IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 2, February 2016.
www.ijiset.com
ISSN 2348 – 7968

Prosody Conversion from Neutral speech to Emotional speech

Ms. Snehal S.Patil[1], Prof. Dr. Mrs. Shaila D. Apte [2]

[1] M.E. Student, Rajarshi Shahu College of Engineering, Tathawade
Pune, Maharashtra, India

[2] Professor, Rajarshi Shahu College of Engineering, Tathawade
Pune, Maharashtra, India

Abstract

There is increase in interest on emotional speech synthesis. Emotion conversion aims at transforming the emotions expressed by voice, which is usually neutral, in to target emotion. It is well known that prosody coveys a great amount of emotional charge. So the emotion conversion system is voice conversion system with special focus on prosody level. In neutral speech, prosodic features can be manipulated to express different emotions. Analysis-Synthesis methods can be used to generate the emotional speech. Pitch modification techniques with LPC and PSOLA synthesis are used to get emotional speech. A speech data base is locally generated for ten different speakers. 20 sentences with neutral emotion as well as with different emotions i.e. questions and exclamatory are recorded. Pitch analysis and pitch modification has been done on these sentences. LPC and PSOLA synthesis are applied to get emotional speech. Thus LPC and PSOLA based Synthesis method is used with Pitch modification to convert the neutral speech signal into question and exclamatory voice. Results for both the synthesis techniques were compared. It is observed that LPC synthesis gives better MOS of 4.7 and intelligibility index of 70%.
.

Keywords: *prosody, LPC Synthesis, PSOLA Synthesis speech analysis, Pitch Analysis, Pitch modification.*

1. Introduction

Prosody is one of the key components of Speech Synthesizers, which allows implementing complex weave of physical, phonetic effects that is being employed to express attitude, assumptions, and attention as a parallel channel in our daily speech communication. In general any communication is collection of two phases: Denotation, which represents written content or spoken content and Connotation, which represent emotional and attention effects intended by the speaker or inferred by a listener. Prosody plays important role in guiding listener for speaker attitude towards the message, towards the listener and towards the complete communication event.

From listener point of view, prosody consists of systematic perception and recovery of speaker intentions based on:

a)Pauses: To indicate phrases and separate the two words

b) Pitch: Rate of vocal fold cycle as function of time

c) Rate: Phoneme duration and time

d) Loudness: Relative amplitude or volume.

Prosody is one of the most important components of human spoken communication. A correct prosody in a text-to-speech system contributes to a better quality in terms of intelligibility, naturalness and pleasantness. It carries different kinds of information. Some of them are strongly related with the text itself (e.g. syntactic structure, semantic disambiguation) but a significant part is not (e.g. emotional, intention). Then, it is not possible to generate such prosody capable to transmit them only from text. Most widely used speech synthesis applications assume a reading style. Spoken communication is different from written communication. When a human writes a text he is aware that all information will be contained in words themselves. However, when speaking, humans use additional acoustic cues in order to transmit more information. Prosody is thus even more relevant in spoken communication than in written text. Therefore, a greater effort must be done in prosody modeling when dealing with spoken rather than reading style speech synthesis.

2. Literature Review

Over the years, considerable research has been done on prosody conversion. Literature survey has been carried out for different approaches of speech synthesis include Gaussian Mixture Models (GMM), Linear Mixture Model (LMM), Classification and Regression Tree (CART) and Hidden Markov Model (HMM). Simplified Inverse Filter Tracking (SIFT) algorithm is also used for pitch extraction.

3. Problem Definition

In neutral speech, prosodic features can be manipulated to express different emotions. Analysis-Synthesis methods can be used to generate the emotional speech.

ISSN 2348 – 7968

Pitch modification technique with LPC and PSOLA synthesis are used to get emotional speech. A speech data base is locally generated for ten different speakers. 20 sentences with neutral emotion as well as with different emotions i.e. questions and exclamatory are recorded. Pitch analysis and pitch modification has been done on these sentences. LPC and PSOLA synthesis are applied to get emotional speech.

Thus LPC and PSOLA based Synthesis method is used with Pitch modification to convert the neutral speech signal into question and exclamatory voice.
Results for both the synthesis techniques were compared.

4. System Implementation and results

Below is the block diagram explaining the process of system implementation.

