Effect of Windowing on the Calculation of MFCC Statistical Parameter for Different Gender in Hindi Speech

Dheeraj Rana
School of IT
GGSIP University
Delhi, India

Anurag Jain
School of IT
GGSIP University
Delhi, India

ABSTRACT
This Paper finds the effects of windowing on the values of mean of first 12 MFCC features excluding energy coefficient for different gender. PRAAT software is used for conducting this experiment which uses Hamming windowing technique by default, Standard low values of window and frame size is used as standard for comparison of MFCC values at increased window and frame sizes by computing Average Deviation from standard values. The main aim of carrying out this experiment is to find out whether all 12 basic MFCCs vary uniformly or not when the window size and subsequently frame size are increased.

To carry out the experiment, a speech database of 8 speakers (5 males & 3 females) is prepared. Each speaker recorded 15 sentences in two emotional states viz. Natural and Anger. The experiments are performed for 7 different cases of window and frame size.

General Terms
MFCC, Emotions, Pattern

Keywords
MFCC, Window size, Frame size, Hamming Window, Mean, Average Deviation

1. INTRODUCTION
For every speech recognition, emotion recognition or speaker identification system the first step is to extract features from speech samples that can be used for further processing of the sample, feature extraction is necessary as voice contains infinite information like emotion, gender and speaker identity . So, generally feature extraction is the process of losing some information contained in the speech and extracting only relevant information required for analysis and sufficient for classification. Mel Frequency Cepstral Coefficients (MFCCs) are widely used for emotion identification, speaker identification, speech recognition and provide satisfactory results i.e. success rates of more than 60 % [3] . MFCC is based on known variation of the human ear's critical bandwidth with frequency, MFCC has types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000 Hz [6]. First two basic steps of MFCC computation are Framing and Windowing. Framing is the process of dividing the speech signal into small frames typically in the range of 5 to 50 milliseconds. The next step windowing is the process to window each frame to reduce discontinuities and leakage at start and end of each frame [1]. MFCC features are calculated for each frame. Detailed explanation of all the steps involved in the computation of MFCCs is available in [1], [6]. For the purpose of speech recognition and speaker identification the mean values of the MFCCs may not prove much helpful as the great detail of individual frames is lost in the process. But the emotional state in a speech sample is the characteristic inherited in the whole sample So, many conventional emotion recognition techniques are based on the mean values of the MFCCs and can successfully recognize emotion with accuracy of more than 60% [5]. MFCC features when analyzed frame wise should be computed with small window size (5-30 ms) because speech signal behaves as quasi stationary in short periods of time that is desirable for distinguishing between vowels, voiced fricatives and voiceless fricatives [4]. But while analyzing the mean of MFCCs then whole speech signal behave as one entity and large window lengths are also used for MFCC computation. This Paper is focused on effect of window size on the mean values of the 12 basic MFCCs. To limit the experimental results to be influenced only by the window size and not by emotional state in the speech sample, this experiment is conducted for two emotional states viz. Natural and Anger. These two emotional states are chosen as these are totally opposite and can be distinguished from each other with an accuracy of around 98% [5]. PRAAT is the software used for conducting this experiment. PRAAT is a standalone program that can be used for analyzing, synthesizing and manipulating speech through an object oriented interface or its programmable scripting language [13]. PRAAT uses hamming windowing technique for calculating MFCC features by default. Selecting the optimal window function for speech processing is still an open challenge [2]. Hamming window function is described mathematically in [1], [6]. These windows have reasonable sidelobe and mainlobe characteristics which are required for DFT computation. Hamming window is shown diagrammatically in Figure 1. Hamming window based baseline system provide very good performance for Speaker Verification Systems [2].

