African Journal of Mathematics and Computer Science Research Vol. 3(10), pp. 255-266, October 2010
Available online at http://www.academicjournals.org/AJMCSR
ISSN 2006-9731 ©2010 Academic Journals

Full Length Research Paper

Transformation of emotion based on acoustic features of intonation patterns for Hindi speech

S. S. Agrawal[1], Nupur Prakash[2] and Anurag Jain[2]*

[1]CDAC (Centre for Development of Advance Computing), Noida KIIT (Kalinga Institute of Industrial Technology), Gurgaon, India.
[2]School of I.T. (Information Technologies), (GGSIP) Guru Gobind Singh Indraprastha University, Delhi, India.

Accepted 29 September, 2010

Changes in intonation patterns may convey not only different meaning but different emotions even if the sequence of speech segments are same in a sentence. The patterns change depending upon structure and emotion of the sentence and require being stored in speech database. It is a difficult and time-consuming task to store all utterances of all the expressive style, which also consumes huge memory space. So there should be an approach that minimizes the time and memory space for emotion rich database. A number of studies in this respect have been done for several languages and models developed. However, for Hindi not many studies have been done. Taking this fact in consideration the intonation patterns have been studied for different languages in this paper and analysed for Hindi language. On the basis of dense research on intonation pattern an algorithm has been proposed for emotion conversion. This algorithm only requires storing neutral utterances in the database and other expressive style utterances can be derived from these neutral emotion. Proposed algorithm is based on linear modification model (LMM), where fundamental frequency (F0) is one of the factors to convert emotions. To perform the experiments, an intonational rich database is maintained for four expressive styles; surprise, happiness, anger and sadness. The perception tests also carried out, where group of listeners were asked to listen to the utterances from database and judge the emotion. This perception test involves classification of the emotions already available in the database by the listener and to judge the quality of converted neutral utterances. The results are analysed for four emotions: happiness, anger, surprise and sadness and performance of the experiment is evaluated. The accuracy of perception test on transformed emotions was found out to be 95% for surprise and 93.4% for sadness 82% for happiness and 96.7% for anger.

Key words: Intonation patterns, intonational database, emotion conversion, fundamental frequency (F0), perception test.

INTRODUCTION

Speech is an important medium not only to convey strictly linguistic content of sentences but also the expressions of attitudes and emotions of the speaker. Pith level, pitch contour, F0 and intensity are said to be important prosodic cues. Intonation pattern is one of the determinants of the emotion conveyed in speech. Mozziconacci et al. (1999) suggests that in the production study, no clear-cut and one to one relationship between intended emotion and intonation pattern were found. But some patterns can occur most often in all emotions. All these experiments have been performed on the Dutch utterances.

These acoustic parameters are useful for correlating the behavior of the person. Several studies have been carried out for interpreting emotions in different languages.

Daniel (1998) has also explained about the Intonation Patterns in detail and connected with prosodic structure of the language.

Abelin et al. (2000) suggests that the degree of stability

*Corresponding author. E-mail: anuragjain76@gmail.com.

in the way different emotions and attitudes are interpreted by the use of prosodic patterns. It also depends on listener's culture and linguistics background of the speakers. They also analyzed the re-occurring relation between acoustic and semantic properties of the stimuli. Conclusion of the same tells that few emotions are interpreted in great degree with the intended emotion.

In other research made by Mozziconacci et al. (2001) argued that production studies are optimally supplemented with perception studies. He also argued that interpreting speech data in the framework of a model of intonation provides a methodological background of intonation phenomena relevant to speech communication.

F0 curve for some of the emotions expressed in Turkish utterances are studied by Meral et al. (1999) and concluded that the most correctly recognized emotion is anger and least recognized emotion is happiness and also there is learning process in interpretation of emotions.

Montero et al. (1999a) classified the emotions in segmental emotion and prosodic emotion and confirm it with the help of preliminary test of concatenate synthesis using only automatic emotional prosody.

Experiments on emotional speech database for Spanish are performed by Montero et al. (1999b) and generated recognition rates to identify emotions for copy synthesis and automatic prosody experiment. They suggested that sad and neutral sentences are the easiest sentences to recognize due to shorter F0.

The degree of stability in the way different emotions and attitudes are interpreted by the use of prosodic patterns, has been studied in Abelin et al. (2000). Emotion also depends on listener's culture and linguistics background of the speakers and it is a very cumbersome process to create a large corpus to develop a good quality speech synthesizer. Due to this reason, ability to add emotions by conversion to synthesized speech corpora is in high demand.

Different models have been proposed for emotion conversion like linear modification model (LMM), Guassian mixture model (GMM) and Classification and Regression Tree (CART). The LMM makes direct modification of F0 contours, syllabic durations, and intensities from the acoustic distribution analysis results. F0 contours contains F0 top, F0 bottom, and F0 mean. Twelve patterns (four emotions with three degrees, "strong," "medium," and "weak") have been deduced from the training set of the corpus by Jianhua et al. (2006). Further analysis shows that the expression of emotion does not just influence this general prosody features, but also affects the sentence stress and more subtle prosodic features.

More focus is made on emotion conversion for spoken English with the help of different techniques using GMM (Inaloglu and Young, 2007a), Regression Tree (Tooher et al., 2008) and Codebook approach (Inaloglu and Young, 2007b), however not much work has been reported for Hindi corpora. The proposed work is based on LMM model where we analyze the natural source utterance and natural target utterance (on training set) are analyzed and on the basis of the prosodic differences obtained between two emotional natural utterances. The source utterance is algorithmically modified and made them as target utterances.

The experiments have been performed on Hindi speech corpora based on intonation patterns for different kinds of emotions like neutral, surprise, anger, happiness and sadness.

SPEECH DATABASE

For performing experiments and verification of results, an emotionally rich database was prepared. For the data-base, ten native speakers were given 25 sentences to generate Hindi utterances in five expressive styles neutral, sadness, anger, surprise and happy. These sentences are emotionally rich and can be spoken in all five emotions. Once the speaker is in emotionally charged mood, each speaker was asked to record the sentences with full emotion at 44.1 KHz sampling rate and 16-bit precision with Mono channel and stored in the computer.

