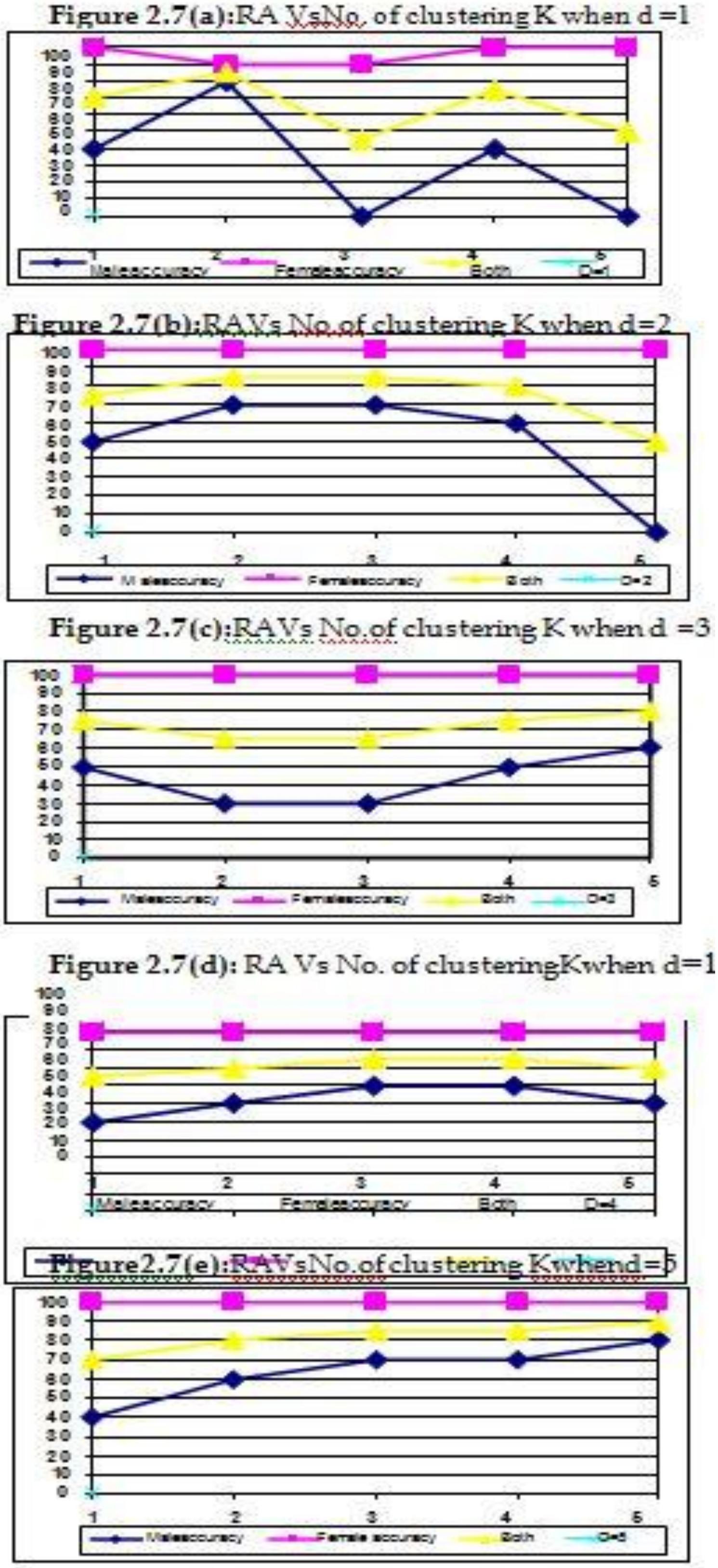


Figure 2.7(a):RA VsNo. of clustering K when d =1

Figure 2.7(b):RAVs No.of clustering K when d=2

Figure 2.7(c):RAVs No.of clustering K when d =3

Figure 2.7(d): RA Vs No. of clusteringKwhen d=1

Figure2.7(e):RAVsNo.of clustering Kwhend=5

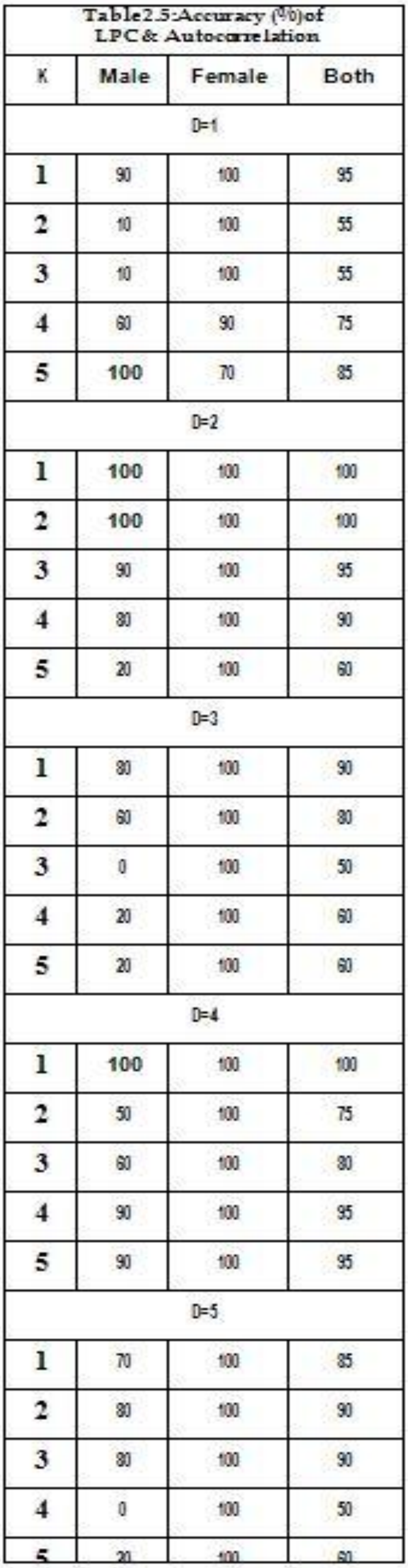


Table2.5:Accuracy (%)of LPC & Autocorrelation

K	Male	Female	Both
D=1			
1	90	100	95
2	10	100	55
3	10	100	55
4	60	90	75
5	100	70	85
D=2			
1	100	100	100
2	100	100	100
3	90	100	95
4	80	100	90
5	20	100	60
D=3			
1	80	100	90
2	60	100	80
3	0	100	50
4	20	100	60
5	20	100	60
D=4			
1	100	100	100
2	50	100	75
3	60	100	80
4	90	100	95
5	90	100	95
D=5			
1	70	100	85
2	80	100	90
3	80	100	90
4	0	100	50
5	20	100	60

Figure2.8: Recognition Accuracy(RA)of LPC & Autocorrelation Vs. No.of clustering Kand varying of Dimension (D) between1to5

