Modeling Prosodic Parameter of 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 ,88888875096.
#2 Department of Electronics Engg., Bharati Vidyapeeth Deemed University college of Engg.Mharashtra,India.411043,9689880817.

ABSTRACT

Emotion of speech expresses the affective state of the speaker.There are different prosodic parameters, e.g., F0, duration, and intensity of the given utterance, innotation pattern which can be modify to generate emotional speech.There were different methods are tried like A linear modification model (LMM), a Gaussian mixture model (GMM) method and a classification and regression tree (CART) method. There are different methods of time scaling like TD-PSOLA, FD-SOLA, WSOLA, HNM. here we have used TD-PSOLA method.In TD-PSOLA,Time and pitch scaling is directly performed on speech segment.spectral and formant properties remains unchanged. TPSOLA gives better result than other time and pitch scaling methods.

Key words: *FD-PSOLA,HNM,TD-PSOLA,WSOLA,ZCR*

corresponding Author: Ms.C .D.Pophale

INTRODUCTION

Speech is a natural way of communication between people. It is a randam naturally occurring signal.The two components of speech coded like , :(i) “What is said” and (ii) “How it is said” are very important. The first component indicates the linguistic information. The second component shows non-linguistic or suprasegmental component which indicates the prosody or emotion state of speaker . i.e. pitch, intensity and speaking-rate rules.

TD-PSOLA is a mostly used technique for prosodic modification of speech signals. There are two types of PSOLA method,TD-PSOLA and FD-PSOLA.But for speech synthesis,TD-PSOLA is mostly used.It is also called as pitch changing algorithem.The quality of TD-PSOLA is depends on methods used for proper pitch marks of voiced segments.

In section III of this paper, we present the detailed methods for pitch marking ,the voiced segments detection and TDPSOLA method. Finally, section IV,we have briefly discuss an evaluation of the results and section V,We concludes the paper.

II. Data Collection

A microphone is used for single channel recording.In an almost noise –free small closed room,20 sentences of different words in neutral emotions are recorded by female students

III. Methodology

A schematic of Diagram of modeling prosodic parameter of neutral speech is shown in Figure 1 That will be same for all emotions only pscale & tscale facters will be change.

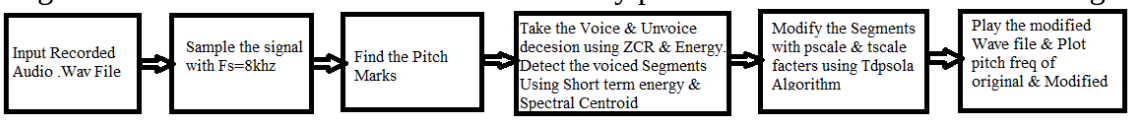

Fig. 1: A Schematic of Synthesizer.

A schematic speech synthesizer is shown in Figure 1. The recorded wave file with neutral emotion is the input to the system. To Extract the pitch marks following steps are followed.

Step1: Pitch tracking and Pitch marking

Pitch tracking will determine the pitch (F0) over a large interval of speech. Each point on pitch track will give a value of fundamental frequency.

Step2: Pitch synchronous analysis

The pitch synchronous analysis is a process ,used to find out the pitch period. The pitch cycle information is used to form frames. A short term energy of each sequeneces within frame is obtained by convolving the squared samples of speech with a Hanning window .The obtained short time energy contour have large amplitude peaks at regular intervals in the voiced regions and small-amplitude peaks in the unvoiced region at irregular intervals.

Step3: Pitch period finding using autocorrelation method

Autocorrelations is the correlation of signal with itself. It is the similarity between samples as a function of the time separation between them. It is a mathematical tool to find repeating patterns and their periods. Autocorrelation methods will list two pitch periods to detect pitch.

Step 4: Find pitch mark

The algorithm for pitch marking is given as:

1. Pre-processing
2. Block creation Process
3. Peak picking process
4. Peak organization into sub frames
5. Dynamic programming

4.1. Pre-processing

Some speech signals have more positive peaks, and some have more negative peaks. First find out voice regions . the average of peak amplitude for original signal for the voiced regions is calculated, Then it is calculated for the inverted speech signal. If the average peak amplitude of inverted signal is greater than of the original signal ,then signal is inverted and used for processing.otherwise polarity of the original signal is maintained for all remaining processing

4.2. Block Creation Process

There are four categories of block for the regions u-u,u-v,v-v,or v-u or a combination of them. blocks are short in time(30-60msec). The block size depends on average pitch period ,It is a multiple of average pitch period and must have either 1 or 2 pitch cycles in region. It varies on the pitch values of the region.

4.3. Peak Identification

This step involves locating peaks of speech signal in the block which will be worked as land marks. Same peak will be end point of previous pitch cycle and starting point of next pitch cycle. .Block is normalized to 1, sliding window is used to find out the largest peak amplitude,which is middle point of window.