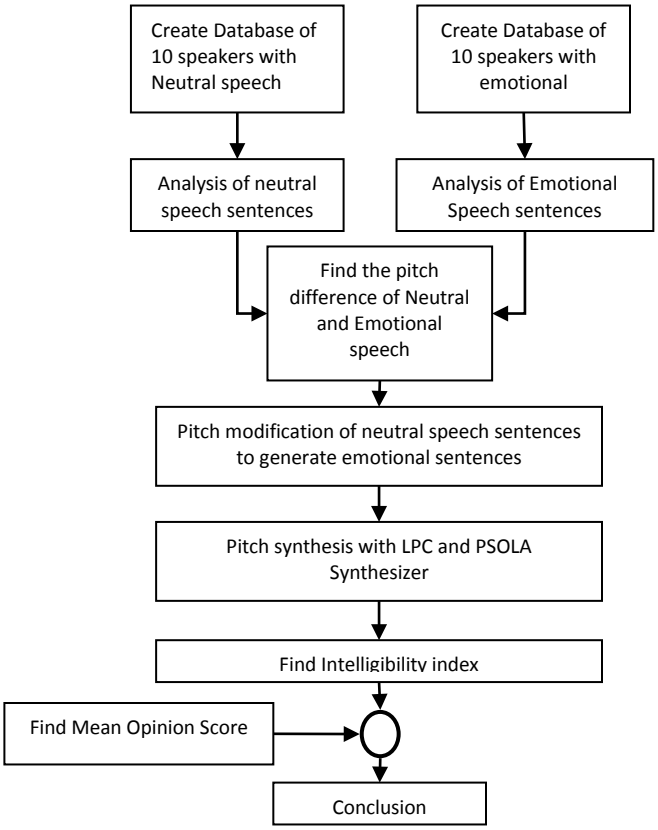

Fig. 1 Algorithm of system implementation.

4.1 Database Preparation

A database consist neutral sentences and emotional sentences. In emotional sentences there are questions and exclamatory sentences. There are around 400 sentences of 10 speakers have been generated. The first step was to list out the sentences of exclamatory and questions followed the recording of each sentences with neutral and with desired emotions of all the speakers. The recordings of sentences were done using the software Audacity Sound Recorder. These words and sentences were recorded at a sample rate of 44100 Hz with a single (Mono) channel. The recordings were done in a silent room with minimal noise. For pitch analysis minimum and maximum pitch is taken as 75 Hz and 500 Hz.
Speakers recorded the words and sentences, with each word and sentence being spoken in two different prosodies like question mark and exclamatory and then optimize it. These words and sentences are emotionally rich and can be spoken in all four other emotions. Once the speaker is in emotionally charged mood, each speaker was asked to record the words and sentences with full emotion and stored in the computer. As a part of the optimization process, the files were subject to three steps:

1) Normalization
2) Noise Reduction

To normalize is to adjust the volume so that the loudest peak is equal to the maximum signal that can be used in digital audio. Two methods were used for mitigating the background noise from the sound one is the manual spectral subtraction was done on each sound file by hand picking the noise regions and the second is the sound was subject to a noise gate to clean up the sound. As a last part of the optimization process, the files were trimmed so that silent regions at the beginning and end of each utterance were eliminated. The resulting database had an average duration for each sentence utterance.

4.2 Speech Analysis

In our experiment we extracted all the pitch values of each and sentences recorded in normal and emotional speech using MATLAB programme and PRAAT. Each recorded sentence has three parts i.e. starting, middle and end using the fact that the pitch values should be continuous. Then we extracted the pitch values sentence from starting and ending point as well as maximum and minimum pitch value between starting and end points of particular sentence. The difference of original emotional and neutral speech has been found..
The database constructed was used to study and analyze patterns and similarities in the pitch contour when one sentence was spoken in exclamatory emotion, as compared to when the same sentence was spoken in the neutral emotion. For the purpose of the analysis, pitch points were computed for the utterances and a pattern was found. An algorithm for the conversion from neutral state to

ISSN 2348 – 7968

emotional state was implemented. For computing the pitch points, the sound files were loaded in PRAAT for extracting the pitch values. The whole process was repeated for all sound files. For the purpose of analysis, each utterance was partitioned into three sections Start, Mid and End part of the sentence. Also, in each part, the pitch points pertaining to the maximum pitch and the minimum pitch were computed. Fig.2 represents waveform comparison of neutral speech and exclamatory speech. Fig.3 represents waveform comparison of neutral speech and question mark speech.