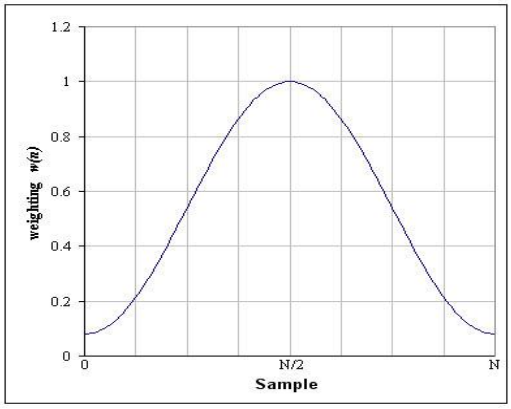

Figure1. Hamming Window Function

2. DATABASE CREATION

India is a country that has a large number of languages and a number of dialects. Hindi is the language spoken by majority of population in India. Hindi is very much phonetic in nature i.e. written symbols and spoken utterances exhibit one-to-one mapping. Detailed explanation about Hindi language can be found in [8]. For performing this experiment, 15 Hindi sentences are given to 8 native speakers (5 males & 3 females) to record utterances in two emotional states Neutral and Anger. Recording of utterances was done in a closed room and noise free environment. For recording a laptop computer, a headphone mic and Audacity software was used. Recorded utterances were then subjected to Background Noise reduction function in Audacity software that minimizes the noise in the recorded utterance to a great extent. The sampling rate used for recording is 16KHz/16bit on mono channel. During recording some portion of the recorded utterance at beginning and end consists mostly as silence portion or some noise. So, these portions are trimmed off the recorded file. In this manner, 150 utterances each of Natural speech and Anger speech are recorded and saved for further processing. Database is divided speaker- wise and gender-wise. Database is stored in different directories on gender basis and each speaker is given an identity for experimental purpose, S1-S5 (Male Speakers) and S6-S8 (Female Speakers). To generate natural emotional feeling of anger in the speakers a long emotional story was dictated to the speakers to induce the same emotional feeling within the speaker. To record neutral utterances no special arrangement needs to be done, any speaker can record an utterance of any sentence long or short in neutral mood. To further check the emotional quality of the utterances, each sentence was uttered 3 times by each speaker and a listening test was also conducted. 10 listeners participated in the listening test, all listeners were in the age group of 17-25 years and with good education. Only those sentences were selected for the experiment whose emotion is identified by more than 75% listeners. Figure 2 shows the list of Hindi sentences used for the speech database.

S.no.	Hindi Sentences	In English Alphabets
1	ये इस महीने का तीसरा सोमवार है।	Ye iss mahine ka teesra somvaar hai.
2	आपको अपनी बाते सोचकर बोलनी चाहिये।	Aapko apni baatein sochkar bolni chaiye.
3	उस आदमी की कद काठी बहुत लम्बी है।	Uss aadmi ki kad kathi bahut lambi hai.
4	इस खेल के विजेता को बडा पुरस्कार मिलेगा।	Iss khel ke vijeta ko bada puraskar milega.
5	हमे एक होनहार आदमी की तलाश है।	Hume ek honhar aadmi ki talaash hai.
6	तुम्हारा बेटा स्कुल जाने लगा होगा।	Tumhara beta school jaane laga hoga.
7	उस पेड पर ऐक कबुतर है।	Uss ped par ek kabootar hai.
8	गुलाव के फूल की खुशबु भीनी होती है।	Gulaab ke phool ki khushboo bheeni hoti hai.
9	ऊपर के कमरे मे कुछ किताबे रखी हुई है।	Upar ke kamre mei kuch kitaabe rakhi hui hai.
10	हमारी परीशा की तैयारी पुरी हो चुकी है।	Hamari pariksha ki taiyari poori ho chuki hai.
11	पूनम घुमने के लिये गोमती एक्सप्रेस आने वाली है।	Poonam ghoomne ke liye gomti express se aane wali hai.
12	मेरा पुराना दोस्त आज भूमि पुजन के लिये आने वाला है।	Mera purana dost aaj bhoomi poojan ke liye aane wala hai.
13	पुनम और पुजा सीधी साधी लडकियाँ है।	Poonam aur pooja seedhi saadhi ladkiya hai.
14	पुनित का घर पुरव दिशा में है।	Punit ka ghar purav disha mei hai.
15	कोयल की आवाज बहुत मीठी होती है।	Koyal ki aawaz bahut meethi hoti hai.

