Emotion Generation using LPC Synthesis

[1]Mrs.Sulakshana N. Bhatlawande, [2]Prof. Dr. Shaila D. Apte
[1]PG Student, Rajarshi Shahu College of Engineering, Pune, India.
[2]Prof., Department of Electronics and Telecommunication Engineering, Pune, India.
[1]Email: sdapte@rediffmail.com
[2]*Email: sulakshananb@gmail.com*

Abstract—Speech synthesis means artificial production of human speech. A system used for this purpose is called a speech synthesizer. The most important qualities of a speech synthesis system are naturalness and intelligibility. Naturalness describes how closely the output sounds like human speech, while intelligibility is the ease with which the output is understood. Emotion is an important element in expressive speech synthesis.

This paper describes LPC analysis and synthesis technique. The LPCs are analysed for each speech segment and pitch period is detected. At synthesis the speech samples equal to the samples in one pitch period are reconstructed using LPC inverse synthesis. Thus by using LPC Synthesis we can implement pitch modification or duration modification or spectrum modification to introduce emotion in the neutral speech, such as happiness or anger.

Keywords— *LPC Vocoder, Voiced-Unvoiced decision, Pitch, LPC, Pitch modification, Emotion categories.*

*****

I. Introduction

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer and can be implemented in software or hardware with a completely "synthetic" voice output. The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood.

One of the aspects of naturalness most obviously missing in synthetic speech is appropriate emotional expressivity. The overall goal of the speech synthesis research community is to create natural sounding synthetic speech. To increase naturalness, researchers have been interested in synthesising emotional speech for a long time. Emotion is an important element in expressive speech synthesis. One way synthesised speech benefits from emotions is by delivering certain content in the right emotion (e.g. good news are delivered in a happy voice), therefore making the speech and the content more believable. Emotions can make the interaction with the computer more natural because the system reacts in ways that the user expects.

The implementation of emotions seems straight forward at first but a closer look reveals many difficulties in studying and implementing emotions. The difficulties start with the definition of emotions. People can recognise emotional speech but they cannot describe it. Emotions have a biological basis and are therefore evolutionarily shaped. There are at least six emotions (happiness, sadness, anger, fear, surprise, and disgust) that are expressed in the face and recognised in the same way in many cultures.

The remainder of this paper is organized as follows. Section II introduces the model parameters and model parameters of neutral speech are analysed. Section III describes different methods for taking voiced and unvoiced decision. Section IV gives information about the parameters of speech, which are to be analysed and provided as, inputs to the speech synthesizer for speech synthesis. An overview of the proposed approaches that is how we can modify these parameters for Emotional speech synthesis is described in Section V. Section VI describes different Emotion categories like Happiness, Anger, Sadness, and it describes how model parameters changes according to the change in Emotion. Section VII details the use of LPC for Concatenative synthesis.

II. LPC ANALYSIS - SYNTHESIS

The particular source-filter model used in LPC is known as the Linear Predictive Coding model. It has two key components: analysis or encoding and synthesis or decoding. The analysis part of LPC involves examining the speech signal and breaking it down into segments or blocks. Each segment is than examined further to find the answers to several key questions: Is the segment voiced or unvoiced? What is the pitch of the segment? What parameters are needed to build a filter that models the vocal tract for the current segment? The model parameters, which are pitch contour, and duration. Finally speech is synthesized by concatenating the selected synthesis units with modified emotions or by generating the speech output from the model parameters [9].