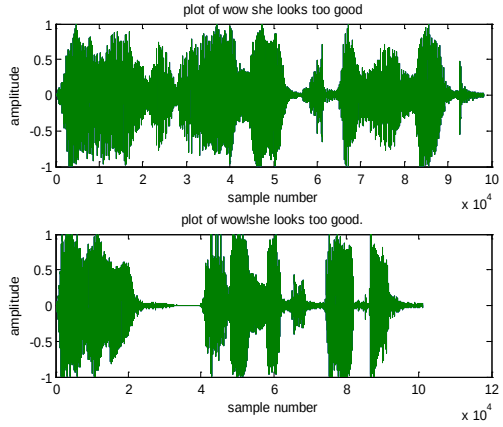

Fig.2: Waveform of neutral and exclamatory speech

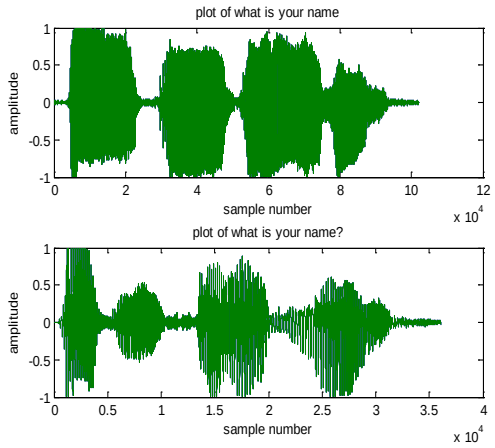

Fig.3: Waveform of neutral and questionnaire speech

4.3 Finding speech contour of neutral and original emotional speech sentences

By using Autocorrelation method, the pitch period of Neutral and Original Emotional speech were found. Here Autocorrelation method was applied to find out the maximum peaks in the waveform. Then distances between two successive maximum peaks values will give pitch period in terms of sample are extracted. Below are the graphs showing pitch contour of neutral speech and original emotional speech.

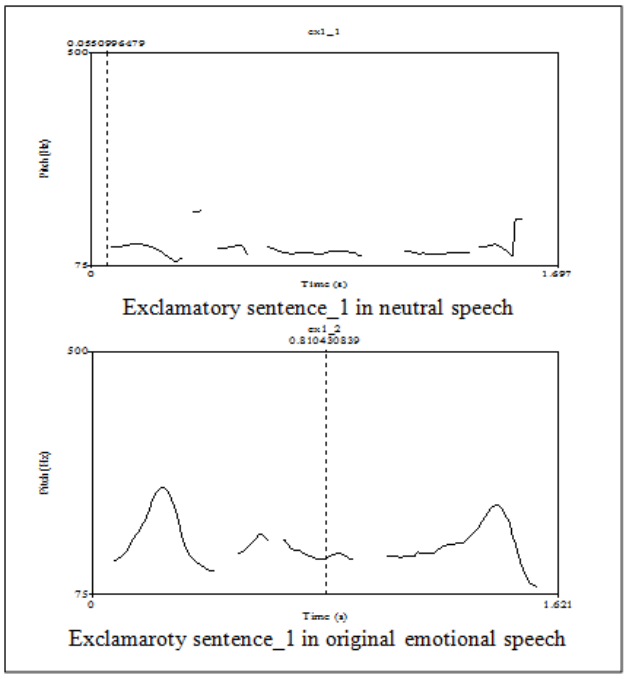

Fig.4: Pitch contours of exclamatory sentence

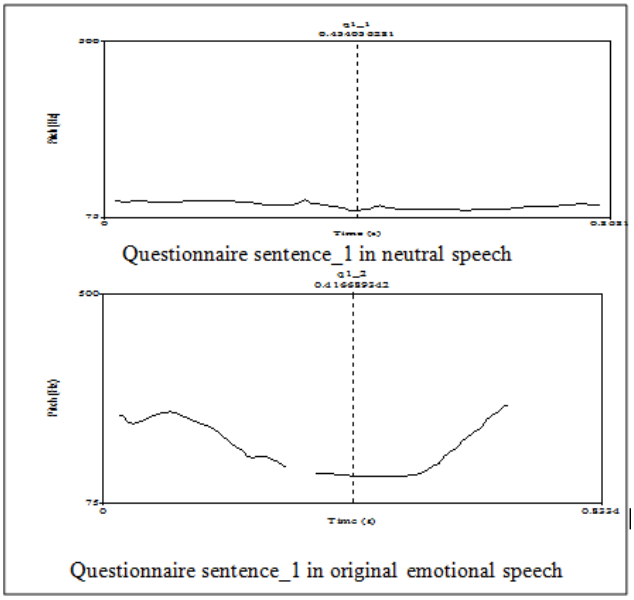

Fig.5: Pitch contours of questionnaire sentence

ISSN 2348 – 7968

4.4 Pitch modification and synthesis with LPC and PSOLA synthesizer

The Neutral speech sentences were taken for pitch modification. As per previous observation, the differences are used to modify the neutral speech in to emotional speech to get desired output. Thus the neutral speeches were modified to emotional speeches. The LPC and PSOLA synthesis were applied on the modified emotional speeches. Below are the graphs shows the pitch ck2ontours of original emotional speech and modified and synthesized emotional speech for both LPC and PSOLA synthesis.

