IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)
e-ISSN: 2278-1676,p-ISSN: 2320-3331, Volume 9, Issue 2 Ver. II (Mar – Apr. 2014), PP 27-31
www.iosrjournals.org

Embedding Prosody into Neutral Speech

Ms. C.D.Pophale, Prof.J.S.Chitode
[1](, Department of Electronics Engg., Bharati Vidyapeeth Deemed University college of Engg.Mharashtra,India.411043).
[2](, Department of Electronics Engg., Bharati Vidyapeeth Deemed University college of Engg.Mharashtra,India.411043).

Abstract:*This paper presents and discusses development of emotion specific in Marathi Speech. The emotions are classified into four categories i.e anger, happinesss, sadness, surprise. There are different methods of speech synthesis are discussed, e.g.LMM, CART, GMM. Sentences of different and same words in all emotion are recorded by students of drama club. Prosody related features and spectral related features were analyzed to synthesize neutral speech. 17 mel cepstral coefficients (MFCCs) were studied as prosody related featurs consisted of vocal tract frequency (f1), speech energy, duration.It is seen that value of all prosodic parameter have highest value for anger speech and lowest value for neutral speech.*

Keywords: *Emotional speech, prosody analysis, speech synthesis,GMM,LMM,TTS*

I. INTRODUCTION

RECENTLY, more and more efforts have been made in the research for expressive speech synthesis, among which emotion is a very important element [12], [13]. Some prosody features, such as pitch variables (F0 level, range, contour, and jitter), and speaking rate have already been analyzed [14], [15].There are also some implementations in emotional speech synthesis. For instance, Mozziconacci [7] added emotion control parameters on the basis of tune methods, resulting in higher performance. A typical system was produced by Campbell [9], who created an expressive speech synthesis from a corpus gathered over five years and gave impressive synthesis results. Schroeder [10] and Eide [11] generated an expressive text-to-speech (TTS) engine which can be directed, via an extended speech synthesis markup language, to use a variety of expressive styles from about 10 h of "neutral" sentences. There are different prosody conversion methods are introduced. Which aim at the transformation of the prosodic parameters, e.g., F0, duration, pitch and Energy of the given utterance. To generate emotional speech, linear modification model (LMM), a Gaussian mixture model (GMM) method and a classification and regression tree (CART) method were tried. The LMM makes direct modification of F0 contours (F0 top, F0 bottom, and F0 mean), syllabic durations, and intensities from the acoustic distribution analysis results. The GMM method attempts to map the prosody features distribution from a "neutral" state to the various emotions,

While the CART model links linguistic features to the prosody conversion. The GMM method attempts to map the prosody features distribution from a "neutral" state to the various emotions, while the CART model links linguistic features to the prosody conversion.[16].

Emotions are expressed in speech, face, gait and other body languages explicitly by human beings along with internal physiological signals such as muscle voltage, blood volume pressure, skin conductivity and respiration. The vocal expressions are harder to regulate than other explicit emotional signals. So, it is possible to know the actual affective state of the speaker from her/his voice without any physical contact. But exact identification of emotion from voice is very difficult due to several factors. The speech consists broadly of two components coded simultaneously :(i) "What is said" and (ii) "How it is said". The first component consists of the linguistic information pronounced as per the sounds of the language. The second component consists of non-linguistic or paralinguistic or suprasegmental component which includes the prosody of the language i.e. pitch, intensity and speaking-rate rules to give lexical and grammatical emphasis for the spoken messages and the prosody of emotion to express the affective state of the speaker. In addition, speakers also possess their own style, i.e. a characteristic articulation rate, intonation habit and loudness characteristic. Thus, isolation of the affective information i.e. the emotion, from voice is not easy.[17]

In this paper ,the study of vocal emotion is carried out using speech samples of Indian language: Marathi which is native language of the state of mshtrahara.The total number of mono-sounds(vowel ,semivowel and consonants) in these language is up to 40.The present work investigates how the speech parameter e.g. vocal tract frequeny,spectrum energy ,fundamental frequency ,duration represent the emotion like sadness,happiness,surprise,anger.