LPC analysis is usually conducted by a sender who answers these questions and usually transmits these answers onto a receiver. The receiver performs LPC synthesis by using the answers received to build a filter that when provided the correct input source will be able to accurately reproduce the original speech signal. Essentially, LPC synthesis tries to imitate human speech production. This diagram is for a general voice or speech coder (vocoder). All voice coders tend to model two things: excitation and articulation. Excitation is

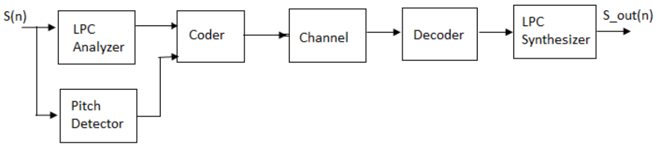

Fig 1 LPC Vocoder

the type of sound that is passed into the filter or vocal tract and articulation is the transformation of the excitation signal into speech. In the study of text-to-speech system Hidden Morkov Model (HMM)-based speech synthesizer utilizing glottal inverse filtering for generating natural sounding synthetic speech was described. The study presented a method to extract and model individual parameters for the voice source and the vocal tract, and a method to reconstruct a realistic voice source from the parameters using real glottal flow pulses [11].

The process of decoding a sequence of speech segments is the reverse of the encoding process. Each segment is decoded individually and the sequence of reproduced sound segments is joined together to represent the entire input speech signal. The decoding or synthesis of a speech segment is based on the information that is transmitted from the encoder. The speech signal is declared voiced or unvoiced based on the voiced/unvoiced determination bit. The decoder needs to know what type of signal the segment contains in order to determine what type of excitation signal will be given to the LPC filter.

III. Methods used for taking a voiced or unvoiced (V/UV) decision:

In any speech voiced vowels possess maximum energy, voiced consonants comes the next whereas unvoiced components possess the least energy. So, for voiced components , periodic airflow coming out of the vocal cords can be described by a periodic pulse train with its period T, hence F0 [= 1/T] is called the pitch of speech signal. Hence, pitch is a factor directly related to resonance of vocal cords.

When we plot any vowel, it shows periodicity. The rate of repetition of the pattern is known as the fundamental frequency (F0) or pitch frequency. Each repetition or period of these patterns corresponds to one glottal cycle, or one cycle of vocal fold opening and closing in the larynx. Pitch frequency is the fundamental frequency of vibrations of the vocal cords. This frequency generated by vocal cords in the form of the filter to produce a speech signal. Thus, speech is basically a convolved signal. Fundamental frequency is related to a voiced speech segment.

With purely unvoiced sounds, there is no fundamental frequency in the excitation signal and hence we consider the excitation as white noise. The air flow is forced through a vocal tract constriction occurring at several places between the glottis and mouth. Unvoiced sounds are found to be more silent and having less energy. They are less steady than voiced sounds.

When a speech sentence is recorded and analysed, we find that there are some silence parts in the utterances where the wave appears like grass, which means no energy is present or it is very less and is equivalent to small noise. The unvoiced segment has somewhat higher amplitude than the silence part and a voiced segment has even higher energy. The following methods are used for taking a voiced or unvoiced (v/uv) decision. The segment is analysed to decide if it is voiced or unvoiced segment. This is because the voiced or unvoiced part will be processed separately.

A. *Zero Crossing Rate (ZCR):*

The zero crossing count is an indicator of the frequency at which the energy is concentrated in the signal spectrum. Voiced speech is produced as a result of excitation of the vocal tract by the periodic flow of air at the glottis and this segment shows a low zero crossing count. Unvoiced speech is produced due to excitation in the form of turbulence in the vocal tract and shows a high zero crossing count. The zero crossing count of silence is expected to be lower than that for unvoiced speech, but quite comparable to that for voiced speech. However the energy for the silence part will be quite low as compared to the voiced segment.

B. *Pre-Emphasized Energy Ratio:*

Voiced and unvoiced segments can be discriminated using the normalized pre-emphasized energy ratio defined by equation,

$$P_r = \frac{\sum_{i=1}^{N} | s(i) - s(i\text{-}1) |}{\sum_{i=1}^{N} s^2 (i)}$$

Where P_r is the pre-emphasized energy ratio and $s(i)$ and $s(i\text{-}1)$ are the i^{th} and $(i\text{-}1)^{th}$ signal samples, respectively. The variance of the difference of the difference between adjacent samples for voiced speech will be very small and that for unvoiced speech will be higher. When we take the plot of pre-emphasized energy ratio we observe that, when a signal is voiced, the pre-emphasized energy ratio has a value very close to zero whereas for an unvoiced segment, the pre-emphasized energy ratio is high. If the pre-emphasized energy ratio is infinite value it is detected as silence part of the signal. Hence, using this parameter, we can identify the voiced, unvoiced, and silence parts of an utterance [1]. Fig.1 shows plot of an input speech signal and plot of pre-emphasized energy ratio of speech signal. Using this measure we can differentiate between voiced, unvoiced and silence part of input speech signal. Fig.2 shows plot of voiced part of a signal. Fig.3 shows plot of unvoiced part of a signal.