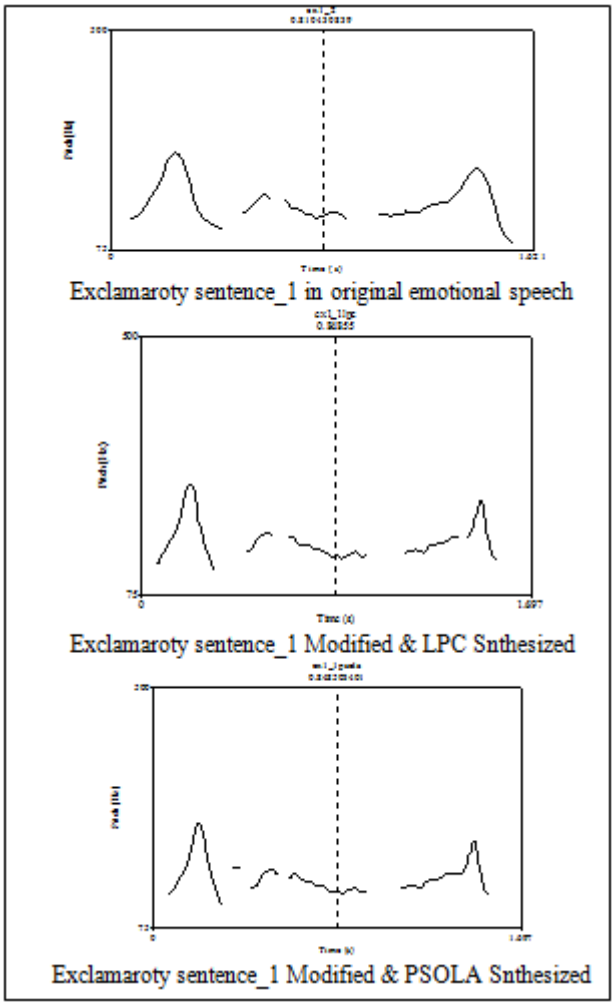

Fig.6: Modified and Synthesized questionnaire sentence

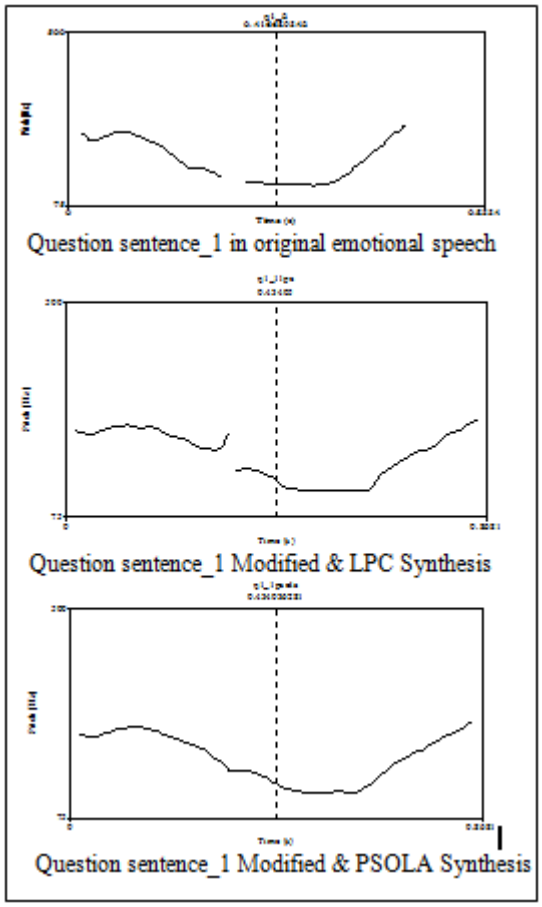

Fig.7: Wave files of original emotional speech Vs modified emotional speech

The graphs shows that the pitch contours of original emotional speech sentences and modified speech sentences are having same pattern and more similarities in terms of pitch frequencies. Consequently the results would be accurate after pitch modification and synthesis. The accuracy can be found in terms of Intelligibility index in next step.

The graphs of wave files of original emotional sentences and modified synthesized sentences are as below,