There is another database which is directly associated with the main module of emotion conversion. The database is used to keep the pitch point values for the utterances, already present in the Speech Database. Six to eleven pitch points have been computed for all utterances. The numbers of pitch points are based on number of syllables present in the sentence and resolution frequency (fr). If sentence has less number of syllables, less pitch points may be there and with the comparison of resolution frequency and pitch points, pitch points can be further reduced to some reasonable quantity.

Resolution frequency is the minimum amount by which every remaining pitch point lies above or below the line that connects two neighbors pitch points. The commonly available PRAAT analysis software tool is used for this purpose. Nearly two hundred utterances of neutral speech were carried out for training and rest of the utterances was kept reserve for testing. Various expressive style utterances are considered for choosing the listeners for perception test. Forty listeners were appeared for the perception test and 100 random utterances were given to them. Only 20 were selected on the basis of their perception towards the utterances.

F0 BASED INTONATION PATTERN OF HINDI

The changes in the F0 patterns in the initial and final positions were carefully studied by Jain et al. (2008). The maximum and minimum F0 value along with the standard deviation and average F0 values and mean slope have also been computed. The effects of the emotions on F0 curve were studied in respect of rise, fall and hold pattern.

Happiness

From the observation of F0 curve for utterances of happiness it was found that the hold pattern appears at the end of the sentences and rise and fall pattern appears at the beginning of the sentences (Figure 1). For some cases fall and rise patterns were also found at the initial position of the sentence.

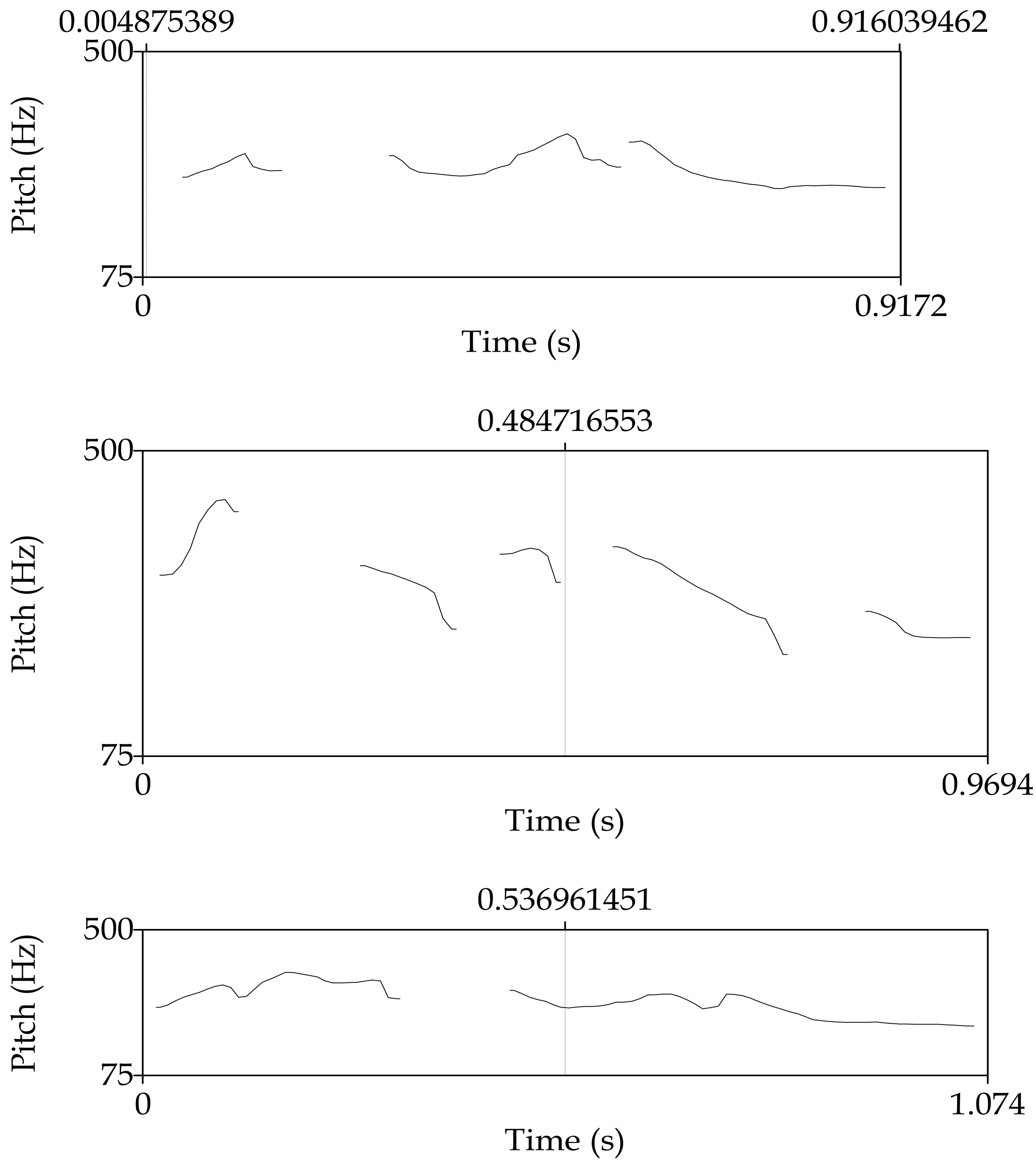

Figure 1. F0 patterns for happiness state.

Anger

The trend of F0 curve in the utterances of anger is towards fall at the end of the sentences and towards rise - fall in the beginning of the sentences. In some cases fall - rise pattern is observed in the beginning of sentences (Figure 2).

Sadness

Through the analysis of the F0 contour of utterances of sadness that the rend is towards fall or hold at the end of sentences and towards fall and rise in the beginning of the sentences (Figure 3).