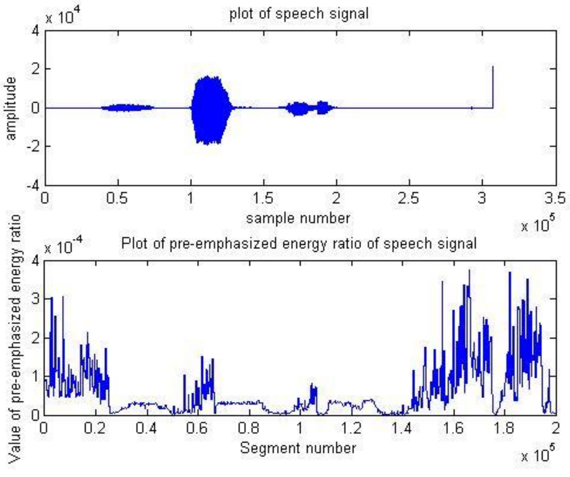

Fig. 2. Plot of Input Speech signal with neutral emotions and Plot of Pre-emphasized energy ratio of speech signal

C. Low Band to Full Band Energy Ratio:

Voiced and unvoiced segments can be discriminated using the low band to full band energy ratio. In a voiced speech signal, the energy is concentrated in the low band. We can measure the ratio of energy in the first 1 kHz band to the full band.
This ratio is high for a voiced speech and quite low for unvoiced speech [1].

When we take the plot of pre-emphasized energy ratio we observe that, when a signal is voiced, the low to full band energy ratio is attains its maximum value whereas for an unvoiced segment, the low to full band energy ratio is low. For the silence part, the low band to full band energy ratio is zero. Hence, using this parameter, we can identify the voiced, unvoiced, and silence parts of an utterance.

IV. ANALYSING SPEECH PARAMETERS

Voiced speech is generated when the excitation comes from a periodic pulse train generated by vocal cords. These vocal cords vibrate with their natural frequency of vibration like a tuning fork and generate pulses at regular intervals. We can extract the parameters related to the vocal cords specially the fundamental frequency. The parameters related to the vocal tract, functioning as a circular waveguide are formants, LPC, etc.

Speech synthesis requires pitch detection and a (V/UV) decision making algorithm as the essential elements. This task requires a combination of signal processing and feature extraction. All the present time domain algorithms for pitch period measurement detect the peak, which is not only positioned at the correct pitch, but also at its integer multiplies, thereby creating a possibility of getting multiple and half-pitch errors. This possibility can be eliminated using the parallel processing approach. Because of parallel functioning pitch estimators, the estimation can be done every 5ms which is found to be an appropriate interval. In case of unvoiced segment, this coincidence will not occur. This indicates that no pitch is present and the segment is unvoiced.

A. *Pitch Period*

We can use a variety of methods, such as autocorrelation, average magnitude difference function (AMDF), etc., for finding the pitch period. Autocorrelation Method for Finding Pitch Period of a Voiced Speech Signal: Autocorrelation is the correlation of a signal with itself. So it is a measure of the similarity between samples as a function of the time separation between them. It can be considered as a mathematical tool to find repeating patterns and their periods [7]. Fig. 4 and Fig.5 show plot of voiced part of a signal and plot of correlation of a signal, which gives pitch period which the period between two successive peaks in terms of no. of samples.

