Study of Speaker's Emotion Identification for Hindi Speech

Sushma Bahuguna
BCIIT, New Delhi, India
sushmabahuguna@gmail.com

Y.P Raiwani
Dept. of Computer Science and Engineering,
HNB Garhwal University Srinagar,
Uttrakhand, India.
yp_raiwani@yahoo.com

***Abstract* - Emotion based speaker Identification System is the process of automatically identifying speaker's emotion based on features extracted from speech waves. This paper presents experiment with the building and testing of a Speaker's emotion identification for Hindi speech using Mel Frequency Cepestral Coefficients and Vector Quantization techniques. We collected voice samples of Hindi speech sentences in four basic emotions to study speaker's emotion identification and it was found that with proposed emo-voice model we are able to achieve accuracy of 73% of speaker's emotion identification in speech out of 93% of the total speech samples provided to the system.**

Keywords: **Emo-voice model, MFCC, prosodic features, spectral features, Vector Quantization.**

I. INTRODUCTION

The human speech contains and reflects information about the emotional state of the speaker. Emotion plays an important role in verbal communication and interaction allowing people to express them their views. Human computer interaction could be more effective when the accurate emotional information in speech could be identified [1, 2, 3]. These applications can then be used in areas such as health, call centers, education etc. where there is more use of human computer interaction. There have been several researches carried out to identify emotional state from speech for different languages. For performing experiment in Hindi speech we collected voice samples of five male and female speakers of different age groups expressing sentences in Hindi each frequently used in everyday communication in four basic emotions namely Happy(H), Natural (N), Sad (S), Anger (A). Emotional speech databases of 20 sample sentences in Hindi are used for emotion expressions Table 1.

Speaker	Age (Yrs)	Gender	Emo#
Spk1	34	Female	4
Spk2	40	Male	4
Spk3	14	Male	4
Spk4	25	Female	4
Spk5	31	Female	4

Table [1]: Specifications of the voice sample

II. FEATURE EXTRACTION

Prosodic and Spectral features extracted from speech are used in emotion identification. Each speaker has unique physiological characteristics of speech production and speaking style and speaker-specific characteristics are reflected in prosody. It is generally recognized that human listeners can better recognize speakers. In most of the ASR-free approaches, pitch contour dynamics are represented using parameters derived from linear stylized pitch segments, which has the advantage that features are derived directly from the speech signal [4]. Spectral features are represented by MFCC and prosodic features are represented by pitch and energy contours [5]. Feature extraction is the process of reducing data while retaining speaker discriminative information. Our task is to train an emo-voice model for each speaker using the corresponding sound file. We have used MFCC coefficients and efficient classifying method Vector Quantization for performing text-independent identification.

A. Mel Frequency Cepstrum Coefficients

Mel Frequency Cepstrum Coefficients (MFCC) processor is mainly used to emulate the behavior of the human ears. The steps for computing MFCC are shown in Figure [1]. It is a representation of MFCC calculation

process [6] which shows the digital speech signal of s1_natural_01.wav analog file. In the first step of MFCC calculation , preprocessing covers digital filters and signal detection. Next in frame blocking, the speech signal is blocked into frames of N samples, the adjacent frames are separated by M (M < N), where N = 256 (which is equivalent to ~ 30 ms windowing and facilitate the fast radix-2 FFT) and M = 100 [7, 10].