In our experiment, some utterances are also demonstrating slightly hold pattern in the beginning of utterance. Overall it is observed that F0 curve have only fall-fall trend throughout the pattern.

Normal

F0 curve of normal utterances, falls at the end of the utterances and the trend is towards rise and fall in the beginning of the sentences. We also got few hold, fall and rise patterns in the beginning of sentences (Figure 4), but occurrence of these patterns are very less. In most of the cases fall is observed towards the end of

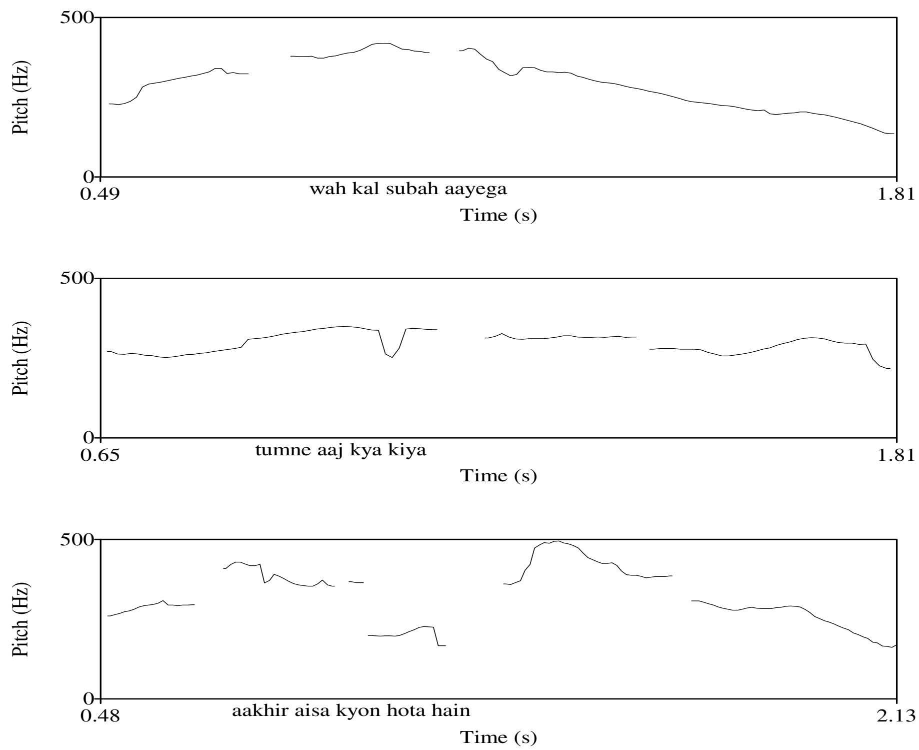

Figure 2. F0 patterns for Anger state.

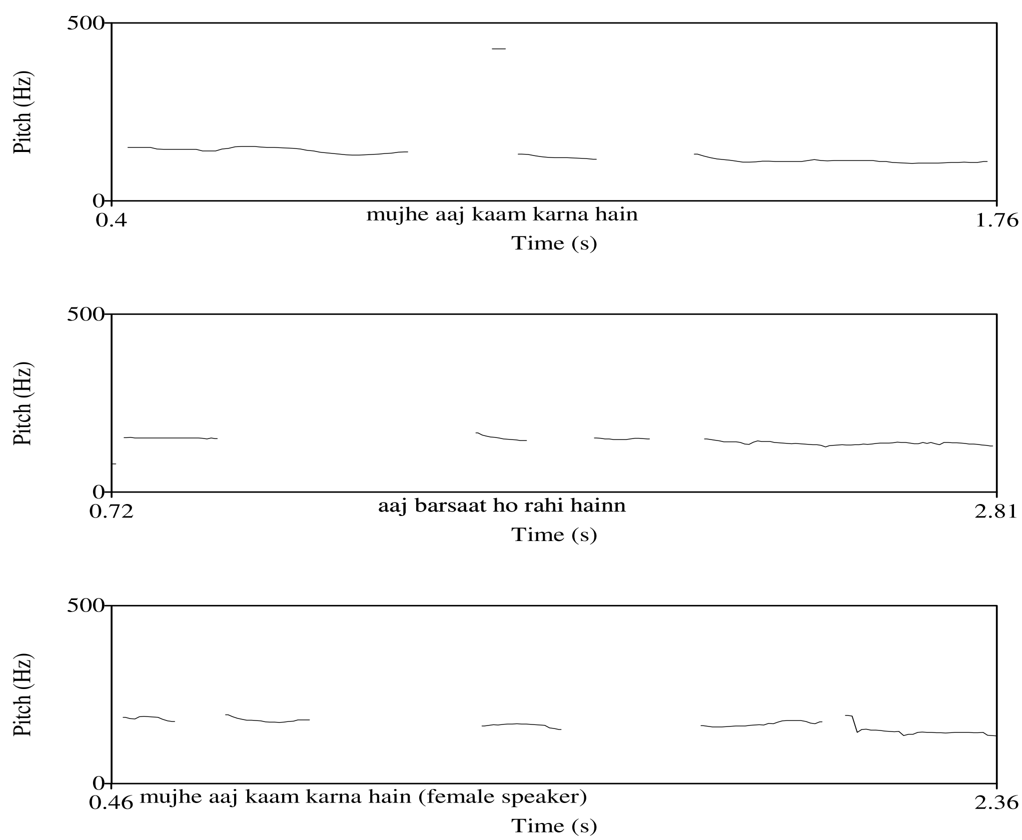

Figure 3. F0 patterns for sadness state.