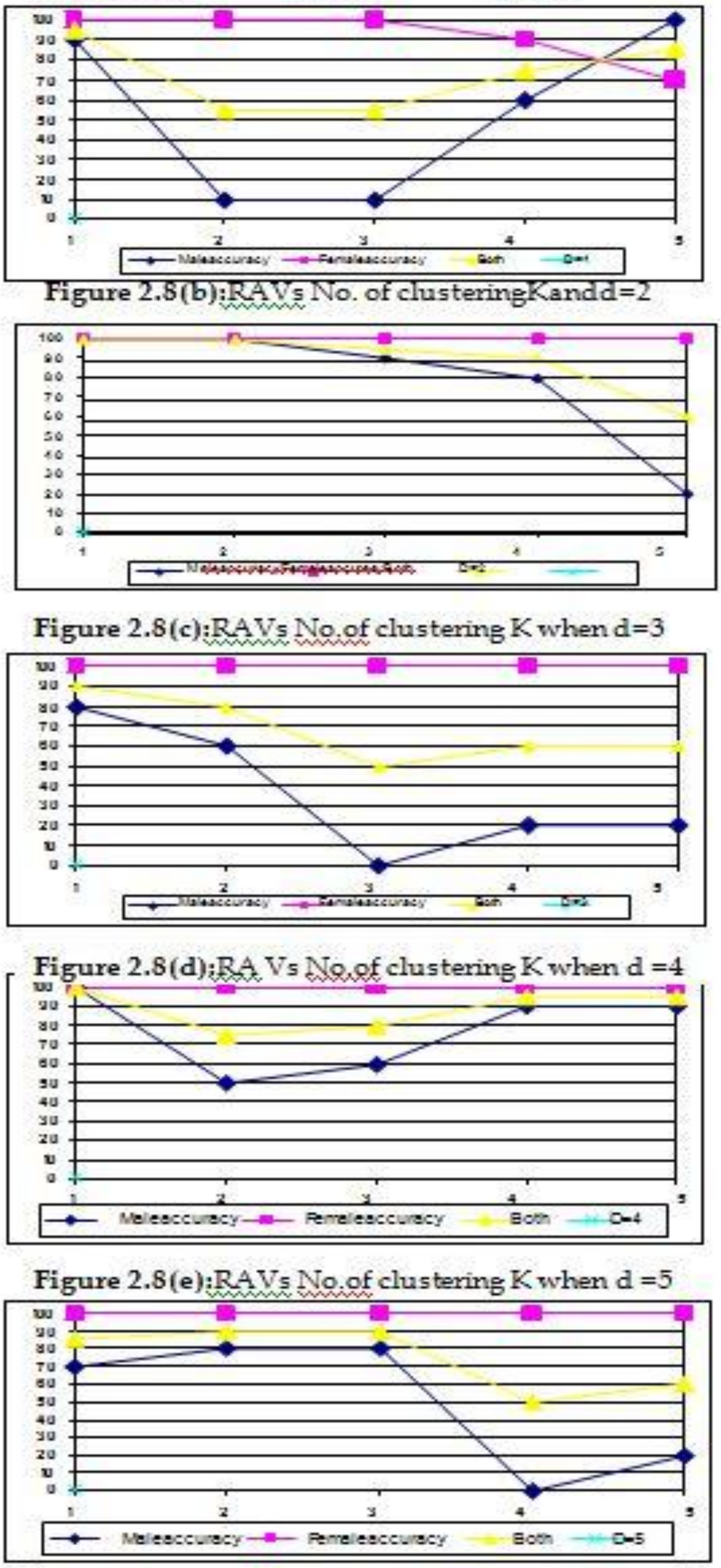


Figure2.8(a): RA Vs No.of clustering Kand d=1

Figure 2.8(b):RAVs No. of clusteringKandd=2

Figure 2.8(c):RAVs No.of clustering K when d=3

Figure 2.8(d):RA Vs No.of clustering K when d =4

Figure 2.8(e):RAVs No.of clustering K when d =5

It has to be taken into consideration that we have implemented 25 experiments to evaluate the Accuracy Techniques which we have used in our study for gender separation. These experiments depend on the number and dimensions of clustering. The variation of the number and dimension plays a very important role in the accuracy of gender separation due to the diversity of the feature extraction of every technique as showing the charts and Tables (2.1 to 2.5) that represent the accuracy of MFCC, Auto correlation, Cepstrum, and LPC techniques respectively.

4. CONCLUSION AND FUTURE WORK

Speech is considered as the essential form of communication between humans. It plays a central role in the interaction between human and machine, and between machine and machine. The automatic speech recognition is aimed to extract the sequence of spoken words from a recorded speech signal and so it does not include the task of speech understanding, which can be seen as an even more elaborated problem. Because of the fact that the goal of speech recognition is still far away from the optimal solution for higher accuracy, the gender separation system has been proposed to enhance the performance of speech recognition through building separate gender models by limiting search from whole space of acoustic models that can further lead to improve accuracy of speech recognition.

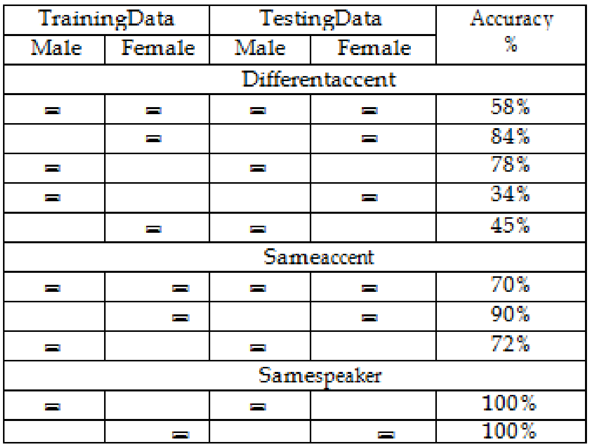
