IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735. Volume 8, Issue 1 (Sep. - Oct. 2013), PP 42-47
www.iosrjournals.org

Pitch Detection of Speech Synthesis by Using Matlab

Abhishek Nandy
(Electronics and Communication Engineering, Abacus Institute of Engineering & Management, India)

Abstract: *In speech synthesis, machine is developed which can accept text and convert into natural sounding speech. Applications of speech synthesis include speech output from computers, reading machine for the visually challenged people. The difference between text to speech synthesizer and any other talking machine (e.g., cassette player) is, it could be trained for any speaker's voice in a fully automatic way. Three main approaches to speech synthesis: articulator synthesis, formant synthesis, and concatenate synthesis. I am carrying out with concatenate synthesis approach. In addition, text-to-speech (TTS) conversion system based on time-domain pitch-synchronous overlap-add (TD-PSOLA) method, has been employed to perform prosody (includes pitch, duration of a speech) modification.[7] To assure good quality of synthetic speech accurate estimation of pitch-period and pitch-marks are necessary for pitch modification. Pitch marking is divided into two tasks; pitch detection and location determination. LPF and some nonlinearity are being used for pitch-detection; peak-valley decision method is used to determine the appropriate parts of speech for used in pitch-mark estimation. In each pitch period, two possible peaks/valleys are searched and one dynamic programming is run to obtain pitch-mark.*

Keywords: *auto-correlation, concatenate, pitch-detection, pitch-mark, TTS*

I. Introduction

A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech. Stephen Hawking is one of the most famous people using speech synthesis to communicate. Concatenative synthesis is based on the concatenation (or stringing together) of segments of recorded speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech. However, differences between natural variations in speech and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches in the output. There are three main sub-types of concatenative synthesis. Concatenate speech synthesis produces speech by concatenating small, pre-recorded units of speech. With longer units, the naturalness increases, less concatenation points are needed. The most widely used units in concatenated synthesis are diphones. At synthesis time, the diphone waveforms are processed through an analysis- synthesis system which is based on a representation of the speech signal by its short-time Fourier transform (STFT) at a pitch synchronous sampling rate. The synthesis part of the system works by overlap-adding the modified short-term signals and it ensures a smooth concatenation of the diphone waveforms. Overlap and add (OLA) methods which have been proposed to manipulate some of the speech parameter while maintain a high degree of naturalness.

In the PSOLA method every signal is correspond to one pitch period of the speech signal. This implies a preliminary segmentation of the speech signal into individual pitch periods. In this method, input speech is represented in speech digital waveform with successive "pitch-marks" distributed along the time scale.

Two analysis windows centered on two successive p-marks should overlap by sufficient amount. The analysis short term signals are obtained by multiplying the signal by the analysis window centered on the corresponding p-marks.

In addition, this method has been used to perform the prosody modification according to the target prosodic information. Prosodic information of a speech includes its pitch (fundamental frequency, f_0), intensity, duration. In TD-PSOLA method it is required to obtain a pitch mark for each pitch period in order to assure a good quality of synthesis speech. If the pitch mark is denoted at the signal with the largest amplitude, but largest peak may not correspond to the largest one in the next period. This is shown in the following figure:

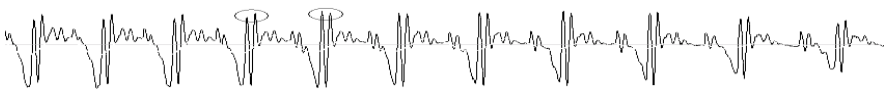


Fig1: It shows that largest peak does not correspond to the largest one in the next period

This will create an unpleasant speech after the method is used. So that in order to determine a pitch mark two highest peaks in each period are searched.

To find out the pitch mark position, I have to first find the pitch period of a speech signal. I use a low pass filter and autocorrelation method to find the pitch of a signal. After that I use an adaptable filter, peak-

valley decision method, and a dynamic programming to determine the pitch mark of a speech signal. With the purpose of spectrally flattening the signal I applied some sort of nonlinearities to the speech signal before the autocorrelation method.

The autocorrelation of a sequence is correlation of a sequence with itself, the autocorrelation of a sequence x(n) is defined by[6]

$$\phi_x(m) \equiv \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)x(n+m).$$

II. Method Used:

2.1 Proposed Algorithm:

I define some symbols which are used in the algorithm:

N: frame size in sample.
$s_m[n]$: the voiced speech of the m[th] frame, where $0 \le n < N$.
$SF_m[k]$: the frequency response of $s_m[n]$, where $0 \le k < N$.
$YF_m[k]$: the pass band frequency response $SF_m[k]$, where $0 \le k < N$.
$o_m[l]$: the adaptive filter's output signal of the m[th] frame, where $0 \le l < N$.