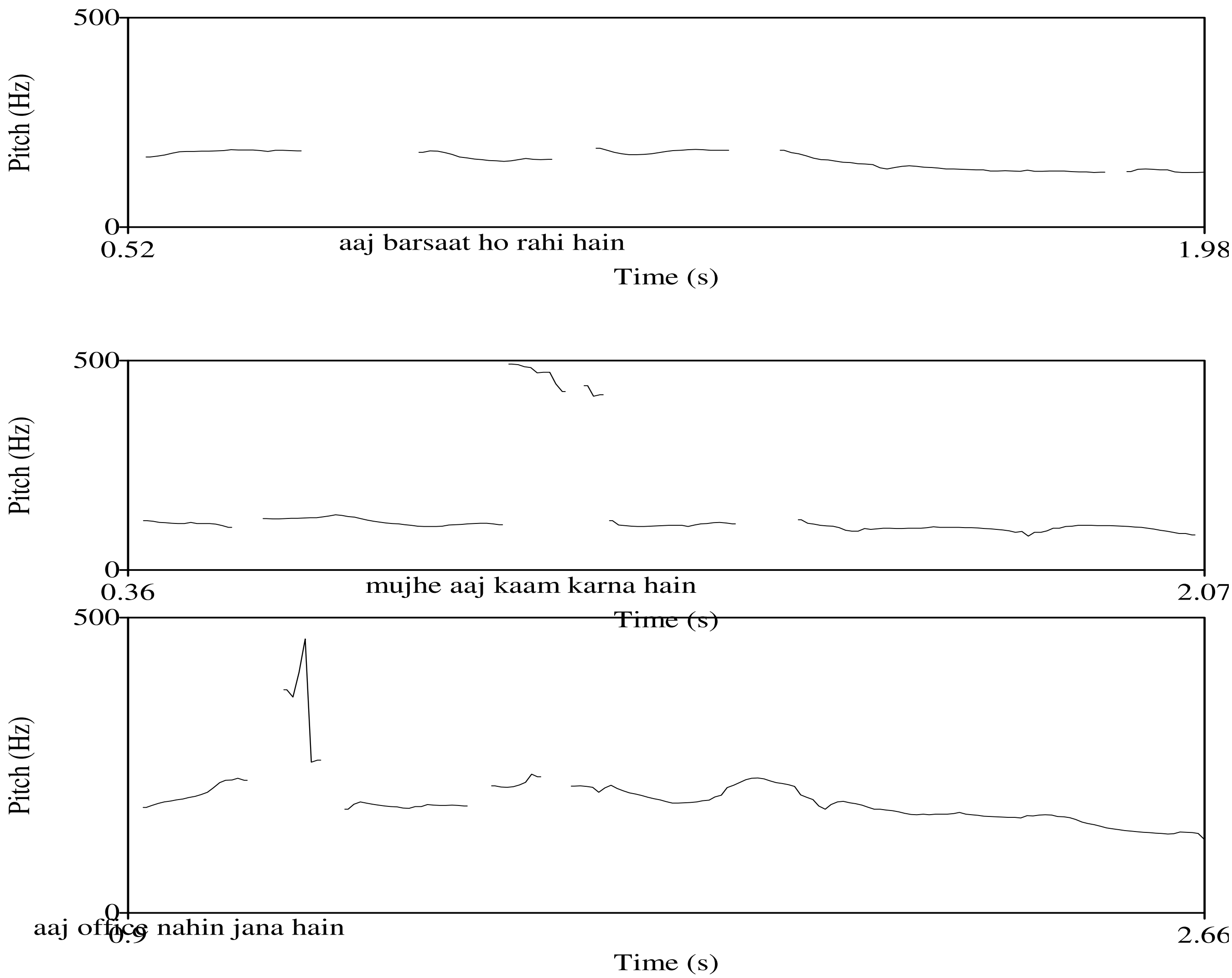

Figure 4. F0 patterns for normal state.

sentence irrespective of the speaker.

Surprise

For surprise emotion for F0 curve it is observed to have rise and fall pattern in the beginning of the sentence and rise pattern towards the end of the sentence (Figure 5). But for some cases hold pattern is also observed towards the end of the sentence. In our experiment, most of the utterances of surprise emotion have evolved in the form of question based surprise state.

Statistical calculation

Statistical analysis was done for F0 curves generated using five emotions generated from different speakers and depicted in (Table 1). Based on generated values from the experiment, surprise has the highest average F0 value, while the normal state has the lowest one. Pitch variation is lowest for normal state and highest for surprise state. Standard deviation is almost same and lowest for sadness and normal state while highest for surprise state.

F0 BASED EMOTION CONVERSION

Before proposing the algorithm for emotion conversion from neutral emotion to target emotion, a speech corpus was analyzed on the basis of F0 patterns.

Two methods are proposed for the desired emotion conversion.

(i) Emotion conversion at sentence level.
(ii) Emotion conversion at word level.

In these methods pitch points (Pi) were studied for the desired source emotion (Neutral) and target emotion and then the difference between corresponding pitch points were evaluated after normalization. This serves as an indicator of the values by which, pitch points of source speech utterance must be increased or decreased to convert it to target utterance**.** When a sound file is selected, pitch analysis is performed, using timestamp and minimum and maximum pitch parameters. The information of the resulting pitch contour is used to posit glottal pulses where the original sound contains much energy and then pitch contour is converted to a pitch tier with many points. For pitch analysis step length is taken as 0.01 s and minimum and maximum pitch is taken