In this paper, 17 Mel Frequency Cepstral Coefficient (MFCC) are estimated which helps to estimate the duration of vowels,semivowels,and consonenst.HMM window frame of 30 msec is used to estimate the vocal tract frequency.MATLAB 10 software is used for programming

II. Data Collection

The following equipment and software are used in the recording process: (i) a wearable headphone-mic set for single channel recording, (ii) a notebook computer with onboard sound interface having 16-bit depth and 44.1kHz sampling frequency, (iii) an almost noise-free small closed room.20 sentences having different words in all emotion s are recorded by male and female students of drama club . And 20 sentences having same words in all emotions are also recorded. The subjects are asked to rehearse their acting a few times before final recording. The distance between the Microphone and the mouth of the subject and the Line-in Volume of the notebook are simultaneously adjusted such that the speech waveforms are not clipped while recording.

III. Listening Test

A listening test of the emotional utterances is carried out with the help of 6 randomly selected volunteer human judges i.e. listeners (3 Males and 3 Females) for the Marathi language database.

IV. Feature Extraction

For each utterances following pre processing tasks are done:(i)all utterances are sampled at 8.1khz sampling frequency,(ii)the frame duration is chosen to be of 30msec.Seventeen MFCC coefficients and energy feature are computed from each hamming window frame using 17 triangular mel frequency filter banks. 23 formant frequencies are estimated. Teager() function from MATLAB 10 software is used to estimate the spectrum energy.Duretion for vowel(d11),semivowel(d12),consonant-(d13) are estimated using lpc coefficients. A. Feature Extraction of MFCCs The first purpose to explore the spectral features by using the Mel-frequency cepstral coefficients (MFCCs) is that they have been widely employed in speech recognition due to superior performance when compared to other features. [1] The Mel-frequency cepstrum is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. For each speech frame of 30 ms, a set of Mel-frequency cepstrum coefficients was computed. Fig. 1 shows the MFCC feature extraction process containing following steps:

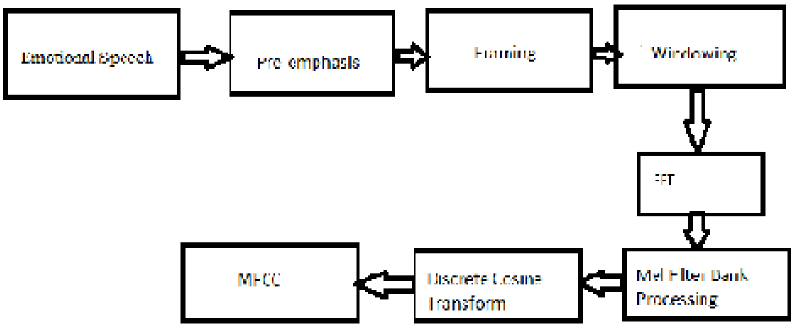

Fig.1.Mel- Frequency Cepstral Coefficients

Step 1: Pre–emphasis

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

$Y[n] = X[n] * 0.95 X [n-1]$ …… (1)

Step2: Framing

It is a process of segmenting the speech samples obtained from the analog to digital conversion (ADC), into the small frames with the time length within the range of 20-40 ms. Framing enables the non stationary speech signal to be segmented into quasi-stationary frames, and enables Fourier Transformation of the speech signal. It is because, speech signal is known to exhibit quasi-stationary behaviour within the short time period of 20-40 ms.

Step3: Windowing

Windowing step is meant to window each individual frame, in order to minimize the signal discontinuities at the beginning and the end of each frame. Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines.

The Hamming window equation is given as:

If the Hamming window is defined as W (n), $0 \le n \le N- 1$ where, N = number of samples in each frame, Y[n] = Output signal, X (n) = input signal ,W (n) = Hamming window, then the result of windowing signal is shown below:

$Y(n) = X(n)\ \text{X}\ W(n)$

$W(n)=0.54-0.46\cos(2\pi n/N-1)$ ………$0 \le n \le N- 1$…(2)

Step4: FFT

FFT converts each frame of N samples from the time domain into the frequency domain. The Fourier Transform is to convert the convolution of the input pulse and the vocal tract impulse response in the time domain. This statement supports the equation below:

$Y(w) = FFT [h(t) * X(t)] = H(w) * X(w)$ ……..(3)

Step5: Mel Filter bank and Frequency wrapping:

The mel filter bank consists of overlapping triangular filters with the cut off frequencies determined by the center frequencies of the two adjacent filters. The filters have linearly spaced centre frequencies and fixed bandwidth on the mel scale. Then, each filter output is the sum of its filtered spectral components. Following equation is used to compute the Mel for given frequency f in HZ: F(Mel) = [2595* log 10 [1+F] 700] ………(4)