The algorithm is as follows:

Step1: Use FFT to transform the signal $s_m[n]$ to obtain the frequency response $SF_m[k]$.
Step2: Find the position k_p of the spectral peak for $SF_m[k]$ by searching the first forty points of $|SF_m[k]|$.
Step3: Decide on the filter's pass band. Let $YF_m[k] = SF_m[k]$ if $3 \le k \le k_p+2$, otherwise let $YF_m[k] = 0$.
Step4: normalize $YF_m[k]$ by multiplying a scale of $Max_k(|YF_m[k]| / |YF_m[k_p]|)$.
Step5: Use IFFT to transform the normalized $YF_m[k]$ to the time domain.

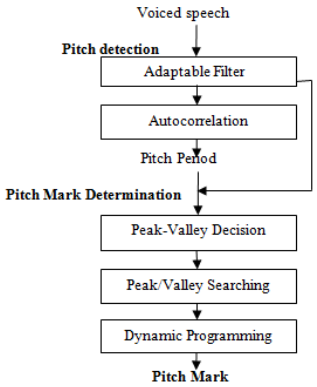
Next step in the pitch mark determination is the peak-valley decision method and dynamic programming.

In the peak-valley decision method I used to calculate the two costs by summing the amplitude of s[q], where q represent the position of the two extreme point of o[·] over each pitch period. In that case o[·] represent the filtered output and s[.] represent the input speech signal.

2.2 Design & Description:

For a synthesis scheme based on the TD-PSOLA method, it is required to obtain a pitch mark for each pitch period in order to assure good quality of synthesis speech. The pitch mark is the reference point for the overlap between the speech signals. There are two major tasks in determination of pitch mark: i) pitch detection ii) location determination. To do that, I follow the design that based on an adaptable filter and a peak-valley method. The block diagram of that method is shown below. In that block diagram I take voiced speech as the input signal because only the periodic parts are of interest. In pitch detection part I use low pass filter and nonlinearity blocks to find out the pitch period of a signal. After filtration of the signal, it goes pass through the autocorrelation block and I get the required pitch period. The required block diagram is shown in section 2.2.1.

2.2.1 Block diagram of the proposed pitch marking method:



III. Simulation Results Obtained:

To implement the structure which is given to the block diagram, I required designing some element namely low pass filter, two nonlinear blocks, and a correlator. Using the MATLAB programming I design the blocks that are required to implement my job.

The first one which I design is a low pass filter, to design a low pass filter; I use some of the following filter specification:

i) FIR, Linear phase, Digital Filter
ii) Pass band of 0 to 900 Hz
iii) Stop band beginning at 1700 Hz
iv) Filter has an impulse response duration of 25 samples
v) The filter pass band was flat to within ± 0.03, and the stop band response was down at least 50 dB[6]

After using those specifications to design a low pass filter, I get the following filter's response graph. In the Y-axis of the graph I plot the filter's absolute magnitude part and in the X-axis I plot the frequency which is normalized in nature.

3.1 Nonlinear Correlation processing:

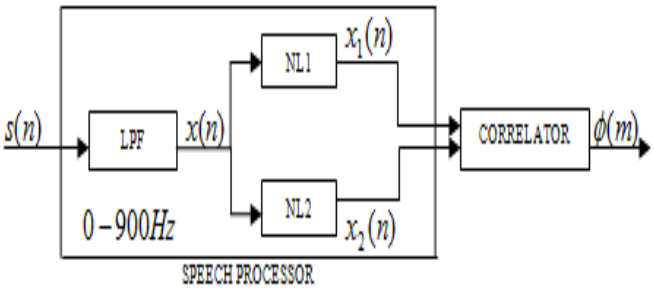


Fig2: Block diagram of non-linear correlation processing

In this block diagram speech signal[S(n)] is first low pass filtered, the output is then used as input to two nonlinear processor, labeled as NL1 and NL2 in the block diagram. The nonlinearities used in each path may or may not be same.[6] It has also been argued that such a correlation will be most appropriate in a variety of actual situations in pitch detection.[2]

3.2 Peak-Valley decision and dynamic programming:

I use the following two equations to calculate the costs:

$$C_{peak} = \frac{1}{N_{peak}} \sum_{n=1}^{N_{peak}} s[Pos_{peak}[n]] \qquad C_{valley} = -\frac{1}{N_{valley}} \sum_{n=1}^{N_{valley}} s[Pos_{valley}[n]]$$

Where the symbols are defined as follows:

C_{peak}= cost estimated at the peaks of o [·].
N_{peak}= total number of the peaks of the o [·].
Pos_{peak}[n] = position of the nth peak of o [·].