Figure 2. List of Hindi Sentences used for speech database

3. EXPERIMENTS

Experiments are divided in two categories emotion-wise and two subcategories gender-wise

i) Effects of windowing in Neutral emotion
   a) on male speakers
   b) on female speakers
ii) Effects of windowing in Anger emotion
   a) on male speakers
   b) on female speakers

For finding the effects of windowing on MFCCs, 7 different cases of window and frame sizes are considered for the experiment. During the experiment it was found that the frame size within a limit does not have much impact on the values of mean of MFCCs if the window size is at least 3 times of the frame size. So, frame size can also be increased to speed up the computation without affecting the outcome of the experiment. Frame size was not increased above 50 milliseconds during experiments. Seven different cases of window and frame size considered for experiment are designed by nearly doubling the size of window and frame (not more than 50 ms) for successive cases.

Designed cases are as follows:-

(W: *window size*, F: *frame size*)

1) (W: 0.015 ms, F: 0.005 ms) (standard case)
2) (W: 0.03 ms, F: 0.01 ms)
3) (W: 0.06 ms, F: 0.02 ms)
4) (W: 0.1 ms, F: 0.04 ms)
5) (W: 0.2 ms, F: 0.05 ms)
6) (W: 0.4 ms, F: 0.05 ms)
7) (W: 0.5 ms, F: 0.05 ms)

4. METHOD USED

4.1 Feature Extraction

PRAAT uses only 6 input variables with no authority to change the detailed implementation of calculating MFCC. PRAAT has a function *To MFCC* (6 input parameters) that can be invoked by an object of sound type. This function returns 1-24(max) number of coefficients plus one energy coefficient. Energy coefficient is not included in this study. Input values to the function *To MFCC* used for this experiment are :-

1)	Number of coefficients	12
2)	Window length (in sec)	Variable
3)	Time step (in sec)	Variable
4)	Position of first filter (in mel)	100 Hz
5)	Distance between filters (in mel)	100 Hz
6)	Maximum frequency (in mel)	0 (N.A.)

MFCC values calculated is saved in the format of Spreadsheet file for carrying out further analysis using Microsoft Excel.

4.2 Analysis of Data

MFCCs obtained for first case of low window and frame size are taken as standard and MFCCs for cases of increased values of window and frame size are then compared with MFCCs of standard case for analyzing effect of window size on their values. Steps followed for analysis of MFCC data obtained are (in continuation) :-

i) Mean of each MFCC is calculated for each sentence (15 sentences).
ii) Mean of 15 resultant values of each MFCC is calculated.
iii) Above two steps are repeated for all 7 cases of window and frame size.
iv) Average deviation is calculated of MFCCs obtained for cases 2 to 7 from MFCCs of first case (standard low values of window and frame size) .
v) Average deviation of each MFCC is calculated by formula

$$\frac{1}{6}\sum_{n=2}^{7}(ABS\,(MFCC(n) - MFCC(1))$$

vi) Repeat steps i) – v) for each speaker and different emotion.
vii) Take mean of the average deviations of each speaker (gender- wise).

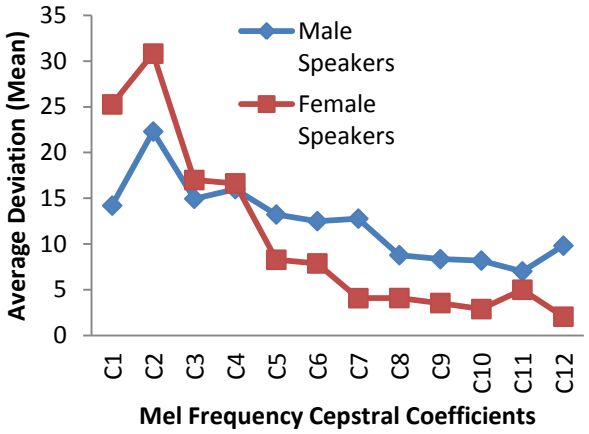

Figure 3(A) Average Deviation of MFCCs by changing window and frame size for Neutral emotion