The next step in the processing is to window every frame to minimize the signal discontinuities at the start and end of each frame. We define the window as

$w(n), 0 \le n \le N-1$, where N = number of samples in each frame

$y_l(n) = x_l(n)w(n), \quad 0 \le n \le N-1$, where *y(n)* = Output signal, *x (n)* = input signal, *w(n)* = Hamming window [12]. The result of windowing signal is

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

Fast Fourier Transform converts each frame of N samples from the time domain into the frequency domain. Here we take the Discrete Fourier Transform (DFT) of each frame, which is defined on the set of N samples $\{x_n\}$, as follow:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \qquad k = 0,1,2,..., N-1$$

Next step shows the mel-frequency scale which represents linear frequency spacing < 1000 Hz and a log spacing > 1000 Hz, based on non linear perception of frequencies of audio signals by human ear. Thus for each speech wave with actual frequency, (*f)* Hz, a subjective pitch is measured on a scale called the 'mel' scale. The *mel-frequency* scale is *F(Mel)=[2595 * log 10 [1 + f] 700]* [7,10,11]. After this step using DCT the real numbers (log mel spectrum and their logarithm) are converted back into time domain, to get the mel frequency cepstrum coefficients (MFCC) [7,10].

The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis, $\tilde{S}_0, k = 0,2,..., K-1$, we can calculate the MFCC's, $\tilde{C}_n$, as [11]