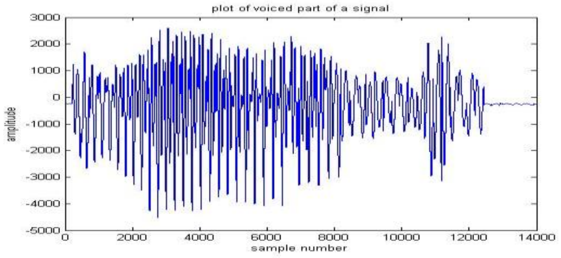

Fig. 3. Plot of Voiced part of a signal

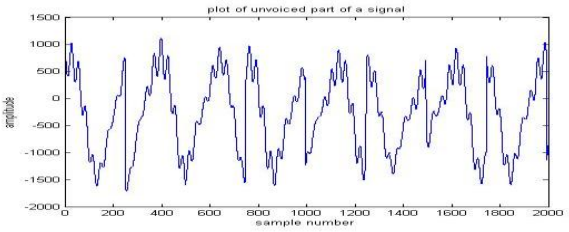

Fig. 4. Plot of Unvoiced part of a speech signal

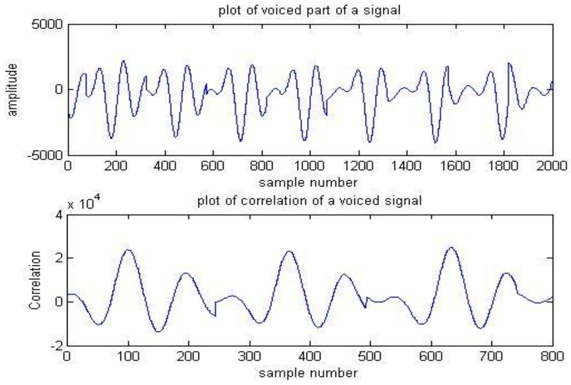

Fig. 5. Plot of Voiced part of a signal And Plot of correlation of a signal

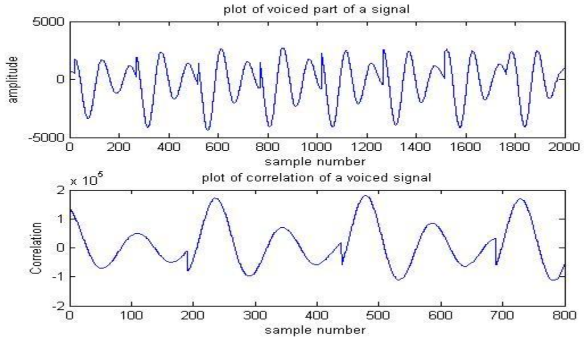

Fig. 6. Plot of Voiced part of a signal And Plot of correlation of a signal

B. Linear Predictive analysis (LPC)

Linear prediction based speech analysis has received considerable attention in the past four decades. In linear prediction, the speech waveform is represented by a set of parameters of an all-pole model, called the linear predictive coefficients (LPC) [5],[6] , which are closely related to speech production transfer function. The LPC analysis essentially attempts to find an optimal fit to the envelope of the speech spectrum from a given sequence of speech samples.

Speech signal is a random process. It has high autocorrelation. Hence it is possible to predict the next sample of speech from previous samples, though exact prediction is not possible. The error in the prediction is called prediction error [1].

Consider a voiced speech segment. The speech signal sample can be expressed as, Equation in Z-domain,

$$s[n] = \sum_{k=1}^{p} \alpha_p(k) s[n-k] + Au[n]$$

Where u (n) is a unit sample sequence or impulse representing excitation for voiced signal. A is a gain represented in the gain contour for impulse excitation. Thus Au (n) is the input to system and α_p are the prediction coefficients.

$$H(z) = \frac{S(z)}{U_g(z)} = \frac{A}{1+\sum_{k=1}^{p} \alpha_P(k)_k z^{-k}}$$

There are three different approaches for finding linear prediction coefficients (LPC).