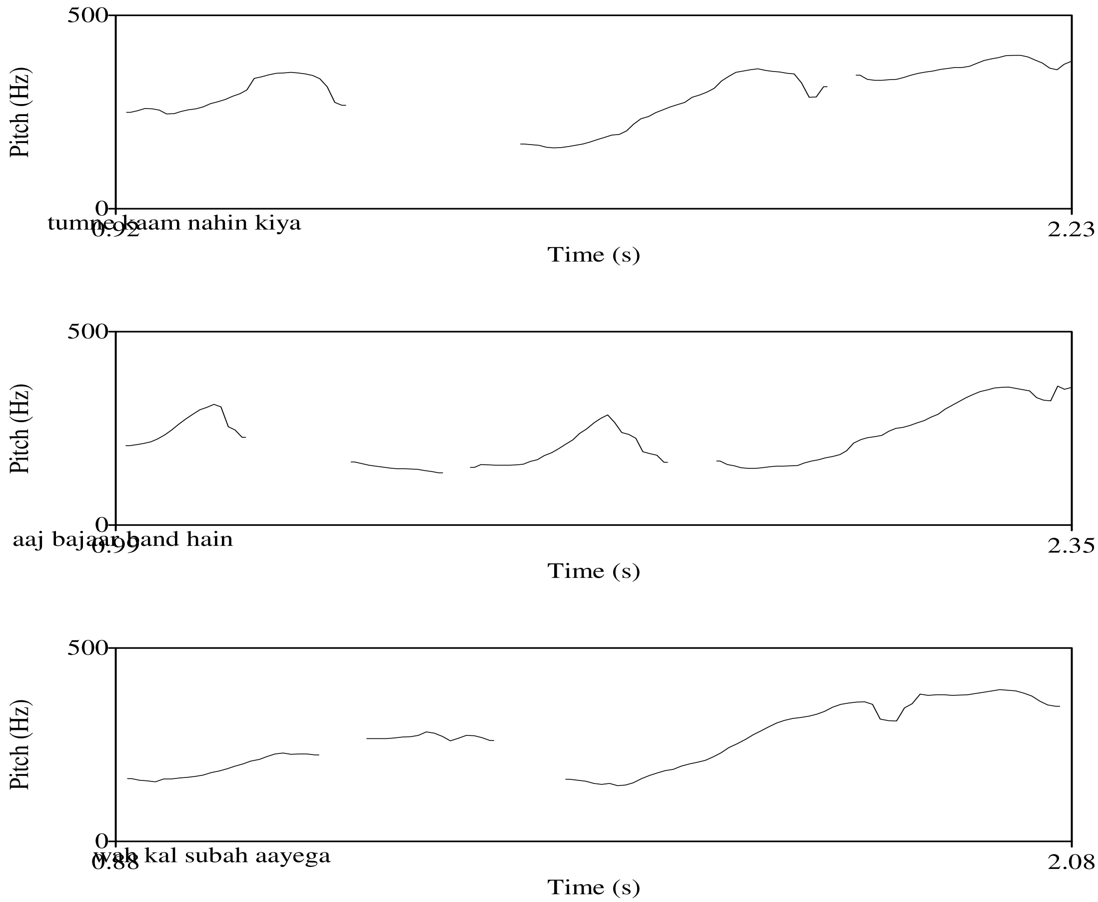

Figure 5. F0 patterns for surprise state.

Table 1. F0 patterns for happiness state.

F0	Happiness	Anger	Sadness	Normal	Surprise
Maximum	474.04	405.82	495.94	463.35	495.17
Minimum	220.91	231.44	194.71	179.10	198.83
Average	290.31	296.3	278.72	270.96	335.91
Standard deviation	55.48	54.49	82.81	81.27	86.58
Pitch Variation	253.12	174.38	301.25	284.25	296.34
Mean absolute Slope (Hz/Sec)	543.6	506.2	530.68	493.2	490.3

as 75 and 500 Hz, respectively. Then stylization process is performed to remove the excess pitch points and then valid numbers of pitch points were noted in the database. After comparison between source and target emotion training set, pitch points are divided into four groups and the initial frequency is set as x_1, x_2, x_3, and x_4 respectively. On the basis of observation of training set y_1, y_2, y_3 and y_4 is added to the subsequent "x" values. In some cases, Pitch point number also matters and is important to decide the transformed F0 value, on which the algorithm is implemented.

All x_i and y_i are in Hz. To select y_i values, Difference table is constructed. On the basis of y_i values given in difference table, those y_i values are taken in consideration on which maximum number of utterances. So for performing the experiment, we have not calculated the mean of x_i and y_i, these values are generated after the rigorous analysis of pitch patterns of neutral and emotional utterances. To construct Difference table, only 8 pitch points are considered for Normal - Surprise emotion conversion. Since the different table is too big to fit in the paper so only 4 pitch points are mentioned in Table 2.

Sentence based emotion conversion

The proposed algorithm for Emotion conversion at sentence level is given below.

Select desired sound wave form
Convert speech waveform in pitch tier
// Stylization
For all P_is

Table 2. Difference table for normal-happiness emotion conversion at sentence level.

Pitch point 1	Range difference (y) (Hz)	-60	+80	+200
	Utterance Frequency	83%	15%	2%
Pitch point 2	Range difference (y) (Hz)	-40	+40	>+100
	Utterance Frequency	86%	8%	6%
Pitch point 3	Range difference (y) (Hz)	-20	+25	>+80
	Utterance Frequency	5%	87%	8%
Pitch point 4	Range difference (y) (Hz)	-10	+60	>+80
	Utterance Frequency	17%	70%	13%

Select Pi, that is more close to straight line and compare with resolution frequency (fr)
if distance between pi and straight line > fr
Stop the process
else
Repeat for other P_is
Divide the pitch points in four groups
For each group
group[i] = x_i+y_i || x_i-y_i
Remove existing pitch points
Add newly calculated pitch points in place of old pitch points.

Algorithm 1. Sentence level transformation

Table 3 depicts the list of pitch points involved (after normalization) in different emotions for a particular training set utterance for a Hindi sentence. Figures 6 and 7 elucidate the variation in the pitch points involved in natural neutral emotion and natural surprise emotion respectively.

Emotion conversion at word level

In order to derive a formula for general emotion conversion, the modified version of the "algorithm 1" was implemented on every word of the utterance, under the same circumstances as that of the previous algorithm.

The modified algorithm is as follows:

Select desired sound wave form
Convert speech waveform in pitch tier
Apply Stylization
Divide the word's pitch points in groups
For each group
*Group [i] = x_i+y_i || x_i-y_i + k*C \\ for some points constant factor 'C' is to be added with multiple of fifty as approximation. k = +1 or -1*
Remove existing pitch points
Add newly calculated pitch points in place of old pitch points.

Algorithm 2. Word level transformation

Table 4 depicts the list of pitch points involved (after normalization) in different emotions for a particular training set utterance "Kal bazaar jana hain". Figures 8 and 9 elucidate the pitch points involved in natural neutral emotion and natural surprise emotion respectively.