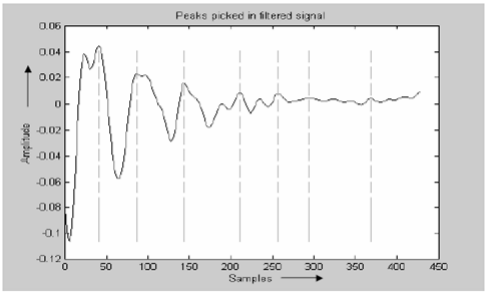

Fig.3 Peaks identified in peak locating process in v-u region

4.4 Dynamic programming

Dynamic programming is used to classify the peaks into subframes and to find out the lowest cost set of pitch period markers.

We have assumed that the number of subframes are equal to number of pitch periods. blocks are divided into overlapping subframe.The size of subframe is of 2*pp. a block will contain 4 subframes. In such way peaks are decided and classified into subframes.

After that two matrices are formed, matrix X and matrix Y.Matrix X stores the local cost values with size of *m* rows and *n* columns. the matrix Y , stores the transition values with size of have *m*m* rows and *n-1* columns.

the local cost values is given as 1- normalized amplitude of peak for subframe.

The contents in matrix Y are a absolute difference between a peak location in current subframe and peak location in next subframe, Thus matrix X represent a local cost of each candidate pitch marker. Matrix Y represent transition costs required for selecting a certain peak in one frame and a particular peak in the next frame.

Dynamic programming is used to find out the lowest cost path or pitch markers through the above two matrice.Finaly we get *n-1* number of total markers; Each marker will be end point of last subframe and starting point of next frame.

Step 5:Detection of voice and unvoiced

ZCR means zero crossing count method and short term energy method is used for voice and unvoice decision.

For voiced speech short-term energy is high and zero-crossings are low, and for un-voiced speech the short-term energy is low and zero crossings are high and both are approximately zero for silence[8].

A deffination of ZCR is

$$\boldsymbol{Zn = \sum_{m=-\infty}^{\infty}\{sgn[x(m)] - sgn[x(m-1)]\}w(n-m)}$$

A definition of Short-time energy can be given as:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \quad \text{i.e.,(2)}$$

5.2 Detection of voice segment

Signal Energy and Spectral centroid this two featurs are used for detection of voice segment.

a Signal Energy can be defined as : for each frame *i* the energy is calculated as

$$E(i) = \frac{1}{N}\sum_{n=1}^{N} |x_i(n)|^2 \quad \text{i.e.,(3)}$$

where $xi(n)$ the audio samples of the i_ith frame, of length *N*.

This feature is used for detecting silent periods in audio signals.It is high for voice segment and low for unvoice segment.

b. Spectral centroid: The spectral centroid, is defined as the center of"gravity" of its spectrum .

It is calculated as

Spectral centroid = $\sum_0^{n-1} x(n)\, f(n) / \sum_0^{n-1} x(n)$ i.e.,(4)

$f(n)$ is the FFT for frame *n* and *x(n)* is the index of frequency. Spectral centroid values are lower for unvoice segment means for lower frequencies and high for voice segment or for higher frequencies.

Two threshold values are computed for example T1 and T2 respectively.Calculated values of spectral centroid and energy for sequences are compared with threshold.then decision is taken of voice and unvoice segment.

Step 6:TD-PSOLA

6 TD-PSOLA algorithm

The TD-PSOLA algorithm will allow pitch modification without changing spectral and formant properties of signal. the time duration modification will change the rate of speech signal [6]. The maximum correlation the between the present and previously synthesized segment is achieved by the proper position of the segment.

TD-PSOLA works pith-synchronously. Per pitch period, there must be one analysis window. A the length of segment will be multiple number of pitch periods to preserve the periodicity of overlapped synthesized segments.

For PSOLA, pitch marks are first find out, pitch marks means the peaks of pitch period are marked. For unvoiced, the pitch marks are placed with a constant spacing. For voiced, pitch scaling is performed.To increase the pitch , this synthesized segments are added with more overlap , or less overlap for lowered pitch. This procedure is performed on the speech signal directly. the shape of the waveform is preserved and the formant structure remains unchanged.

IV Result and Discussion

In this database "Kiti vela Sangital? Abhyas kara ..kara…kara " word is uttered by female speaker.

Figure 5 shows pitch variation and figure 6 shows spectral centroide and short term energy of input neutral speech signal.