1) Autocorrelation method. (Levinson –Durbin algorithm)
2) Covariance method.
3) Lattice structure method. (Burg's algorithm)

Fig. 6 and Fig. 7 show plot of LPC for different segments of voiced part of a signal.

V. Modification Parameters

Emotional simulation is achieved by a set of parameterized rules that describe manipulation of the following aspects of a speech signal: 1) Pitch changes 2) Duration changes 3) Spectrum modification

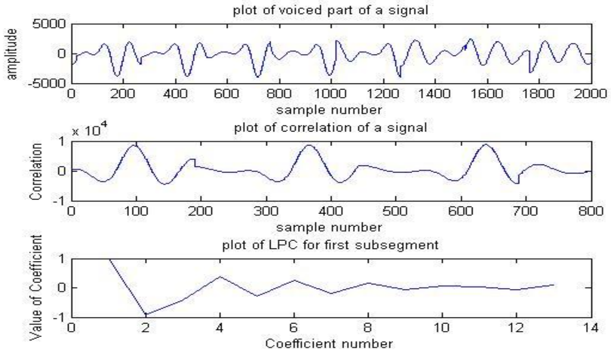

Fig. 7. Plot of LPC's for different sub-segments of Voiced segments of Speech signal

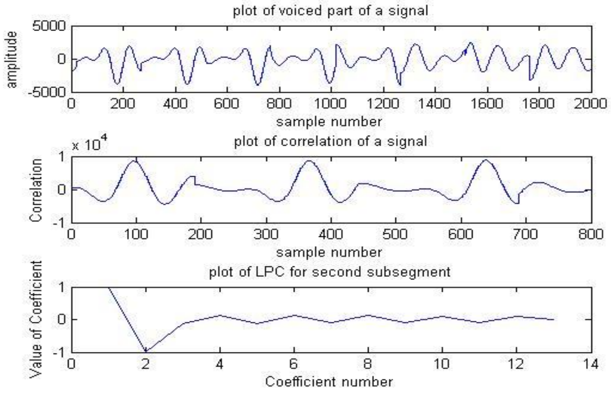

Fig. 8. Plot of LPC's for different sub-segments of Voiced segments of Speech signal

A. *Pitch Modification:*

The following modifications are provided for pitch feature changes pitch level modification, modification in pitch range, pitch value modification and alteration of pitch contour.

- Pitch Level: The modification in the pitch level means the overall level of the F0 (fundamental frequency contour is shifted by multiplying all pitch values with a rate factor (rate= 0 implies no change). When the rate value is high, the pitch values undergo stronger changes than when the rate value is low. This was chosen confirm to the human logarithmic hearing.

- Pitch Range: The pitch range refers to the dynamic range of pitch values. The variation in pitch range can be achieved by a shift of each F0-value by a percentile of its distance to the mean F0-value of the last syllable. If range=0, all pitch values become the last syllable's mean pitch value. If the range value is high, the pitch value is shifted from the mean pitch value of the last syllable by a large amount, thereby increasing the dynamic range of pitch values.

Pitch Variation: A pitch variation on the syllable-level is achieved by the application of the pitch range algorithm on each syllable separately. The reference value in this case is the syllable's mean pitch.

- Pitch Contour: A speech signal is a random signal. The voiced component of speech exhibits pseudo periodicity, which means that the waveform has a periodic nature but the period does not remain exactly constant, it varies to a very small extent. In LTV speech production model the periodic pulse train generated from the vibrations of the vocal cords acts as an excitation for the voiced speech signal. However, if we track the pitch period over the entire voiced segment, we find that there is a small variation in the pitch period.

The contour of variations of the pitch period is termed as pitch contour. That is the pitch contour is the graph of pitch variations over the utterance. The pitch contour of the whole phrase can be designed as a rising, a falling or a straight contour. The pitch contour also contains some information related to the spoken word and the speaker. The contours are characterized by a gradient. As a special variation for the generation of happy speech, the wave model can be used. Here, the pitch values for the main stressed syllables are raised, and the pitch values for the syllables that lie equally distanced in between, are allowed. It is characterised by the maximum amount of raising and connected to a smoothing of the pitch contour, because all F0 values get linearly interpolated. A rising, falling or level contour can be assigned to each syllable-type.