ISSN 2348 – 7968

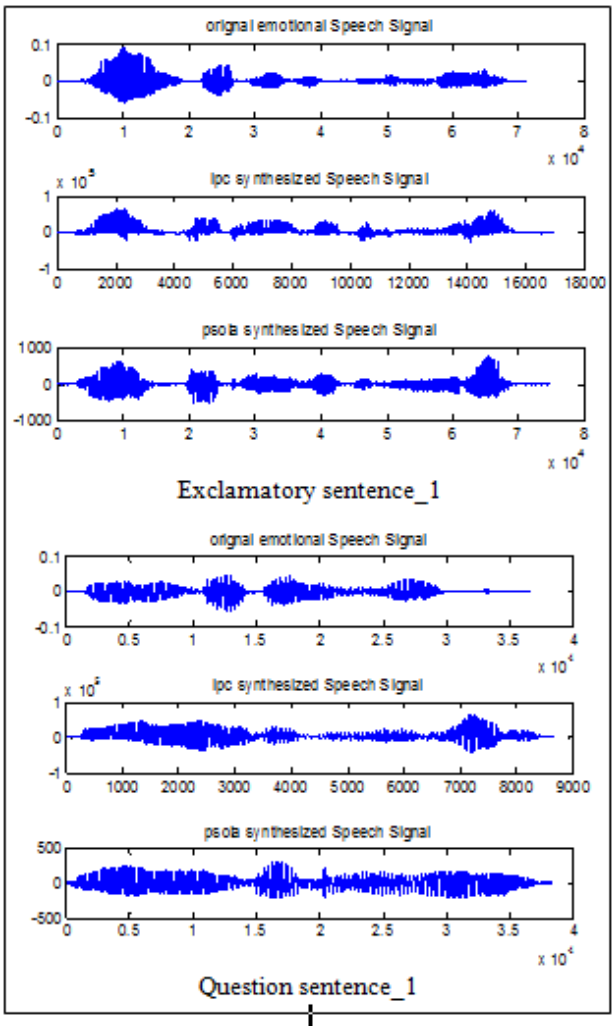

Fig.8: Wave files of original emotional speech Vs modified emotional speech

4.5 Finding Intelligibility Index

As explained in system implementation part, Articulation Index or Intelligibility Index will define the accuracy of the modified speech with respect to original emotional speech. Intelligibility index for both sentences which have been synthesized through LPC and PSOLA synthesis is calculated.

Below are the calculation results for all the 10 speakers for both exclamatory and questioner sentences.

Table 1: Intelligibility Index for Exclamatory sentences

Exclamatory Sentences (LPC Synthesis)										
Sentence No.	Speakers									
	1	2	3	4	5	6	7	8	9	10
1	76%	82%	60%	82%	79%	80%	86%	62%	91%	100%
2	38%	60%	68%	63%	77%	91%	84%	54%	86%	79%
3	63%	81%	81%	82%	59%	68%	80%	87%	51%	68%
4	61%	61%	88%	72%	58%	81%	83%	61%	48%	87%
5	45%	92%	97%	91%	82%	69%	80%	97%	89%	78%
6	77%	82%	86%	86%	53%	71%	78%	75%	68%	82%
7	54%	51%	30%	51%	88%	45%	76%	99%	81%	69%
8	100%	82%	82%	48%	66%	69%	75%	84%	69%	89%
9	79%	82%	93%	89%	86%	69%	62%	96%	71%	76%
10	68%	71%	99%	76%	77%	91%	61%	78%	45%	72%
Avg. Value	**66%**	**75%**	**78%**	**74%**	**73%**	**73%**	**77%**	**79%**	**70%**	**80%**

Exclamatory Sentences (PSOLA Synthesis)										
Sentence No.	Speakers									
	1	2	3	4	5	6	7	8	9	10
1	14%	34%	23%	24%	30%	17%	32%	39%	18%	28%
2	13%	31%	40%	16%	35%	15%	28%	37%	25%	24%
3	10%	36%	33%	16%	31%	13%	29%	38%	28%	11%
4	12%	20%	35%	21%	26%	34%	50%	38%	25%	12%
5	10%	26%	45%	18%	25%	15%	35%	27%	30%	19%
6	16%	26%	37%	25%	24%	19%	43%	27%	13%	29%
7	11%	21%	9%	28%	23%	17%	31%	36%	34%	28%
8	28%	27%	39%	25%	44%	25%	45%	24%	15%	30%
9	24%	21%	36%	30%	15%	8%	49%	43%	19%	33%
10	11%	13%	32%	33%	35%	28%	25%	33%	17%	10%
Avg. Value	**15%**	**26%**	**33%**	**24%**	**29%**	**19%**	**37%**	**34%**	**23%**	**23%**

Table 2: Intelligibility Index for Questionnaire sentences

Questionnaire Sentences (LPC Synthesis)										
Sentence No.	Speakers									
	1	2	3	4	5	6	7	8	9	10
1	87%	46%	92%	72%	54%	47%	61%	89%	83%	75%
2	78%	58%	88%	75%	81%	67%	84%	59%	80%	68%
3	82%	93%	100%	99%	82%	55%	65%	62%	66%	47%
4	69%	78%	98%	99%	83%	83%	88%	78%	64%	67%
5	77%	86%	96%	89%	83%	80%	85%	63%	69%	55%
6	75%	75%	88%	89%	87%	66%	61%	55%	82%	83%
7	63%	74%	90%	82%	85%	64%	82%	58%	93%	80%
8	79%	84%	85%	93%	68%	69%	74%	74%	89%	66%
9	93%	82%	88%	98%	84%	85%	72%	65%	83%	64%
10	59%	82%	98%	99%	80%	88%	80%	96%	81%	78%
Avg. Value	**76%**	**76%**	**92%**	**89%**	**79%**	**70%**	**75%**	**70%**	**79%**	**68%**