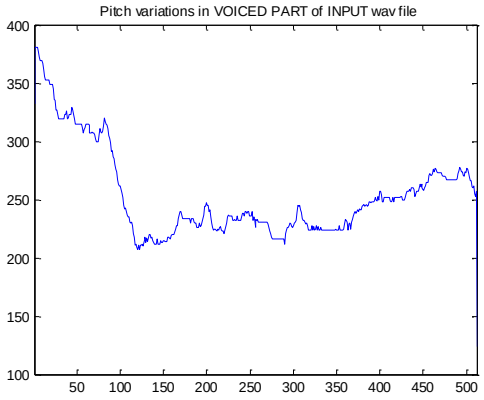

Fig.5: Pitch variation of voiced part of input Neutral speech.

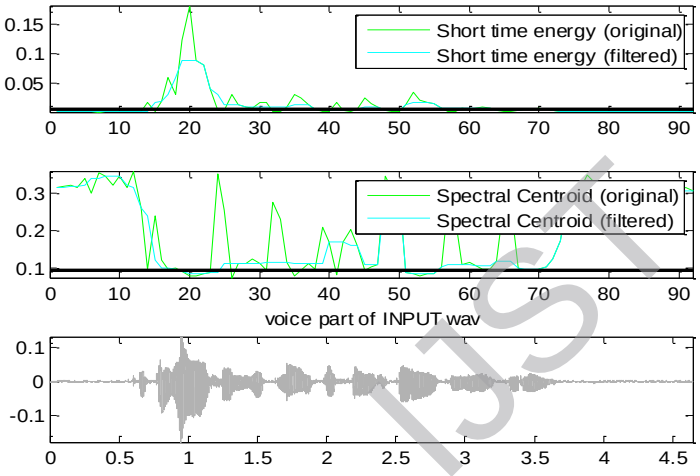

Fig.6:Short term energy & Spectral centroid of original and filtered signal of input Neutral Speech

Figure 7 shows variation of pitch for modified signal with respect to original signal for emotion sad.

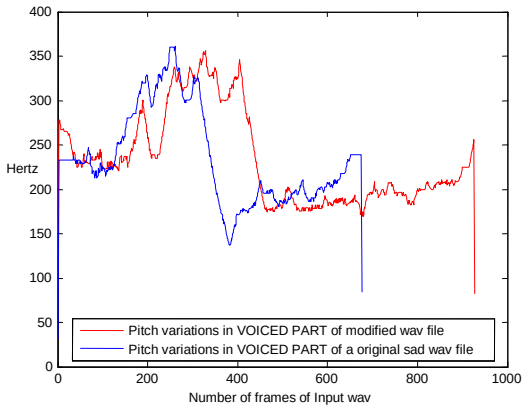

Fig .7 :Pitch variation of original and modified signal of emotion sad.

Figure 8 shows short term energy and spectral centroid for modified signal with respect to original signal for emotion sad.

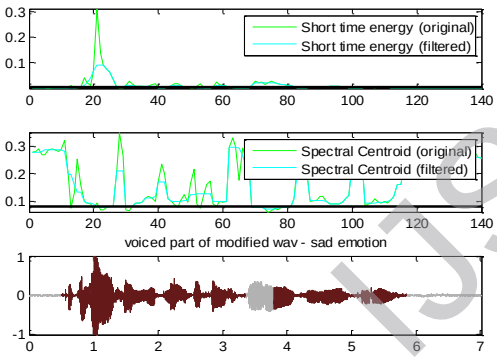

Fig.8:Short term energy and spectral centroide of modified signal of emotion Sad.

Figure 9 shows variation of pitch for modified signal with respect to original signal for emotion happy

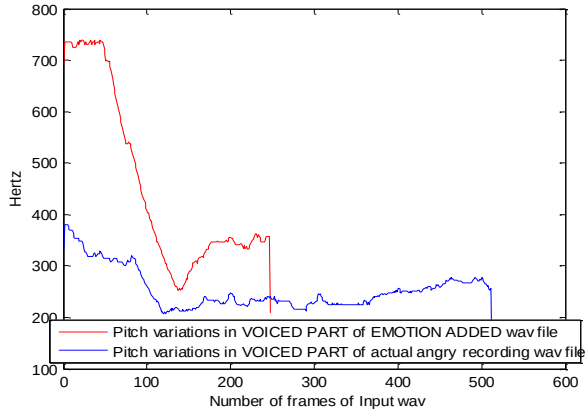

Fig.9: Pitch variations for same database1 with emotion happy.

Fig.10:Short term energy and spectral centroide of modified signal of emotion happy.