Step6: Discrete Cosine Transform

It is used to orthogonalise the filter energy vectors. Because of this orthogonalization step, the information of the filter energy vector is compacted into the first number of components and shortens the vector to number of components

IV. Result And Discussion

Following figures shows graph between gain Vs Frequency. Fig.2.shows the gain vs frequency graph for the emotion anger,the words are”Kay kartoy he?Kiti vela tech tech sangayach?” Fig.3 shows the graph for emotion happiness and the words are,”Kiti sunder phul aahe he?”.fig.4 shows the emotion surprise and the words are ,”Kay?tyzyakade paise nahit?”.fig.5 shows the emotion sadness and the words are ,”Mala kahich karata yet nahi”

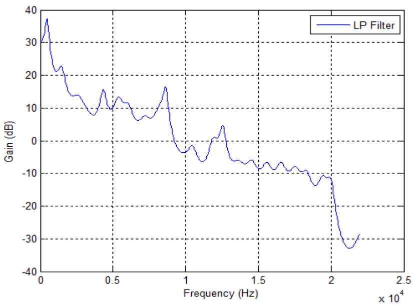

Fig.2 Anger

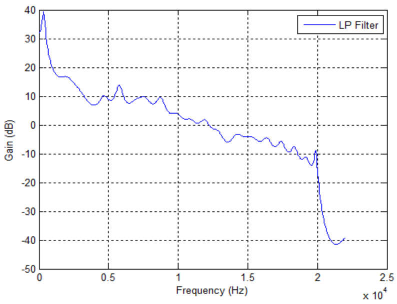

Fig.3.Happiness

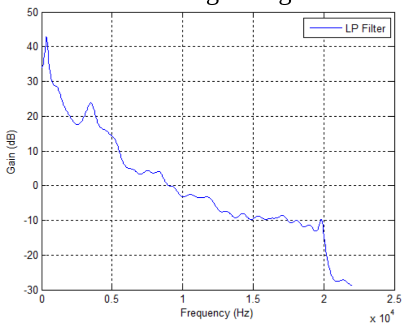

Fig.4 Surprise

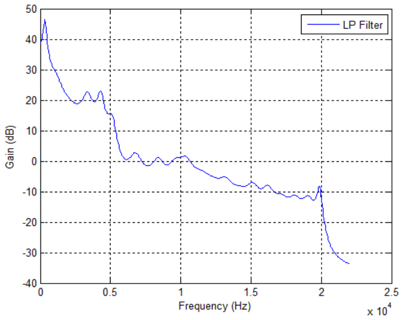

Fig.5.Sadness

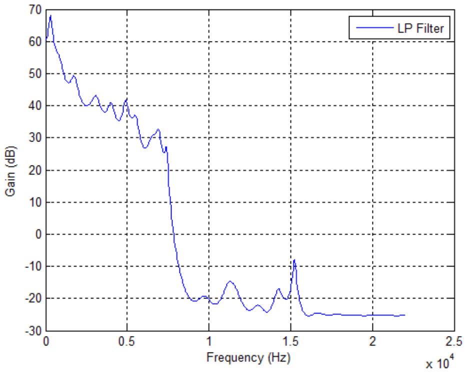

Fig.6 Neutral

Fig 7 to 11 shows how the amplitude varies for different emotion for the same sentences Which are mentioned above. For anger, it shows maximum amplitude, for happiness and surprise it has near about same amplitude. Neutral having lowest amplitude.