Questionnaire Sentences (PSOLA Synthesis)										
Sentence No.	Speakers									
	1	2	3	4	5	6	7	8	9	10
1	12%	19%	24%	10%	42%	18%	32%	39%	19%	32%
2	19%	31%	24%	32%	27%	9%	30%	37%	15%	31%
3	29%	21%	38%	15%	31%	19%	40%	22%	9%	18%
4	28%	4%	46%	23%	21%	19%	42%	40%	21%	9%
5	18%	20%	37%	25%	33%	15%	25%	21%	15%	19%
6	24%	15%	34%	20%	27%	9%	31%	19%	42%	19%
7	27%	10%	34%	18%	34%	21%	32%	28%	44%	15%
8	26%	15%	26%	16%	31%	15%	32%	41%	32%	9%
9	30%	27%	28%	15%	37%	17%	37%	42%	31%	21%
10	20%	21%	45%	34%	26%	14%	24%	47%	29%	40%
Avg. Value	**23%**	**18%**	**34%**	**21%**	**31%**	**16%**	**32%**	**34%**	**26%**	**21%**

It is observed that the Intelligibility index of modified and LPC synthesized sentences are around 70% and above while Intelligibility index of modified and PSOLA synthesized sentences are ranging from 15% to 30%.

The graphs of average values of each speaker for LPC and PSOLA synthesized sentences were plotted and trade line for linear relation is plotted as below,

ISSN 2348 – 7968

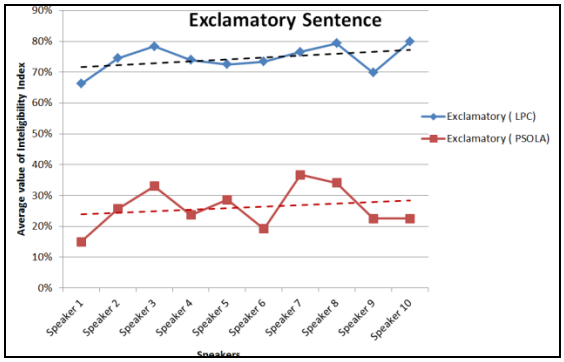

Fig.9: Intelligibility Index for exclamatory sentence

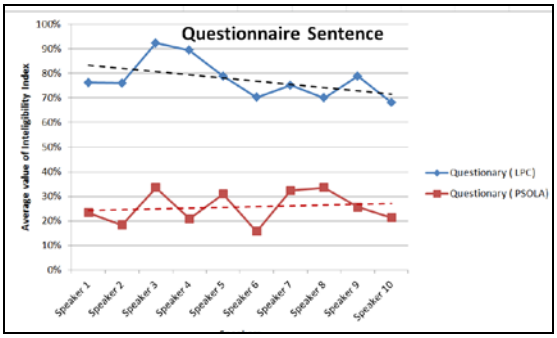

Fig.10: Intelligibility Index for questionnaire sentence

4.6 Mean operating score test

The modified emotional and synthesized sentences of exclamatory and questionnaire of each speaker have been played to each Listener. Thus each sentence can be played to all 10 listeners for their opinion. The average mark of ten listeners for each modified sentence synthesized with LPC as well as PSOLA is calculated. Below are the results for the MOS test.

Table 3: MOS test for LPC & PSOLA synthesized exclamatory sentences

Modified & LPC Synthesized Exclamatory Sentences _average values of 10 Listeners										
Speakers	Sentences									
	1	2	3	4	5	6	7	8	9	10
1	4.4	4.4	4.3	4.3	4.3	4.4	4.7	4.2	3.8	4
2	4.4	4.1	4.1	4.3	4.1	4.4	4.1	4.2	3.7	3.9
3	4	4.2	4.1	4.4	4	4.3	4	4.2	3.8	4
4	4.2	3.8	4.1	4.3	4.1	4.4	4.1	4.2	3.9	4
5	4.3	4.2	4.2	4.2	4.2	4.3	4.6	4.2	3.8	4.1
6	4.4	4.2	4.1	4.1	4.2	4.3	4.7	4.1	3.7	3.9
7	4.4	4.3	4.3	4.1	4.3	4.1	4.7	4.2	3.8	4
8	4.4	4.4	4.3	4	4.1	4.1	4.7	4	3.8	3.7
9	4.4	4.4	4	4	4.3	4.2	4.6	4.1	3.7	4
10	4.4	4.4	4.3	4.3	4.3	4.4	4.7	4.2	3.8	4
Avg. Value	**4.33**	**4.24**	**4.18**	**4.2**	**4.19**	**4.29**	**4.49**	**4.16**	**3.78**	**3.96**