Although the speech data used in this study are collected from different nationalities with different accents (Arabs, Russians, Americans and Indians) and recorded in different sampling rate and channels. High accuracy of gender recognition is obtained by using the techniques which have been mentioned so far.

When the data is mixed from both genders, the accuracy resulted in 58%, and we consider this as allow results of accuracy. But when we separate the gender type, an increase of accuracy is obtained. This appears obviously through the results achieved by the same gender of female which is 84%, while the accuracy results of the same gender of male is 78%. In the step followed, we test the accuracy within the same gender and same speaker and the accuracy resulted is 100%.

In addition, for the training data of male and testing data of female, the accuracy of speech recognition resulted in 45%, while training data of female and testing data of male is resulted in 34%. Moreover, the results of same accents (Indian Accent) appear to be somewhat different for both genders. Their accuracy is 70%. And when they are separated the accuracy of male is 72 % whereas the female's is 90%. To sum up, we conclude that the gender separation and accent plays an important role in increasing the rate of accuracy results of speech recognition. When applying K-means algorithm as pattern recognition for the extraction of features, the results of the experiments for estimated pitch value through Autocorrelation and Cestrum, haves hown 100 % accuracy. MFCC also has show100 % accuracy while giving 90% when LPC is used to estimate formant frequencies. Consequently, we conclude that pitch and MFCC features are more suitable and strongly advised to distinguish the gender type rather than formant frequency alone because it may be more sensitive to noise and accent variation than pitch value and MFCC features.