RESULTS AND EVALUATION

The proposed algorithms were evaluated through experiments carried out on PRAAT software. The perception test is carried out for validation of results.

Experimental results

For this process, predefined speech training set was chosen and randomly neutral utterance is selected. For example, *"Kal tumhe phansi ho jayegi"* was considered

Table 3. Pitch points (in Hz).

Pitch points	Neutral	Surprise	Happiness	Sadness	Anger
Pt1.	276.4	349.5	162	326.4	448.6
Pt2.	319.7	389	357.5	302.6	452.6
Pt3.	205.7	244.5	195.2	471.3	273
Pt4.	211.3	217.9	420.3	115.6	348
Pt5.	331.7	255.6	252.5	118.5	447.9
Pt6.	262.3	414.1	484.3	242.2	435
Pt7.	246.9	261.7	143.2	343.7	399
Pt8.	141.8	492.1	188.9	207.3	379.2
Pt9.	171.3	324.2	107.4	141.9	229
Pt10	125.4	375.9	220.2	279.9	226.6
Pt11	205.2	399.2	264.9	280.1	261.6

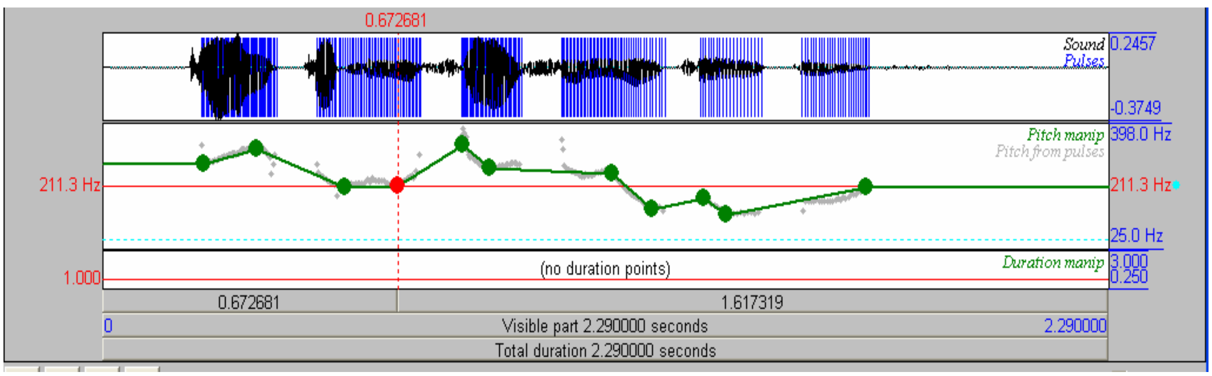

Figure 6. Pitch points for natural normal emotion.

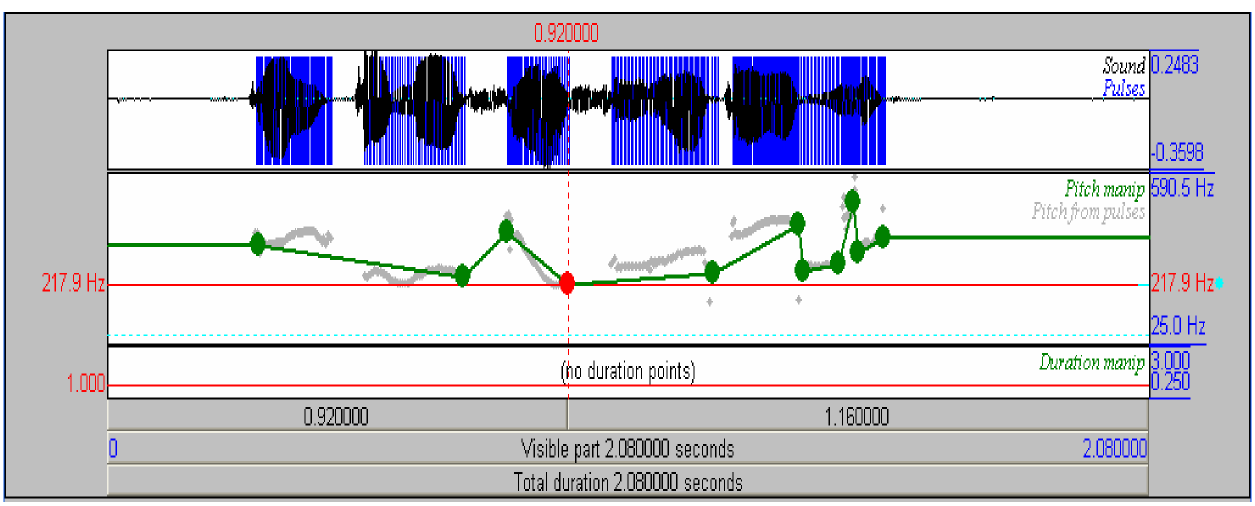

Figure 7. Pitch points for natural surprise emotion.

and the results are shown in Figure 10 and Table 5. In Figure 10, upper picture shows the natural surprise utterance and lower picture displays the transformed surprise utterance. Table 5 elucidates the conversion algorithm pitch points wise. The differences between transformed surprise and sadness pitch point values

Table 4. Pitch points comparison.