B. *Pitch modification using the analysis and synthesis method:*

Pitch modification can be implemented on similar lines using any analysis and synthesis techniques which extract the pitch period, such as, channel vocoder[2], [3],etc. The synthesis filter determines the short-term spectral envelope of the synthesized speech and is characterized by the linear prediction (LP) coefficients obtained from LP analysis on the input speech. These coefficients are commonly called LPC coefficients, which may refer generically to any of several different but equivalent parameter sets that specify the synthesis filter[3],[1].

C. *Analysis- Synthesis Technique used for Duration modification:*

The speech rate can be modified for the whole phrase. The rate change can be executed by changing the duration of the phonemes. If the duration in consequence of the length reduction is shorter than the frame rate, the phoneme gets dropped.

Consider a channel vocoder. In this each segment of input speech is analyzed using a bank of band–pass filters called the analysis filters. The energy at the output of each filter is estimated at fixed intervals and transmitted to the receiver.

At the receiver, the vocal tract filter is implemented by a bank of band-pass filters. The bank of filters at the receiver, known as the synthesis filters, is identical to the bank of analysis filters. Based on whether the speech segment was deemed to be voiced or unvoiced, either a pseudo noise source or a periodic pulse generator is used as the input to the synthesis filter bank. The period of pulse input is determined by the pitch estimate obtained for the segment being synthesized at the transmitter. This glottal flow pulse is interpolated in the time domain with a cubic spline interpolation algorithm [4], in order to achieve a specific fundamental period, and the energy (gain) of the pulse is equalized to the energy measure by the transmitter. The input is scaled by the energy estimate at the output of the analysis filters [1].

Fig.9 shows Block schematic of Channel vocoder transmitter. And Fig 10 shows Receiver for channel vocoder. It parameterizes excitation and spectrum separately. That is the fundamental frequency and the spectrums are parameterized, so they will remain unchanged.

D. Spectrum Modification:

Spectrum modification can be done using homomorphic coder and using sinusoidal coder. There is a possibility of male-to-female voice transformation if the fundamental frequency and vocal tract spectrum both are modified.

To establish the level of human performance as a baseline, we first measure the ability of listeners to discriminate between original speech utterances under three conditions: normal, fundamental frequency and duration normalized, and LPC coded. Additionally, the spectral parameter conversion function is tested in isolation by listening to source, target, and converted speakers as LPC coded speech. The results show that the speaker identity of speech whose LPC spectrum has been converted can be recognized as the target speaker with the same level of performance as discriminating between LPC coded speech. However, the level of discrimination of converted utterances produced by the full VC system is significantly below that of speaker discrimination of natural speech [1].

VI. EMOTION CATEGORIES

A common way of describing emotions is by assigning labels to them like emotion denoting words. There are a number of researchers that have compiled lists of emotional words. Of course some of these terms are more central to a certain emotion than others. Also different emotion theories have different methods for selecting such basic emotion words. To describe different forms of basic emotions like hot and cold anger, finer grained emotion categories are used.

When modelling emotions in speech synthesis two main approaches have been identified. One can model a few emotional states as closely as possible. Recent work in unit selection has used this approach with three full blown emotions. In this research a separate database was recorded for happy, angry, and sad tone of voice. Although this type of method almost guarantees good results, it is impossible to generalise the voice to other types of emotional expression, build a general speech synthesis system that allows control over various voice parameters and therefore should be able to model almost any type of expression.