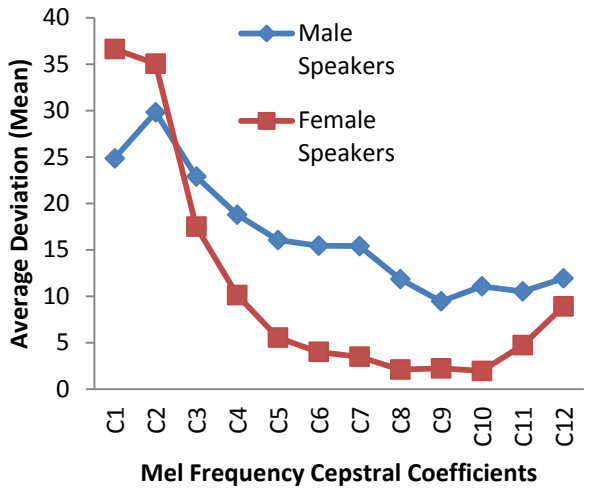

Figure 3(B) Average Deviation of MFCCs by changing window and frame size for Anger emotion

5. RESULTS AND DISCUSSION

The results of the experiments i) and ii) conducted are depicted in Table I and is plotted in Figure 3(A) and Figure 3(B). Findings obtained after the study of the data produced in the experiments shows that values of low dimensional MFCCs change frequently as compared to high dimensional MFCCs for both male and female speakers and for both emotions i.e. not all MFCCs vary uniformly with change in the window size. Low dimensional MFCCs vary with high degree in anger emotion than in neutral emotion for both male and female speakers.

For female speakers, change in the values of low dimensional MFCCs is very high as compared to the high dimensional MFCCs for both emotions but more clearly visible in anger emotion. For male speakers, this discrepancy between low and high dimensional MFCCs also exist but pattern is nearly same for both emotions.

Summary of the results is that mean value of coefficients C1 to C6 and C12 are relatively more sensitive to change in the window size than the coefficients C7 to C11. This different behavior of coefficients can cause considerable alteration to outcome of experiments but it is concluded in [5] that coefficients C1 to C6 have higher accuracy than C7 to C12 in

recognizing emotions, thus the large variation seen between the values of coefficients at increased window sizes for different gender and emotion suggest that better result in emotional classification can be provided by taking mean of MFCCs at large values of window size.

6. CONCLUSION

In this paper, we found the effects of changing window size on the calculation of MFCC mean and the results presented in Table suggest varying sensitiveness among MFCCs to window size and found how taking mean of MFCCs at large values of window size will increase accuracy in emotion recognition. There are other patterns also found through the data in the table that may be used for gender identification or even may be for emotional classification between some selected emotions by comparing MFCC mean values calculated using a small window size and a large window size.

Table I

Average Deviation of 12 MFCCs (Mean) over 15 different Hindi Language sentences calculated at different values of window length and frame length from standard (window,frame) length
***Standard (W,F)* : (0.015, 0.005) in sec ; *Sp(G)*: Speaker Id (Speaker Gender)**

A) Emotion: Neutral

Sp(G)	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
S1 (M)	12.066	17.95	16.738	15.357	13.785	14.183	13.517	11.258	8.3034	8.9277	7.6419	11.629
S2 (M)	5.9849	21.023	6.7079	7.868	10.662	8.1231	10.283	7.414	5.4039	5.9289	3.6204	4.0091
S3 (M)	20.217	25.304	17.831	17.959	15.268	13.725	15.58	8.6753	8.1743	9.4797	6.0509	11.089
S4 (M)	21.626	23.964	21.434	19.816	14.295	13.678	11.515	8.822	9.3905	9.7437	10.212	11.663
S5 (M)	11.056	23.237	11.997	18.984	12.084	12.778	12.961	7.6986	10.426	6.821	7.5025	10.658
Mean	14.19	22.296	14.941	15.997	13.219	12.497	12.771	8.7735	8.3396	8.1802	7.0055	9.8095
S6 (F)	24.909	32.666	15.848	18.439	8.5146	9.6444	4.9474	4.2217	4.8266	3.9225	8.0192	1.2447
S7 (F)	29.512	29.666	14.063	11.113	3.2223	2.3356	1.3522	1.1085	1.0389	2.6545	1.2787	3.1458
S8 (F)	21.361	30.114	21.067	20.32	13.082	11.63	5.9155	6.9245	4.7468	2.0605	5.6802	1.7371
Mean	25.261	30.815	16.993	16.624	8.273	7.8699	4.0717	4.0849	3.5375	2.8792	4.9927	2.0425