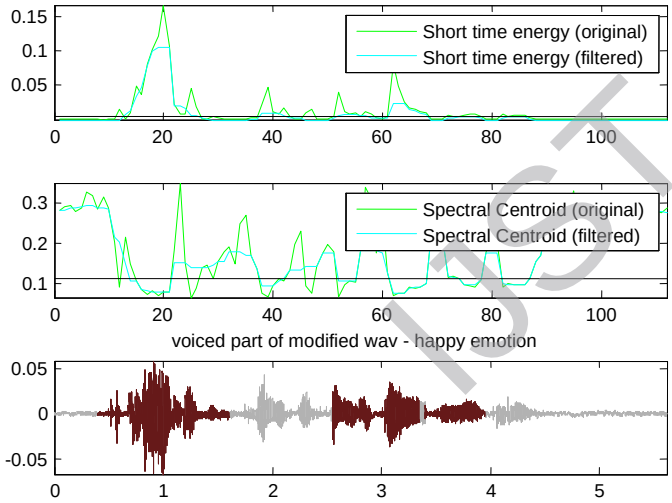

Fig 10:Short term energy, Spectral Centroid & Voiced part of Modified File with Emotion Happy.

Figure 11 Pitch variations of Voiced part of original and modified signal for emotion Surprise

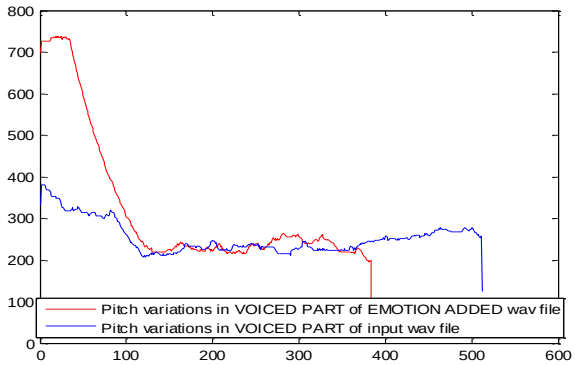

Fig 11:.Pitch variations of Voiced part of original and modified signal for emotion Surprise

Figure 12 shows Short term energy and spectral centroid of modified signal of emotion Surprise.

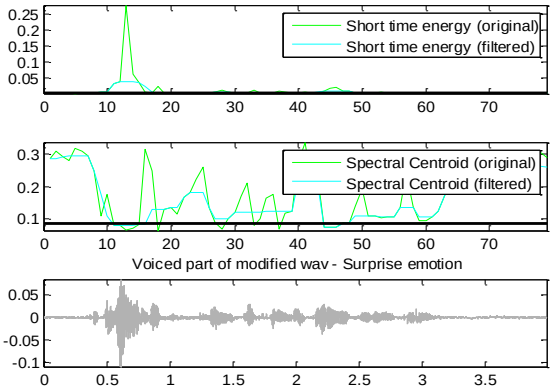

Fig 12:Short term energy and spectral centroid of modified signal of emotion Surprise.

Figure13 shows Pitch veration of Original and modified speech signal for emotion anger.

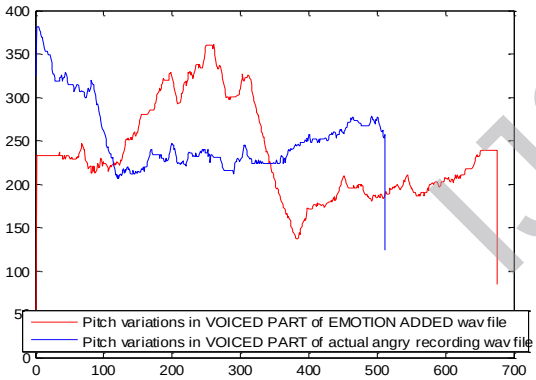

Fig .13: Pitch veration of Original and modified speech signal for emotion anger.

Figure 14 shows Short term energy and spectral centroid of modified Signal with emotion Anger

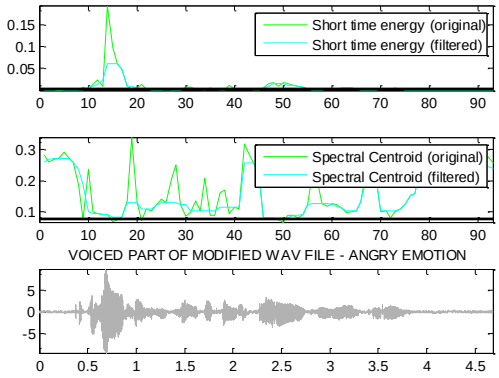

Fig 14:Short term energy and spectral centroid of modified Signal with emotion Anger

CONCLUSION

We have implemented a speech synthesizer system .Work done is divided in to three stages, first pitch mark detection, second voice/unvoice detection ,voice segment detection and third time scale & pitch scale modification using TDPSOLA algorithm.

A speech signal in all emotion(Neutral,Angry,Surprise,Sad,Happy) using same & different Marathi words are recorded .A suitable pitch period finding method is implemented i.e.Autocorelation method to find pitch period.And then pitch mark correspondence is found, and using TDPSOLA,pitch of neutral signal is modified.All these program is done usig MATLAB 10 software.