Modified & LPC Synthesized Questionnaire Sentences_ average values of 10 Listeners										
Speakers	Sentences									
	1	2	3	4	5	6	7	8	9	10
1	4.4	4.4	4.3	4.3	4.3	4.4	4.7	4.2	3.8	4
2	4.4	4.1	4.1	4.3	4.1	4.4	4.1	4.2	3.7	3.9
3	4	4.2	4.1	4.4	4	4.3	4	4.2	3.8	4
4	4.2	3.8	4.1	4.3	4.1	4.4	4.1	4.2	3.9	4
5	4.3	4.2	4.2	4.2	4.2	4.3	4.6	4.2	3.8	4.1
6	4.4	4.2	4.1	4.1	4.2	4.3	4.7	4.1	3.7	3.9
7	4.4	4.3	4.3	4.1	4.3	4.1	4.7	4.2	3.8	4
8	4.4	4.4	4.3	4	4.1	4.1	4.7	4	3.8	3.7
9	4.4	4.4	4	4	4.3	4.2	4.6	4.1	3.7	4
10	4.4	4.4	4.3	4.3	4.3	4.4	4.7	4.2	3.8	4
Avg. Value	**4.33**	**4.24**	**4.18**	**4.2**	**4.19**	**4.29**	**4.49**	**4.16**	**3.78**	**3.96**

Table 4: MOS test for LPC & PSOLA synthesized questionnaire sentences

Modified & PSOLA Synthesized Exclamatory Sentences _average values of 10 Listeners										
Speakers	Sentences									
	1	2	3	4	5	6	7	8	9	10
1	3.90	3.90	3.80	3.90	3.90	3.70	3.70	3.60	3.50	3.60
2	3.90	3.80	3.80	3.90	3.80	4.00	3.80	3.90	3.70	3.70
3	3.60	3.80	3.80	3.90	3.70	3.90	3.70	3.90	3.70	3.70
4	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	3.00
5	3.85	3.68	3.84	3.90	3.89	4.00	3.89	3.90	3.74	3.79
6	3.90	3.80	3.80	3.80	4.00	4.00	4.00	3.80	3.60	3.80
7	3.90	4.00	3.90	3.80	4.00	3.80	4.00	3.90	3.70	3.80
8	3.90	3.90	3.90	3.70	3.90	3.90	4.00	3.70	3.70	3.60
9	4.00	4.00	3.00	4.00	4.00	4.00	4.00	4.00	3.00	4.00
10	3.89	3.89	3.67	3.89	3.44	3.56	3.44	3.89	3.78	3.78
Avg. Value	**3.88**	**3.88**	**3.75**	**3.88**	**3.86**	**3.89**	**3.85**	**3.86**	**3.64**	**3.68**

Modified & PSOLA Synthesized Questionnaire Sentences_ average values of 10 Listeners										
Speakers	Sentences									
	1	2	3	4	5	6	7	8	9	10
1	3.80	4.00	3.90	3.90	4.00	4.00	4.00	3.90	3.70	3.70
2	3.90	3.90	3.80	3.90	3.80	4.00	3.80	3.90	3.70	3.60
3	3.70	3.90	3.70	3.90	3.80	3.90	3.70	3.90	3.70	3.60
4	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	3.00
5	3.95	3.84	3.84	3.63	3.89	4.00	3.89	3.90	3.74	3.73
6	3.90	3.90	3.80	3.80	4.00	4.00	4.00	3.80	3.60	3.80
7	4.00	3.70	3.90	3.60	4.00	3.80	4.00	3.90	3.70	3.80
8	3.90	3.80	3.90	3.70	3.90	3.90	4.00	3.70	3.70	3.60
9	4.00	4.00	3.00	4.00	4.00	4.00	4.00	4.00	3.00	4.00
10	3.78	3.89	3.78	3.89	4.00	4.00	3.67	3.67	3.78	3.56
Avg. Value	**3.89**	**3.89**	**3.76**	**3.83**	**3.94**	**3.96**	**3.91**	**3.87**	**3.66**	**3.64**

As mentioned under system implementation heading, 1 indicates poor rating while 5 indicates best rating. From kabove ratings it is observed that the overall average ratings for all sentences synthesized with LPC synthesizer vary from 3.7 to 4.7 and with PSOLA synthesizer vary from 3 to 4. Further the average of average values for each sentence has been taken to plot the graph. Below is the graphical representation for Average of average values for all the modified and synthesized Exclamatory and Questionnaire sentences.