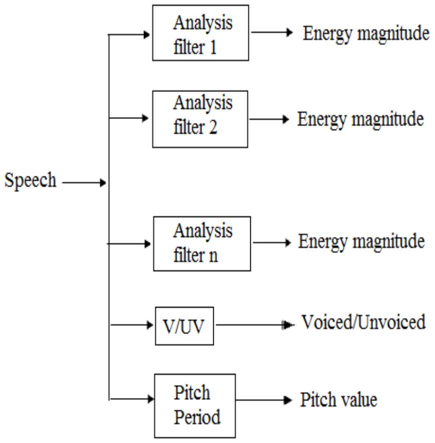

Fig.9. Channel vocoder transmitter block schematic

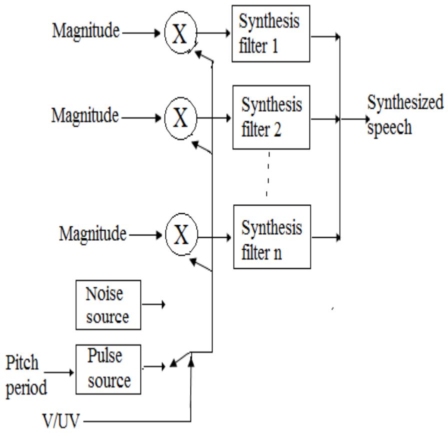

Fig. 10. Receiver for channel vocoder

For emotional speech synthesis, global prosodic parameters are often treated as universal or near-universal cues for emotion. At least formant and time-domain synthesis, prosody rules play an important role in automatically generated emotional expressivity in synthetic speech. The synthesis parameters can be varied to obtain perceptually optimal values for emotion synthesis. The global prosodic settings are F0 level, range, speech tempo and loudness. For example, steepness of F0 contour during rises and falls. Sometimes the speech tempo of vowels and consonants, stressed and unstressed syllables, and the placement of pauses within utterances for emotional speech synthesis. The F0 contour is shown to play an important role in emotion recognition.

The acoustic correlates of emotions in human speech. It is possible to achieve our goal only if there are some reliable acoustic correlates of emotion/affect in the acoustic characteristics of the signal. A number of researchers have already investigated this question. Their results agree on the speech correlates that come from physiological constraints and correspond to broad classes of basic emotions, but disagree and are unclear when one looks at the differences between the acoustic correlates of fear and surprise or boredom and sadness. Indeed, certain emotional states are often correlated with particular physiological states which in turn have quite mechanical and thus predictable effects on speech, especially on pitch, (fundamental frequency F0) timing and voice quality. For instance, when one is in a state of anger, fear or joy, the sympathetic nervous system is aroused, the heart rate and blood pressure increase, the mouth becomes dry and there are occasional muscle tremors. Speech is then loud, fast and enunciated with strong high frequency energy. When one is bored or sad, the parasympathetic nervous system is aroused, the heart rate and blood pressure decrease and salivation increases, which results in slow, low-pitched speech with little high-frequency energy. Furthermore, the fact that these physiological effects are rather universal means that there are common tendencies in the acoustical correlates of basic emotions across different cultures.

The speech had to be both of high quality and computationally cheap to generate. For this we can use a concatenative speech synthesizer, which is an enhancement of more traditional PSOLA techniques (it produces less distortions when pitch is manipulated). It allows to express emotions as efficiently as with formant synthesis, but with more simple controls and the liveliness of concatenative speech synthesis. The price of quality is that minimal control over the signal is possible, but this is compatible with our need of simplicity.

The parameters' values can be obtained for five affects: calm, anger, sadness, happiness, comfort. We can obtain these parameters first by looking at studies describing the acoustic correlates of each emotion then we can deduce some initial value for the parameters and modify them by hand, by trial and error until it give a satisfying result.

Depending on previous research following is a model of intonation contours for the different emotions:

Happiness: The mean pitch of the utterance is high, has a high variance, the rhythm is rather fast, few syllables are accented, the last word is accented, and the contours of all syllables are rising.

Anger: The mean pitch is high, has a high variance, the rhythm is fast, with little variance of phoneme durations, a lot of syllables are accented, the last word is not accented, the pitch contours of all syllables are falling.

Sadness: The mean pitch is low, has a low variance, the rhythm is slow, with high variance of phoneme durations, very few syllables are accented, the last word is not accented, the contours of all syllables are falling.