$$\tilde{C}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right],$$

$n = 0.1..... K-1$

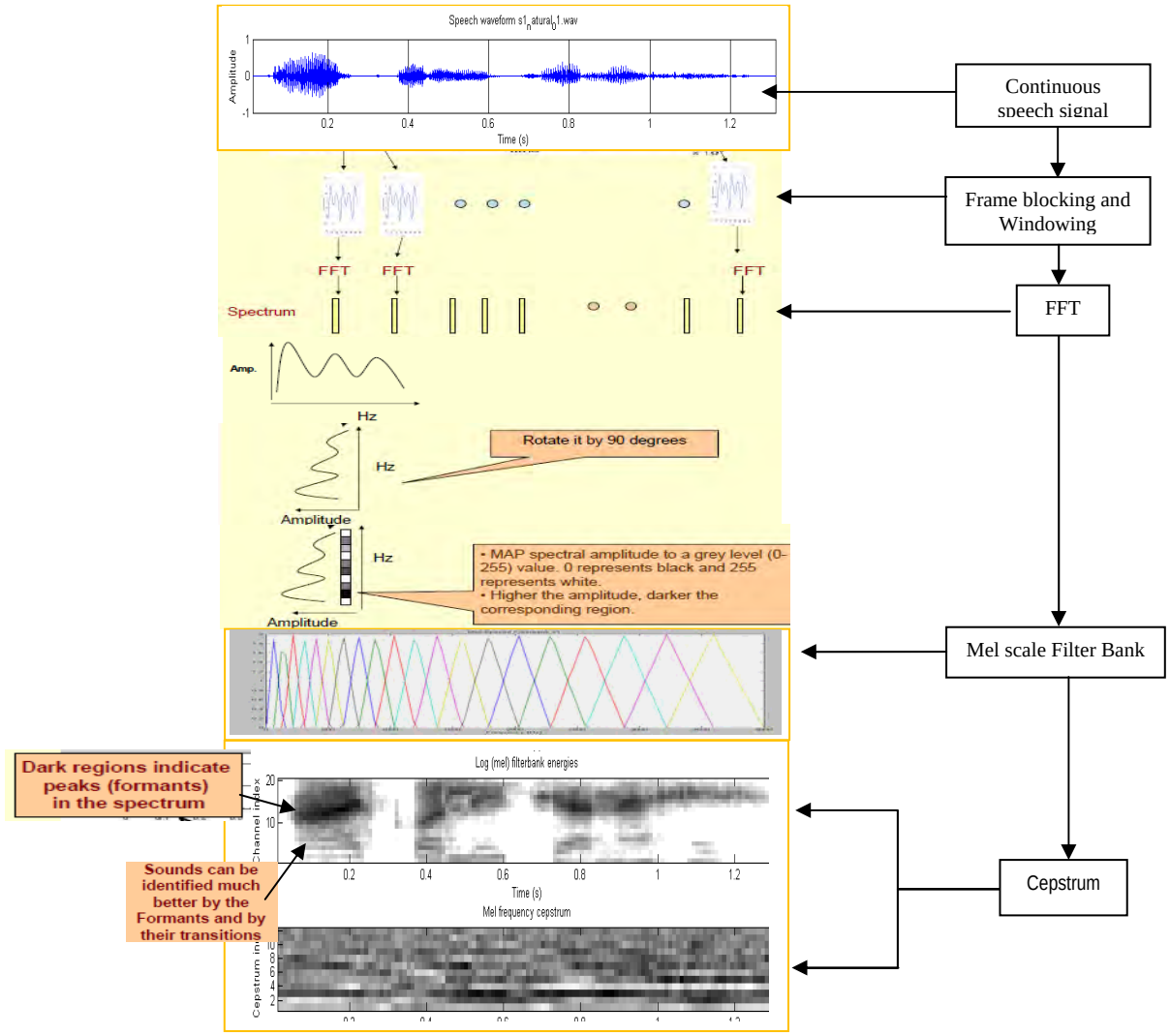


Figure [1]: MFCC Flow chart and diagram

B. Vector Quantization

In this method, VQ code-books consisting of a small number of representative feature vectors are used as an efficient means of characterizing speaker emotion specific features [13, 14]. Figure [2] represents Block diagram of emotion identification system using VQ.

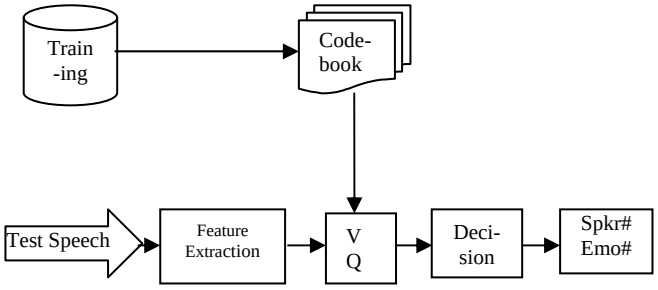


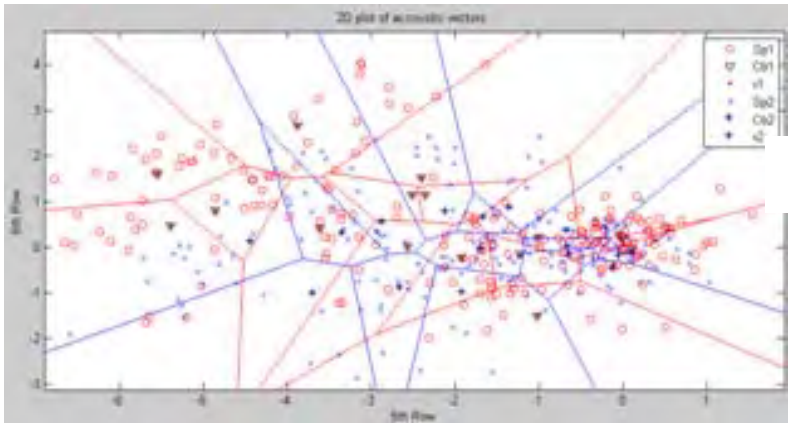
Figure [2]: Block diagram of Emotion identification System using VQ.

A speaker-specific VQ code-book is generated by clustering the training feature vectors of each speaker (Figure [3]). After which the system would have information of the emo-voice characteristic of each (known) speaker. In testing phase, the system recognizes the (assumed unknown) speaker's emotion of each speech file

in the testing folder. The system would then be able to recognize which registered speaker's emotion provides a given utterance from amongst a set of known speaker's emotional speech. In the recognition phase an unknown speaker, represented by a sequence of feature vectors {x1,…, xT }, is compared with the codebooks in the database. For each codebook a distortion measure is computed, and the speaker with the lowest distortion is chosen (Table [2]). One way to define the distortion measure is to use the average of the Euclidean Distances [8]. The Euclidean distance is the distance between the two points that can be measured with a ruler, which can be proven by repeated application of the Pythagorean Theorem.