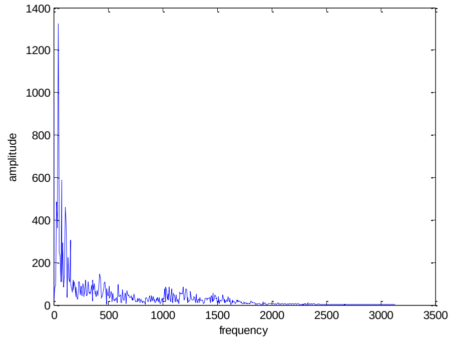

Fig.7.Anger

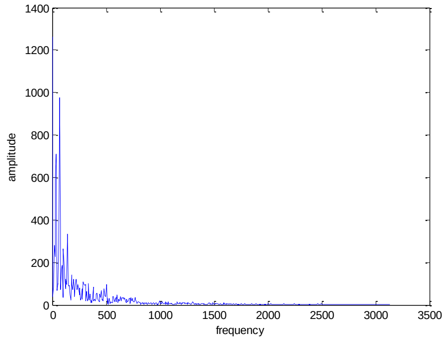

Fig.8.Happiness

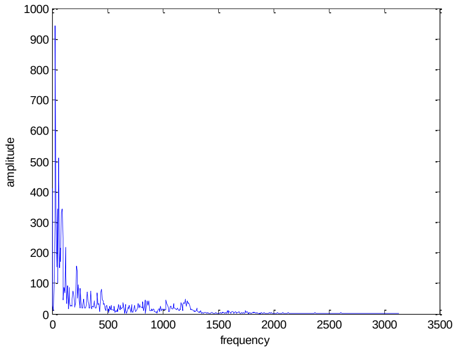

Fig.9.Surprise

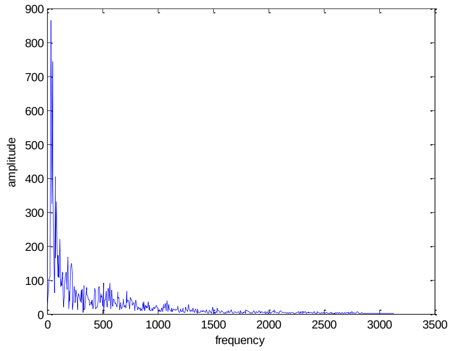

Fig.10 Sadness

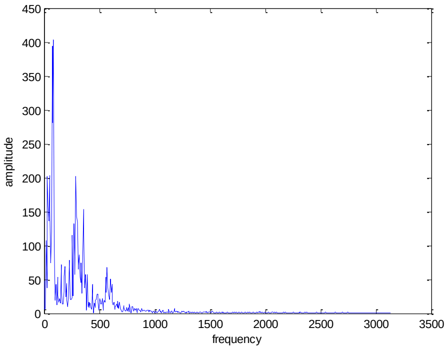

Fig.11 Neutral

The value of parameters Energy, Vocal tract frequency(F1), duration related to emotions angry, happiness, sadness, surprise and neutral are shown in table.

TABLE I. Distribution of Prosodic Parameters IN Different Emotions(pp:Prosodic Parameter)

Emotions PP	Anger	Happiness	Surprise	Sadness	Neutral
E(DB)	155	137	123	106	90
D(ms)	3.9	3.7	3.5	2.4	1.9
F_1(HZ)	42.9	42.9	36.8	30.6	25.9

From above Table it shows that Energy goes on decreasing from anger to neutral speech. Anger speech has highest energy. In most of the cases happiness and surprise have same energy. Sadness have lowest energy. Duration (Vowel+semivowel+consonent) is also goes on decreasing as we go from anger to Neutral speech. Vocal tract frequency of anger and happiness is mostly same. Sadness having lowest frequency.

It is complicated to convert “neutral speech into emotional speech”. Previous result has discussed on various methods [17] like GMM, CART, and Linear modification methods. The GMM method is much more suitable for a small training set, while CART give better output if trained in a large context balanced corpus.

V. Future Work

Our study is going on pitch parameter and LPC vocoder to convert neutral speech into emotional speech.

VI. Conclusion

This paper has described a perception experiment that was designed to make a classification of emotional speech. The classification results help us to achieve more acoustic patterns when synthesizing emotional speech.

Acknowledgment

The authors would like to thank Prof. J.S.Chitode for their advice, and they would also like to thank especially the students of Government Polytechnic, karad for their cooperation with the experiments.

References

[1] N. Campbell, "Perception of affect in speech—Toward an automatic processingof paralinguistic information in spoken conversation," in *Proc. ICSLP*, Jeju, Korea, Oct. 2004, pp. 881–884.

[2] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*.Cambridge, U.K.: Cambridge Univ. Press, 1988.

[3] J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, Eds.,*Progress in Speech Synthesis*. New York: Springer, 1997.

[4] J. Tao, "Emotion control of Chinese speech synthesis in natural environment,"in *Proc. Eurospeech*, 2003, pp. 2349–2352.

[5] Y. Xu and Q. E. Wang, "Pitch targets and their realization: Evidence from mandarin chinese," *Speech Commun.*, vol. 33, pp. 319–337, 2001.

[6] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentence of Japanese," *J. Acoust. Soc. Jpn. (E)*,vol. 5, no. 4, pp. 233–242, 1984.

[7] S. J. L. Mozziconacci and D. J. Hermes, "Expression of emotion andattitude through temporal speech variations," in *Proc. ICSLP*, Beijing,China, 2000, pp. 373–378.