B) Emotion: Anger

Sp(G)	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
S1 (M)	21.442	27.184	25.401	18.289	17.583	17.947	18.55	14.888	11.579	11.676	11.548	14.938
S2 (M)	17.621	28.146	13.157	14.21	17.365	12.785	18.07	15.71	9.8646	13.312	9.6047	9.878
S3 (M)	27.693	27.01	21.038	17.864	15.705	16.438	15.551	9.6146	9.1514	12.691	11.207	14.339
S4 (M)	30.53	32.879	30.61	21.546	14.57	14.419	10.104	8.3645	4.6063	5.5932	7.3193	4.3117
S5 (M)	26.938	33.826	24.234	21.951	14.982	15.596	14.691	10.588	12.043	12.031	12.887	16.143
Mean	24.845	29.809	22.888	18.772	16.041	15.437	15.393	11.833	9.4489	11.06	10.513	11.922
S6 (F)	33.006	35.714	16.528	11.754	1.8897	0.4576	2.0311	0.3606	0.8323	1.0858	1.0986	7.2483
S7 (F)	46.454	32.862	11.12	2.6204	5.0512	4.0982	6.0543	2.3383	4.3525	2.8764	9.1284	17.25
S8 (F)	30.417	36.564	24.831	15.975	9.6758	7.4477	2.3563	3.5856	1.4679	1.8457	3.9386	2.1992
Mean	36.626	35.047	17.493	10.117	5.5389	4.0012	3.4806	2.0948	2.2176	1.9359	4.7218	8.8992

7. ACKNOWLEDGMENTS

The authors are thankful to the Guru Gobind Singh Indraprastha University for providing support and students who participated in the preparation of the speech database.

8. REFERENCES

[1] Jain, A., Prakash,N. and Agrawal, S.S. 2011. Evaluation of MFCC for Emotion Identification in Hindi Speech. In Communication software and Networks (ICCSN) Proceedings, IEEE 3rd International Conference on, May, Xi'an, China.

[2] Sahidullah, Md. and Saha, G. 2012.A Novel Windowing Technique for Efficient Compuation of MFCC for Speaker Recognition. IEEE signal processing letters, vol. 20, no. 2, 2012, pp 149-153.

[3] Kandali, A.B., Routray, A. and Basu, T.K. 2008. Emotion Recognition from Assamese Speech using MFCC features and GMM classifier In TENCON 2008, IEEE Region 10 Conference, 2008, Hyderabad, India.

[4] Kelly, C.K. and Gobl, C. The Effects of windowing on the calculation of MFCCs for different types of speech sounds. In NOLISP'11 Proceedings 5th international conference on Advances in nonlinear speech processing, Nov, 2011.

[5] Sato, N. and Obuchi, Y. " Emotion Recognition using Mel Frequency Coefficients", Journal of Natural Language Processing, Information and Media Technologies, vol. 2, no. 3, pp. 835-848, September, 2007.

[6] Muda, L., Begum, M. and Elamvazuthi, I. , "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Waroing (DTW) Techniques", Journal of Computing, vol. 2, issue 3, pp. 138-143, March, 2010.

[7] Joshi, D.D. and Zalte, M.B. " Recognition of Emotion from Marathi Speech Using MFCC and DWT Algorithms", IJACECT, vol. 2, issue 2, 2013.

[8] Agrawal, S.S. Emotions in Hindi Speech- Analysis, Perception and Recognition. in Speech Database and Assessments. In (Oriental COCOSDA) Proceedings, IEEE International Conference on, October 26-28, 2011, Hsinchu, Taiwan.

[9] Ittichaechareon, C., Suksri, S. and Thaweesak, " Speech Recognition using MFCC", ICGSM, July 28-29, 2012, Pattaya, Thailand.

[10] Zheng, F. , Zhang, G. and Song, Z. " Comparisons of Different Implementations of MFCC", Journal of Computer Science and Technology, vol. 16, no. 6, pp. 582-589, September, 2011.

[11] Khan, S., lslam, M.R. and Faizul, M. Automatic Speaker Recognition In 3rd international conference on electrical and computer engineering (ICECE), December 28-30,2004, Dhaka, Bangladesh.

[12] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[13] http://www.fon.hum.uva.nl/praat/