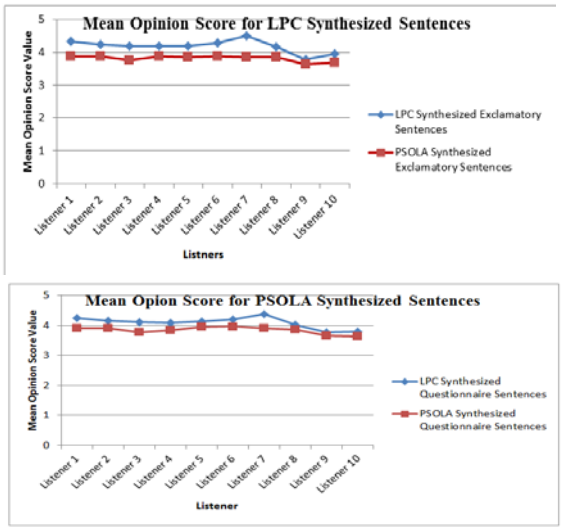

Fig.11 Graph of MOS test rating

The graph patterns are having similar trend for both the synthesis methods and highly depend on the listeners.

ISSN 2348 – 7968

5. Conclusions

Based on the defined algorithm, the pitch modification of neutral speech sentences to desired emotional speech sentences are achieved with considerable accuracy through Prosody conversion.

The pitch modification with LPC and PSOLA synthesizers showed similar pitch contour compare to original emotional speech sentences. However the intelligibility index graph shows that LPC synthesized sentences is having great accuracy and more close to the original emotion speech sentences. However accuracy of PSOLA Synthesized sentences is not up to the mark and having vast differences with original emotional speech sentences.

MOS test values and graph indicates that the sentences synthesized with both LPC and PSOLA methods are with desired emotions. However the LPC synthesized sentences are much better results in terms of auditability and pitch amplitude compare to PSOLA synthesized sentences. Moreover from the graph of MOS test it is observed that the ratings are highly depend on the Listeners ability and would like to recommend that the MOS test should be performed with higher number of listeners to have reduced variations and more accurate results.

Acknowledgments

It is my pleasure to get this opportunity to thank my beloved and respected Guide Prof. Dr. Mrs. Shaila D. Apte, who imparted valuable basic knowledge of Electronics specifically related to Speech Processing.

References

[1]."Speech and Audio Processing "by Dr.Shaila D.Apte, Published by Wiley India Pvt.Ltd.

[2]. Jianhua Tao, member, IEEE, Yongguo Kang and Aijun Li, "Prosody conversion from neutral speech to emotional speech, in IEEE transactions on audio, speech and language processing, vol.14, No.4, July 2006".

[3]. Daniel Erro, Eva Navas, Inma Hernaez and Ibon Saratxaga, "Emotion conversion based on prosodic unit selection in IEEE transactions on audio, speech and language processing, vol.18, No.5, July 2010".

[4]. Gan Zhen, Yu Hong-Zhi, Yang Hong-Wu, "Generation method of Lanzhou dialect speech based on Gaussian Mixture Model".

[5]. Chi Chun Hsia, Student member, IEEE, Chung-Hsien Wu, Senior member, IEEE, and Jian-Qi Wu, "Conversion function clustering and selection using linguistic and spectral information for emotional voice conversion in IEEE transactions on computers, vol.56, No.9, Sept 2007".

[6]. Chung-Hsien Wu, Senior member, IEEE, Chi-Chun Hsia, Chung-Han Lee, and Mai-Chun Lin, "Hierarchical prosody conversion using regression – based clustering for emotional speech synthesis", in IEEE transactions on audio, speech and language processing, vol.18, No.6, Aug 2010.

[7]. Eva Navas, Inmaculada Hernaez, and Iker Luengo, "An Objectiv the and subjective study of the role of semantic and prosodic features in building corpora for emotional TTS", in IEEE transactions on audio, speech and language processing, vol.14, No.4, 2006.

[8]. ArchNA Agarwal, Anurag Jain, Nupur Prakash, S.S.Agrawal, "Word Based Emotion conversion in Hindi Language".

[9]. Ashwin Bellur, K.Badri Narayan, Raghava Krishnon K, Hema A Murthy, "Prosody modeling for syllable –based concatenative speech synthesis of hindi and tamil".

[10]. Huibin Jia,Jianhua Tao, "Prosody modeling for mandarin exclamatory speech" National Laboratory of pattern recognition, institute of automation, Chinese academy of sciences.