[8] J. E. Cahn, "The generation of affect in synthesized speech," *J. Amer.Voice I/O Soc.*, vol. 8, pp. 1–19, Jul. 1990.

[9] Synthesis units for conversational speech—Using phrasal segments, N.Campbell. [Online]. Available: http://feast.atr.jp/nick/refs.html

[10] M. Schröder and S. Breuer, "XML representation languages as a way of interconnecting TTS modules," in *Proc. ICSLP*, Jeju, Korea,2004, pp.1889–1892.

[11] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, "A corpus-based approach to < ahem expressive speech synthesis," in *Proc. IEEE Speech Synthesis Workshop*, Santa Monica, CA, 2002, pp.79–84.

[12] I. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J.Acoust. Soc. Amer.*, pp. 1097–1108, 1993.

[13] R. M. Stibbard, "Vocal expression of emotions in non-laboratory speech:An investigation of the reading/leeds emotion in speech project annotation data," Ph.D. dissertation, Univ. Reading, Reading, U.K., 2001.

[14] E. Moulines and Y. Sagisaka, "Voice conversion: State of the art and perspectives," *Speech Commun.*, vol. 16, no. 2, pp. 125–126, Feb. 1995.

[15] S. McGilloway, R. Cowie, E. Doulas-Cowie, S. Gielen, M. Westerdijk,and S. Stroeve, "Approaching automatic recognition of emotion from voice: A rough benchmark," in *Proc. ISCA workshop Speech Emotion*,2000, pp. 207–212.

[16] Jianhua Tao, Yongguo Kang, and Aijun Li,"Prosody Conversion From Neutral Speech to Emotional Speech", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL14, NO. 4, JULY 2006

[17] Aditya Bihar Kandali#1, Aurobinda RoutrayTapan Kumar Basu," Vocal Emotion Recognition in Five Languages of Assam Using Features Based on MFCCs and EigenValues of Autocorrelation Matrix in Presence of Babble Noise",

[18] N. Amir, "Classifying emotions in speech: A comparison of methods,"in *Proc. Eurospeech*. Holon, Isreal, 2001, pp. 127–130.

[19] G. McLachlan and T. Krishnan, "The EM algorithm and extensions," in *Wiley Series in Probability and Statistics*, New York: Wiley, 1997.

[20] C. Gobl and A. N'1Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Commun.*, vol. 40, pp.189–212, 2003.

[21] R. Tato, R. Santos, R. Kompe, and J. M. Pardo, "Emotional space improves emotion recognition," in *Proc. ICSLP*, Denver, CO, Sep. 2002,pp. 2029–2032.

[22] V. A. Petrushin, "Emotion recognition in speech signal: Experimental study, development and application," in *Proc. ICSLP*, Beijing, China,2000, pp. 222–225.

[23] B. Hayes, *Metrical Stress Theory: Principles and CaseStudies*. Chicago, IL: Univ. Chicago Press, 1995.

[24] Z.-W. Shuang, Z.-X.Wang, Z.-H. Ling, and R.-H.Wang, "A novel voice conversion system based on codebook mapping with phoneme-tiedweighting," in *Proc. ICSLP*, Jeju, Korea, Oct. 2004, pp. 1197–1200.

[25] L. M. Arslan and D. Talkin, "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum," in *Proc. Eurospeech*,Rhodes, Greece, 1997, pp. 1347–1350.

[26] T. Watanabe *et al.*, "Transformation of spectral envelope for voiceconversion based on radial basis function networks," in *Proc. ICSLP*, Denver, CO, 2002, pp. 285–288.

[27] A. Li and H.Wang, "Friendly speech analysis and perception in standard chinese," in *Proc. ICSLP*, Jeju, Korea, 2004, pp. 897–900.

[28] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. ICASSP*, 2005, pp. 9–12.

[29] Y. Chen *et al.*, "Voice conversion with smoothedGMMand map adaptation,"in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 2413–2416.[30] Y. Stylianou *et al.*, "Continuous probabilistic transform for voice conversion,"*IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142,1998.

[31] HMizuno,HMizuno, andMAbe, "Voice conversion based on piecewise linear conversions rules of formant frequency and spectrum tilt," *Speech Commun. 16*, pp. 153–164.

[32] Y. Kang, Z. Shuang, J. Tao, W. Zhang, and B. Xu, "A hybrid gmm and codebook mapping method for spectral conversion," in *Proc. 1st Int.Conf. Affective Comput. Intell. Interaction*, 2005, pp. 303–310.