The Euclidean distance is defined by: $d(x, y_i) = \sqrt{\sum_{j=1}^{k} (x_j - y_{ij})^2}$ where x_j is the j^{th} component of the input vector, and y_{ij} is the j^{th} component of the codeword y_i [9].

Thus, each feature vector in the sequence X is compared with all the codebooks, and the codebook with the minimized average distance is chosen to be the best.



Figure[3]: Result of codewords in 2 dimentional space of two speech files. The codewords are marked for two different speakers speaking same sentence in same emotion. The voronoi regions for speaker1 is separated by red boundary and for speaker2 by blue boundary.

The emotional speech of a speaker corresponding to the VQ codebook with least total distortion is recognized as the emotion of speaker of the input speech. Table [2] shows the sample speech test conducted for 19 speech files in the train database with 7 speech files in the test database.

Sno	Spkr No	Emo Rec	Emo Act	Distortion																		Min dis
				trn_1_a	trn_1_h	trn_1_n	trn_1_s	trn_2_a	trn_2_h	trn_2_n	trn_2_s	trn_3_a	trn_3_h	trn_3_n	trn_3_s	trn_4_a	trn_4_h	trn_4_n	trn_4_s	trn_5_a	trn_5_h	trn_5_n	
1	1	angry	angry	5.687	6.964	7.359	6.584	6.29	6.791	6.463	6.985	9.173	9.038	8.689	8.853	7.99	9.329	8.047	9.12	8.102	9.168	9.074	5.687
2	1	happy	happy	6.94	6.659	7.775	6.878	7.414	7.255	7.62	7.471	8.657	8.608	8.828	9.86	7.868	9.09	8.31	8.64	7.826	7.973	7.866	6.659
3	1	natural	natural	7.047	7.445	6.056	6.708	7.543	7.862	6.708	7.762	8.345	8.931	8.077	9.599	8.438	7.723	6.669	8.795	8.202	9.447	9.147	6.056
4	1	sad	sad	6.365	7.955	6.876	6.063	6.59	6.874	6.782	7.11	7.528	8.43	7.791	9.221	7.855	7.908	7.117	8.272	7.727	8.124	7.852	6.063
5	2	angry	angry	5.687	7.735	7.018	6.651	5.673	6.32	6.076	6.624	8.624	8.754	8.718	9.403	7.779	8.916	7.927	8.558	9.967	9.535	9.556	5.673
6	2	happy	happy	6.365	7.974	6.844	6.798	6.251	5.133	6.046	6.884	8.035	8.052	8.196	9.564	7.696	7.822	7.192	7.434	9.448	9.478	10.17	5.133
7	2	natural	natural	6.752	7.429	6.544	6.135	6.594	7.13	5.719	6.933	8.07	8.413	8.321	9.323	8.157	8.084	6.804	8.211	9.086	9.005	8.989	5.719

Table [2]: Distortion calculated for 7 test speech samples with the 19 trained data.

III. RESULTS

Analysis of building and testing of an automatic emotion based speaker identification system using MATLAB environment show resultant matrix (Table [3]) after the testing each speech file in the test database with the trained vectors generated for computer speaker emotion identification system.

			Predicted Class																			
	SP KR	KR	sp1				sp2				sp3				sp4				sp5			
	KR	Emo	A	H	N	S	A	H	N	S	A	H	N	S	A	H	N	S	A	H	N	S
Known Class	sp1	A	20																			
		H	3	11			1	2						1								
		N			20																	
		S	2			13				3				2								
	sp2	A	1				17	1	1													
		H					1	16														
		N				1			19													
		S				2			1	15				2								
	sp3	A									13		2	5								
		H	1	1							2	12	2	2								
		N				1					3	2	8	4			2					
		S									1	4	5	10								
	sp4	A	1					1	1			1			9	4		3				
		H										1			1	18						
		N															20					
		S						1		2					8	1		8				
	sp5	A																	9	6		
		H																		13		2
		N																			15	
		S																		2	3	7

Table [3]: Matrix - Computer Speaker Emotion Identification.