VII. USE OF LPC FOR CONCATENATIVE SYNTHESIS

LPCs are used for synthesis of segments and concatenative synthesis can be done using synthesized segments. The speech signal is divided into a number of segments of say 10-20 ms duration. If we consider speech having a duration of 10-12. LPC for each segment will be evaluated using the Levinson-Durbin algorithm or the Burg algorithm as explained above. The samples in the segment are predicted and the prediction error is evaluated. The prediction error peaks are used as excitation for speech synthesis of segment.

The synthesized speech is intelligible. The synthesized segments are joined or concatenated to get synthesized speech, which is found to be more intelligible if the speech segments are pitch synchronous. If we use the error signal as it is an excitation for the LPC model, the reconstructed speech signal will be an exact replica of the original speech signal. Instead of transmitting the entire speech residual signal we can extract the excitation information from the error signal and use it for excitation. That is track the peaks in the residual signal and use the pulses so obtained as the excitation to generate speech.

Wu et al. Adopted variable –length units for synthesizing high-quality speech [10].

VIII. CONCLUSION

We have proposed a new Emotional Speech Synthesis System LPC based Analysis-synthesis framework is employed for Pitch modification, Duration modification or Spectrum Modification. There is scope to investigate new methods for analysing Model parameters and synthesising the emotional speech specifying different Expressions or Emotions. The results for Emotional Speech Synthesis are not given here. We are working progressively in this direction.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Prof Dr. Mrs. S. D. Apte, "*Speech and Audio Processing"*, Ref. book by WILEY-INDIA, First Edition: 2012.

[2] A. Gersho, "*Advances in speech and audio compression*," Proc. IEEE, vol. 82, pp. 900-918, 1994.

[3] Roar Hagen, Erdal Paksoy, Allen Gersho, "*Voicing-Specific LPC Quantization for Variable-Rate Speech Coding*", IEEE Tran. On Speech And Processing, Vol.7, No. 5, Sept 1999.

[4] Tuomo Raitio, Antti Suni, Junichi Yamagishi, Hannu Pulakka, Jani Nurminen, Martti Vainio, and Paav Alku.,"*HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering*" IEEE Transactions on audio, speech, and language processing, vol. 19, no. 1, January 2011.

[5] V.Jain and R. Hangartner, "*Efficient algorithm for multipulse LPC analysis of speech*", in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 1984.

[6] Amro EI-Jaroudi, Member, IEEE, and John Makhoul, Fellow, IEEE,"*Discrete All-Pole Modeling*," IEEE Transactions on Signal Processing, Vol. 39, No. 2, Feb 1991.

[7] Lawrence r. Rabiner, Fellow, IEEE, "*On the Use of Autocorrelation Analysis for Pitch Detection*", IEEE Tran on Acoustics, Speech, and Signal processing, Vol ASSP-25, No.1, Feb 1977.

[8] M. Schroder, "*Emotional speech synthesis- a review*", in Proc. Eurospeech'01, Aalborg, Denmark, 2001, Vol. 1, pp. 561-564.

[9] Chi-Chun Hsia, Chung-Hsien Wu, Senior Member, IEEE, and Jung-Yu Wu, "*Exploiting Prosody Hierarchy and Dynamic Features for Pitch Modeling and Generation in HMM-Based Speech Synthesis*".IEEE Transactions on audio, speech, and language processing, vol. 18, no. 8, November 2010.

[10] Chung-Hsien Wu, Senior Member, IEEE, Chi-ChunHsia, Jiun-Fu Chen, and Jhing-Fa Wang, Fellow, IEEE, "*Variable-Length Unit Selection in TTS Using Structural Syntactic Cost*", IEEE Transactions on audio, speech, and language processing, vol. 15, no. 4,pp. 1227- 1235, may 2007.

[11] Tuomo Raitio, Antti Suni, Junichi Yamagishi, Hannu Pulakka, Jani Nurminen, Martti Vainio, and Paavo Alku, "*HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering*", IEEE Transactions on audio, speech, and language processing, vol. 19, no. 1, January 2011.