Word	Pitch points normal (Hz)	Pitch points surprise (Hz)	Pitch points happiness (Hz)	Pitch points anger (Hz)	Pitch points sadness (Hz)
1. Kal	234.6	267.2	259.9	364.5	278
	319.2	394.9	388.8	457.1	304.2
2. Bazaar	205.6	425.2	184.1	319.6	263.3
	185.7	490.8	265.8	276.9	242.3
	111.6	255.8	398.8	378.7	299.8
	117.4		464.8	437.8	
3. Jana	279.3	275.8	368.6	490.8	307
	199.6	209.1	491.8	253.1	299.4
			253.3	403.9	
4. Hain	231.1	390	275.4	361.2	111
	87.6	294.3	165.1	269.1	123.4
		317	238.3		

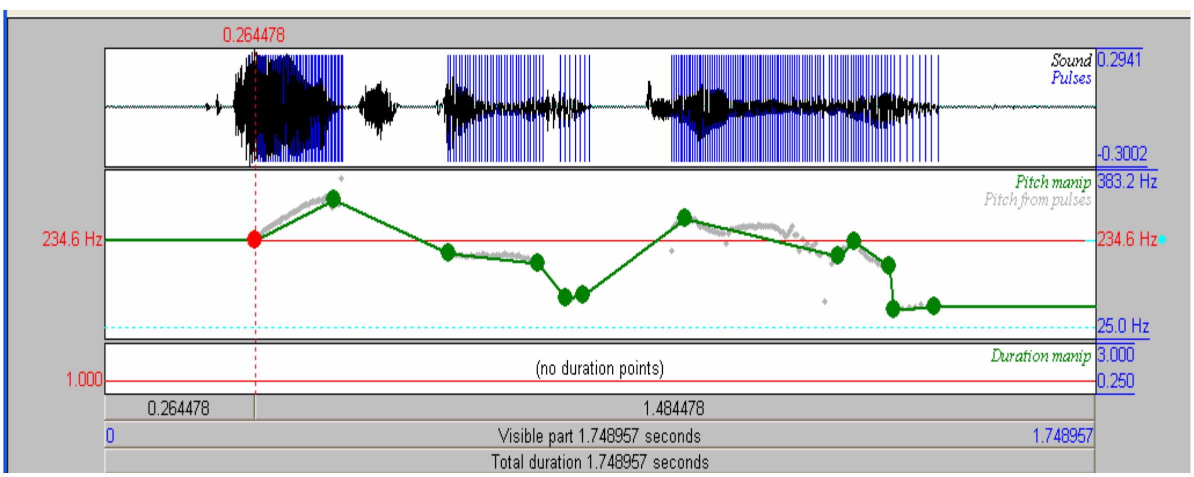

Figure 8. Pitch points for natural normal emotion (word level).

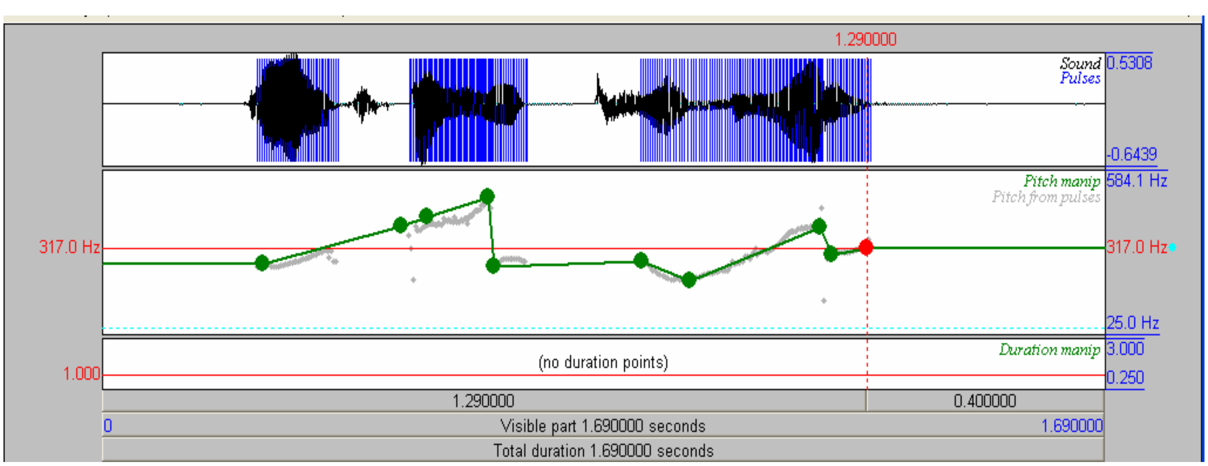

Figure 9. Pitch points for natural surprise emotion (word level).

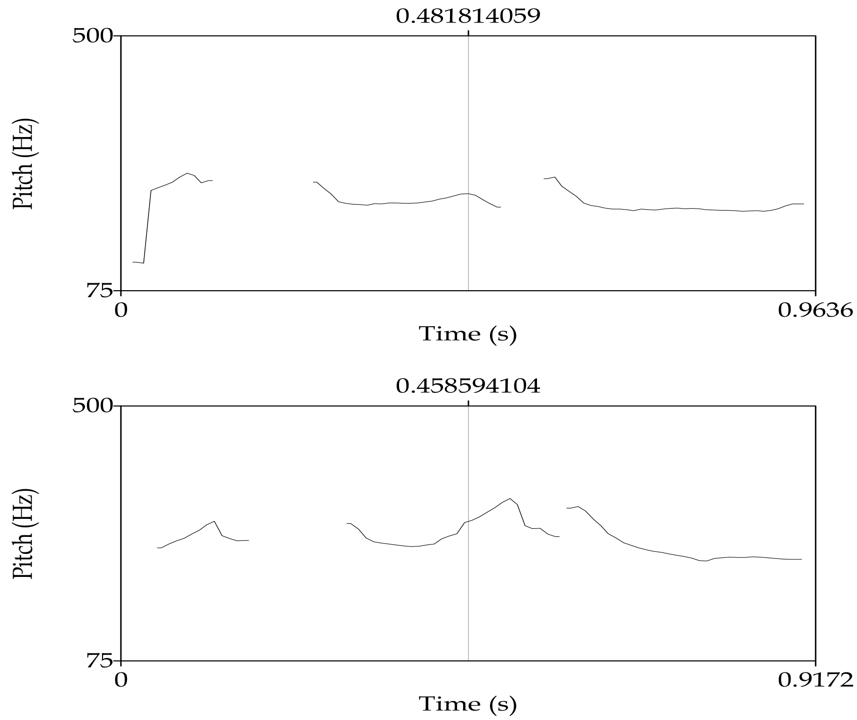

Figure 10. Natural and transformed surprise emotion utterance.

Table 5. Comparison of emotions after transformation for "*Kal tumhe phansi ho jayegi*".