C_{valley}= cost estimated at the valley of o [·].
N_{valley}= total number of the valleys of the o [·].
Pos_{valley}[n] = position of the nth valley of o [·].

The peak-valley decision method is made as follows:

If $C_{peak} > C_{valley}$ then the positive part (peak) of s[·] is adopted for the evaluation of the pitch mark. Otherwise, the negative part (valley) of s[·] is adopted for the purpose.

Once the peak or valley is determined, in my case say the peak is adopted for the pitch mark determination. The positions of the pitch mark are set by picking the peaks of the speech signal. The PSOLA (Pitch Synchronous Overlap and Add) method can be used to produce good quality of speech if the pitch is marked at the highest amplitude of the signal. To do the pitch mark we have to calculate the following two quantities.

For the i^{th} pitch period, P_i, suppose the highest and the second highest peaks are located at L_{i1} and L_{i2}, respectively. The distortion of the pitch period, $dis_i\ (j, k)$ and its accumulation, $A_i\ (j)$ are defined as follows:

$dis_i\ (j, k) = ||L_{ij} - L_{(i-1)k}| - P_i| + g(j, k)$, *for i=2, 3...PN,*
$A_i\ (j) = \min\ \{dis_i\ (j, 1) + A_{i-1}(1), dis_i(j, 2) + A_{i-1}(2)\}$, *for i=2, 3....PN,*

Where PN is the total number of pitch period and j, k= 1, 2; $g\ (j, k)$ is a penalty function, defined by $g(j, k) = 0$, if j=1 or k=1 otherwise, $1/PN$

IV. Graphs of Simulation:

Output response of the low pass filter:

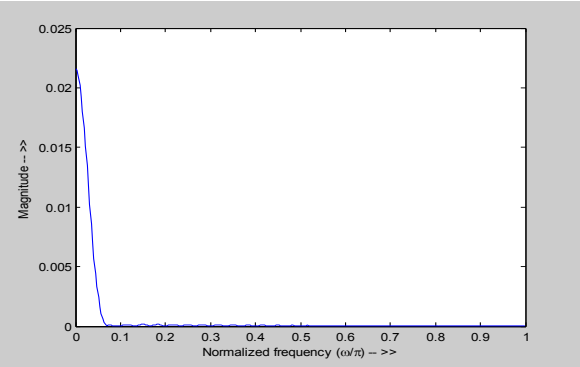


Fig 3: Response graph of the low pass filter .

Input speech and Low pass filtered signal:

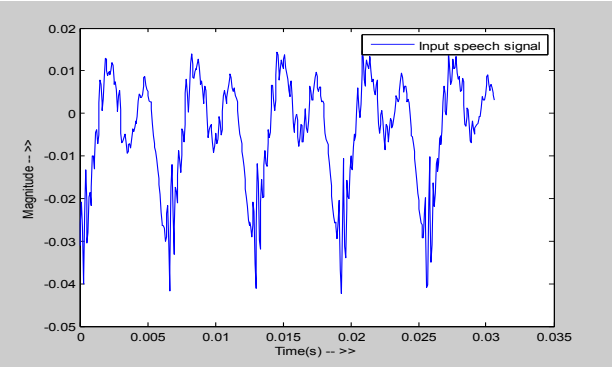


Fig4: Input speech signal.

Non-linearly processed output from the correlation processing:

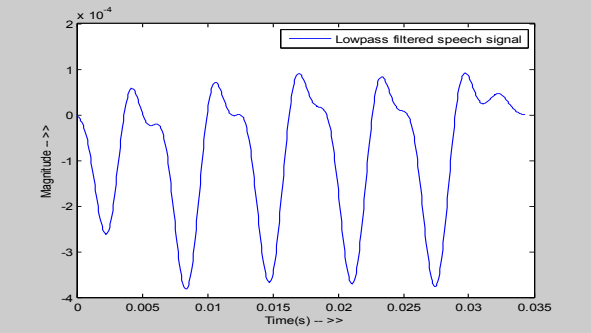


Fig5: speech signal after passing through LPF

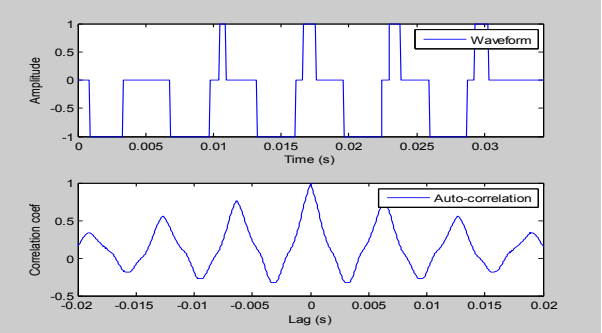


Fig6: nonlinearly processed signal and correlator output

After the correlator block I get the pitch period (f0) of the speech signal. In my work, I get an f0= 155.3398 Hz for a speech signal which I am used in my on-going work.

First figure shows the response of the band pass filter that I am used as my adaptive filter to transform the speech signal into a sin like wave. The speech signal that I am getting after the band pass filter is like smooth sine curves with the high frequency component are cut out. Filtered speech is now passing through a peak-valley decision block to find out which part of the speech is suitable for pitch mark estimation. In fig8 and fig9, the red asterisks denote the positive peaks and the green asterisks denote the valleys of the speech signal. After calculate the cost using my method I see that the positive peak is greater than the negative peak, so I use the positive peak of the speech signal to determine the positions of the pitch marks.

Output response of band pass filter's:

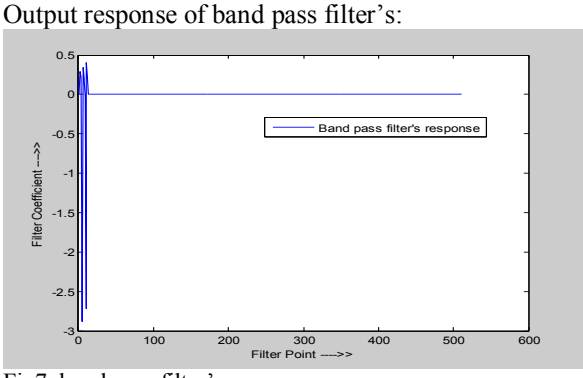


Fig7: band pass filter's response

Peak-Valley detection and pitch mark in a speech signal:

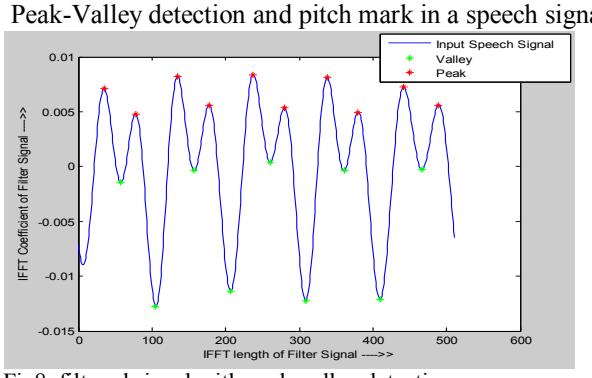


Fig8: filtered signal with peak-valley detection

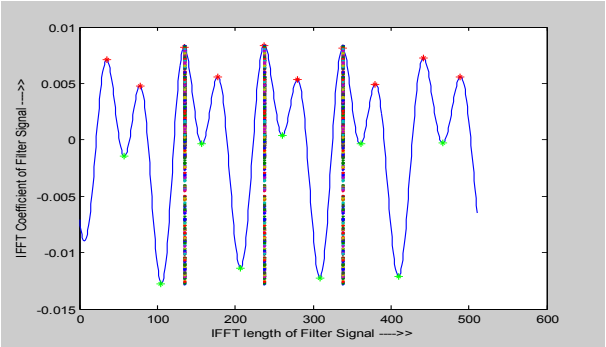


Fig9: speech signal with its pitch mark

V. Conclusion

I have examined method for combining nonlinear processing of the speech signal with a standard correlation analysis to give correlation functions which have sharp peaks at the pitch period. I have seen that the non-linearity provide some degree of spectral flattening which enhance the periodicity peaks in the correlation function, and reduce the correlation peaks due to the formant structure of the waveform. In the following figures I show the difference between unprocessed and non-linearly processed signal:

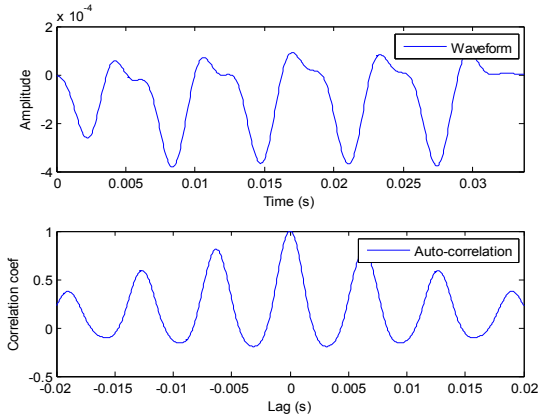


Fig 10: Unprocessed speech signal

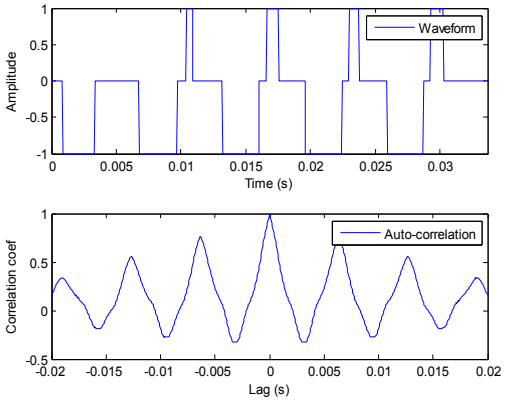


Fig 11: Non linearly processed speech signal

The output of the correlator section gives the pitch of the speech signal which is very important to set the pitch mark of a speech signal. Pitch mark is a reference point for the overlap between speech signals. If I set the pitch mark at the right position it will assure good quality of synthetic speech. To do that a peak valley decision method has been adopted to select either the positive or negative parts of the signal are being used, followed by a dynamic programming to set the pitch mark at the signal.

Acknowledgments:

Though what matters most in a journal are its contents, it is the parts of the whole introduction, methods used, graphs of simulation, simulation results obtained, conclusion etc. that make it an attractive proposition. I have pinned my hopes that the readers would appreciate this approach. I have been fortunate to get help and co-operation from many friends involved in this journal writing effort.

Dr. Anish Deb, Reader & Professor in Calcutta University, Raja bazaar, and Anil Kumar Sharma, Asst. Prof of Abacus Institute of Engineering. & Management, Mogra, took the responsibility of creating the excellencies in formatting the journal paper somewhere. Many thanks to both of you.

And lastly many thanks to Miss Moumita Bachaspati who cheered me in good times, encouraged me in bad times, understood me at all times and for listening to my dreams.

References:

[1] M. M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio Electro Acoust.,* vol. AU-16, pp. 262-266, June 1968.

[2] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electro Acoust.,* vol. AU-20, pp. 367-377, Dec. 1972.

[3] L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a nonlinear smoothing algorithm to speech processing," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-23, PP. 552-557, Dec. 1975.

[4] C. A. McGonegal, L. R. Rabiner, and A. E. Rosenberg, "A semiautomatic pitch detector (SAPD)," *IEEE Trans. Acoust., Speech, and Signal Processing,* vol. ASSP-23, pp. 570-574, Dec. 1975.

[5] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time digital hardware pitch detector," *IEEE Trans. Acoust., Speech, and Signal Processing,* vol. ASSP-24, pp. 2-8, Feb. 1976.

[6] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, and Signal Processing,* vol. ASSP-24, pp. 399-418, oct. 1976.

[7] Lawrence R. Rabiner, Fellow, IEEE, "On the use of Autocorrelation Analysis for pitch Detection," *IEEE Transaction on Acoustics, Speech, and Signal Processing,* vol. ASSP-25, No.1, Feb. 1977.

[8] Jau-Hung Chen and Yung-An Kao, "Pitch Marking Based on an Adaptable Filter and a Peak-Valley Estimation Method," *computational linguistics and Chinese language processing,* vol. 6, No.2, pp. 31-42, Feb. 2001,

[9] Pedro M. Carvalho, Luis C. Oliveira, Isabel M. Trancoso, M. Ceu Viana*, "Concatenative Speech Synthesis for European Portuguese," INESC/IST, *CLUL INESC, Rua Alves Redol, 9, 1000 Lisboa, PORTUGAL {Pedro.Carvalho, Luis.Oliveira, Isabel.Trancoso, mcv}@inesc.pt.

[10] Ramesh Babu, "Digital Signal Processing".