We can depict the following results from the resultant matrix of 372 emotional speech samples (Table [3]):

A. Speaker Identification without Emotions:

- Total correct prediction made speakers wise 344/372 i.e. 92.47%
- The error rate is 28/372 i.e.7.53%

B. Speaker Identification with Emotions:

- The model made 273 correct predictions of Emotions.
- The model made 99 incorrect predictions of Emotions.
- The model scored 372 cases (273 + 99).
- The error rate is 99/372 i.e. 26.61%
- The accuracy rate is 273/372= 73.39%.

In voice authentication there are some user influences that affect the speech and emotion of a speaker, must be addressed like cold, expression and volume, misspoken or misread prompted phrases, previous user activity, background noises etc.

IV. Conclusion

In the study we have used techniques of MFCC and VQ for identification of speaker speaking in different emotions and applied to text independent speaker's identification system. The result shows that with proposed method we are able to achieve 73.39% of speaker's emotion identification in speech by system. The experiment has been performed on small utterances and database could be enhanced to achieve more accuracy. The results of our experiments are limited to recognize the speaker based on the devices used for recording the corresponding speech files.

References

[1] Takashi Fujisawa & norman D. Cook, "*Identifying Emotion in speech Prosody ushing Acoustical Cues of Harmony*." INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004. ISCA 2004

[2] Jian Zhou, Guoyin Wang, Yong Yang, Peijun Chen., *"Speech Emotion Recognition Based on Rough Set and SVM"*, Proc. 5th IEEE Int. Conf. on Cognitive Informatics (ICCI'06), @)2006.

[3] R Cowie et. al., *"Emotion recognition in human-computer interaction"*, Signal Processing Magazine, IEEE 18 (1), 32-80

[4] Leena Marya,, B. Yegnanarayana b, *"Extraction and representation of prosodic features for language and speaker recognition"*, ScienceDirect, Speech Communication 50 (2008) 782–796, Elsevier

[5] K Sreenivasa Rao and Shashidhar G Koolagudi, *"Identification of Hindi Dialects and Emotions using Spectral and Prosodic features of Speech"*, Journal of Systemics, Cybernetics & Informatics;2011, Vol. 9 Issue 4, p24

[6] P. Chakraborty et al., *"A automatic speaker recognition system"*, Neural Information Processing: 14th International Confernce, ICONIP 2007

[7] Digital Signal Processing Mini-Project: - University of Illinois, http://www.ifp.illinois.edu/~minhdo/teaching/speaker_recognition/.

[8] Benjamin J. Shannon and Kuldip K. Paliwal., MFCC Computation from Magnitude Spectrum of Higher Lag Autocorrelation Coefficients for Robust Speech Recognition.

[9] http://www.mqasem.net/vectorquantization / vq. html

[10] *http://eng.najah.edu/sites/eng.najah.edu /files/finalreport_SI.doc*
[11] V. Tiwari, *"MFCC and its applications in speaker recognition"*, International Journal on Emerging Technologies 1(1): 19-22(2010)
[12] Anjugam M,Kavitha M, *"Design and Implementation of Voice ControlSystem for Wireless Home Automation Networks"*International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012.
[13] M. D. Pawar et. al, "Speaker Identification System Using Wavelet Transformation and Neural Network" International Journal of Computer Applications in Engineering Sciences [VOL I,SPECIAL ISSUE ON CNS, JULY 2011]
[14] Patricia Melin et. al.,*" Voice Recognition with Neural Networks, Type-2 Fuzzy Logic and Genetic Algorithms"*, Engineering Letters, 13:2, EL_13_2_9 (Advance online publication: 4 August 2006)