Pitch points	Transformed surprise utterance (Hz)	Transformed happiness utterance (Hz)	Transformed sadness utterance (Hz)	Transformed anger utterance (Hz)
1	342	190	326.1	448.6
2	389.7	293.5	312.8	449.7
3	275	170.9	440.3	258.8
4	249.3	411.3	150.5	339.5
5	463.5	402.5	139.9	380.6
6	275.9	422.3	230.0	456.8
7	317.3	280.5	343.7	399
8	470.3	201.5	216.5	379.9
9	377.8	233.3	150.4	267.8
10	386.7	170.3	279.9	227.9
11	452	265.4	280.1	290.1

have been shown. In this depiction, only the intonation has been taken into consideration. Other factors like stress, duration and syllable position within the word are not considered. Although after carefully listening, the listeners are very much convinced with the transformed surprise and sadness emotion from neutral emotion. It is easily verified in perception test.

Perception test

In this part of the experiment, utterances of different emotions are presented to the listeners. The listeners are asked to identify the emotions of those utterances. The listeners are divided into 3 groups of 5 candidates each. Table 6 presents the perception matrix on the original natural emotional data based on the F0 analysis of intonation patterns of Hindi. The matrix row represents expressed emotions and columns represent perceived emotions. Values are mentioned in percentage. As seen by perception matrix, "anger" is most accurately perceived emotion and least accurately perceived emotion is "happiness". Table does not depict 100% accuracy for each expressed emotion due to listener's confusion about

Table 6. Perception table (values are in %).

Emotion	Happiness	Anger	Sadness	Normal	Surprise
Happiness	64.2	0	0	0	4.32
Anger	2.8	96.3	0	2.12	8.92
Sadness	2.3	1.43	87.3	31.52	1.32
Normal	25.2	0	11.45	66.82	3.54
Surprise	5.6	3.12	1.36	1.89	82.36

Table 7. Transformed perception matrix.

Emotion	Happiness	Anger	Sadness	Surprise	Normal
Happiness	82	0	0	9.2	0
Anger	0	97.6	0	6	0
Sadness	0	0	93.4	0	7.1
Surprise	2.5	5.8	0	95	0

the emotion. For some emotions it exceeds 100% and for some emotions it is below than 100%. It is because some times listener perceives more than one emotion for the same expressed emotion and sometimes he/she was unable to perceive the emotion under the mentioned category of emotions and assumes it to be 0. This phenomenon is due to overlapped features of Pitch range and variation for all the emotions. It is concluded that no emotions has one to one mapping with particular pitch variation and other acoustic parameters. In the next stage the listeners were asked to predict the emotions of system generated transformed emotions to verify the performance of the system. Table 7 presents the perception matrix for the transformed emotional data from original neutral emotion speech. The listeners were given full liberty to predict any emotion. It can be analyzed with Tables 6 and 7 that transformed emotional states has provided good results as compared to natural emotional states and from the results of emotion transformation, we can proof the perfectness of methodology proposed in this paper.

CONCLUSION

An algorithm for emotion conversion has been described which consists of simple and experienced rules for converting F0 parameter and energy contour. In this paper the alignment of pitch points by linguistic rules has not been considered, the future word will focus on the linguistic rules for emotion conversion. Perception test verifies the performance of the algorithm. But the proposed system has scope for further refinements. The F0 and Energy factor have been considered for our experiment; the effect of other factor like Spectrum, Duration, Syllable information etc can be further investigated. The experiment has been performed on 800 utterances and not adequate in terms of numbers. The database should be enhanced to achieve the perfect ness. Since few deviations have been made for Hindi from other languages and can be easily verified with the results of intonation pattern for different emotions hence, it is justified to design Hindi based intonational model where transformation of emotions can be incorporated.

ACKNOWLEDGEMENTS

This research was supported by Guru Gobind Singh Indraprastha University Delhi and Special thanks are conveyed to colleagues and post graduates students who participated in the perception test and speech recording.

REFERENCES

Abelin A, Allwood J (2000). Cross Linguistic Interpretation of Emotional Prosody. Proceedings of the ISCA Workshop on Speech and Emotion.

Daniel H (1998). “Intonation in British English A book chapter of Intonation Systems- A survey of Twenty Languages. Cambridge University Press ISBN-10:052139550X.

Jain A, Prakash N, Agrawal SS (2008). Acoustical and Perceptual study of Intonational Patterns in emotional Hindi Speech. in Proc. O-COCOSDA, Japan.

Jianhua Tao, Yongguo Kang, Aijun Li (2006). Prosody Conversion From Neutral Speech to Emotional Speech. IEEE transactions on Audio, Speech, and Language Processing, 14(4).

Tooher M, Irena Y, Christer G (2008). Transformation of LF Parameters for Speech Synthesis of Emotion: Regression Trees. Speech Prosody pp. 705-708, Campinas, Brazil.

Inaloglu Z, Young S (2007a). A System for Transforming in speech: combining Data-Driven conversion techniques for Prosody and Voice quality. in proc. of INTERSPEECH.

Inanoglu Zeynep, Young S (2007b). Data-driven emotion conversion in spoken English”. Speech Commun., 51: 268-283. http://www.fon.hum.uva.nl/praat (date last viewed 05/08/2010).

Montero JM, Gutierrez-Arriola J, Colas J, Enriquez E, Pardo JM (1999a). Analysis and Modeling of Emotional Speech in Spanish.

ICPhS, pp. 957-960.
Montero JM, Gutierrez-Arriola J, Colas J, Enriquez E, Pardo JM (1999b). Development of an emotional speech synthesizer in Spanish. Eurospeech.
Mozziconacci SJ, Hermes DJ (1999). Role of Intonation Patterns in Conveying Emotion in Speech. ICPhS. pp. 2001-2004.
Mozziconacci SJ (2001). Emotion and attitude conveyed in speech by means of prosody. 2nd workshop on Attitude, Personality and emotions in User-Adopted Interaction, Sonthofen, Germany.