Speech Recognition Accuracy Results: Speech recognition accuracy has been measured under Sphinx system and the results are as follows:

Future work



TrainingData		TestingData		Accuracy %
Male	Female	Male	Female	
Differentaccent				
▬	▬	▬	▬	58%
	▬		▬	84%
▬		▬		78%
▬			▬	34%
	▬	▬		45%
Sameaccent				
▬	▬	▬	▬	70%
	▬		▬	90%
▬		▬		72%
Samespeaker				
▬		▬		100%
	▬		▬	100%

Table: speechrecognition accuracy of different accent, same accent andsame speaker

In future work, we are planning to expand the range of evaluation set and examine the specific types of errors the system makes in the technique that gives low accuracy. We areal so planning to invest the performance of the system to make it applicable, usable as well as useful for the study purpose in speaker- separation and speaker-verification.

Summary

In our experiments, we have been Able to show that speech recognition accuracy could be improved depending on separate gender model (i.e., using a speaker of the same gender). However, it is argued that following a speaker dependency (male/female) itself leads to some extent to the improvement of the speaker accuracy. That would require gender classification before transcription of each new speaker, but our research has shown that this can be done accurately and in real time.

In both the gender separation accuracy and the speech recognition accuracy of Sphinx transcriptions in our work, we have noticed that the accent variability among speakers plays a crucial role in speech recognition accuracy besides gender. Therefore, it seems that any recognition system attempting to be robust to a wide variety of speakers and languages should make some effort to account for different accents that the system lightens counter. Thus, our initial study in gender separation has demonstrated the benefits of combining gender type for improving speech recognition accuracy through MFCC, and autocorrelation technique shaving better performance than others.

5. ACKNOWLEDGMENTS

Thanks to UGC given to me opportunity to do research as part of Post Doctoral Fellow in IIIT Hyderabad.

6. REFERENCES

[1] Breazeal, C. and Aryanda, L. (2000), 'Recognition of affective communicative intent in robot- directed speech,' in 'Proceedings ofHumanoids2000'.

[2] http://www.ece.auckland.ac.nz/p4p_2005/archive/reports 2003/pdfs/p60_ hlai015.pdf.

[3] S. Davis and P. Mermelstein. Compar ison of parametric represent ations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics Speech and Signal Processing, 28:357–366, Aug1980.

[4] ParrisE.S.,CareyM.I., Language Inde pendent Gender Identification, Proceedings of IEEEICASSP,pp685-688,1996.

[5] Linde,Y.,A. Buzo, and R.M. Gray, "An Algorithm for Vector Quantizer Design, "IEEE Trans. on Communication, 1980, COM-28(1),pp.84-95.

[6] Hartigan, J.A., Clustering Algorithm, 1975, New York, J. Wiley.

[7] Gersho, A., "On the Structure of Vector Quantization," IEEE Trans. On Information Theory, 1982, IT- 28,pp. 256-261.

[8] Richard P. Lappmann, Speech recog nition by Machines and Humans, SPEECH Comm., pp. 1-15, 1997.

[9] Santosh K. Gaikwad, Bharti W. Gawali and Pravin Yannawar, A Review on speech recognition technique, International Journal of Computer Application, Vol. 10(3), pp. 16-24, 2010.

[10] M. Prabha, P. Viveka and Bharathasreeja, Advanced gender recognition system using speech signal, IJSET, Vol.6(4), pp. 118-120, 2016.

[11] Chetana Prakash and Suryakanth V Gangasetty, Fourier-Bessel based cepstral coefficient features for text-independent speaker identification, IICA, pp.913-930, 2-11.

[12] MusaedAlhussein, Zalfiqar Ali, Muhammad Imran and Wadood Abdul, Automatic gender detection based on characteristics of vocal folds for mobile healthcare system, Hindawi, pp. 1-12, 2016.

[13] Suma Swamy and K. V Ramakrishnan, An efficient speech recognition system, CSEIJ, Vol3(4), pp. 21-27, 2013.

[14] Preeti Saini and ParneetKaur, Automatic speech recognition: A review, IJETT, Vol (2), pp. 132-136, 2013.

[15] Bhupinder Singh, Neha Kapur and Puneet Kaur, Speech recognition with Hidden Markow model: A review, IJARCSSE, Vol. 2(3), pp. 400-403, 2012.

[16] M.A Anusuya and S.K Katti, Speech recognition by Machine: A review, IJCSIS, Vol6(3), pp. 181-205, 2009.