Pitch Period of Speech Signals
Introduction, Determination, and Changing

Mehdi S. Javadi
Sadjad Institute of Higher Education
E-mail: mehdi_s_javadi@yahoo.com

Abstract-This paper, embarks upon sound processing techniques in the field of pitch determination which is an important factor in this connection. In this discussion, with representation and introduction of four pitch tracking methods, we have studied and compared them. These methods are: time- domain waveform similarity, auto-correlation, AMDF, and frequency-domain harmonics' peaks determination method. In resume, we, finally, will be introduced to pitch changing methods for vocoders use. For this, we will introduce instrumental pitch shifting and modified formant pitch shifting methods and their specifications.

1-Introduction:

A large proportion of present-days vocoders are based upon the analysis of a speech signal into an excitation signal and a vocal tract transfer function. Both the excitation and the transfer function are then described in terms of a small number of slowly varying parameters, from which an estimate of the original speech wave is synthesized. There is need for improvement in our description. However, the remarkably small degradation of speech quality in use indicates that the greater need is for an improved parametric representation of the excitation.

Traditionally, the excitation is regarded as consisting of intervals that are either Voiced (V) or Unvoiced (UV). Such a V/UV dichotomy is clearly an oversimplification, as indicated, for instance, by existence of Voiced fricatives. However it is generally accepted that many improvements in our methods of deriving the excitation signal are possible, even without the embellishment of partial voicing

This paper is organized as follows:
In section 2 we will be introduced to the way speech is generated, section 3 describes Voiced Unvoiced decision, some of pitch extraction methods are represented in section 4, and finally in the last section, section 5, pitch changing techniques are discussed.

2-Speech Generation:

Speech generation procedure starts with a flow of air produced by lungs. This flow passes through Glottis Consists of vocal cords. In vowel sounds such as /a/ and /e/, air flow causes these cords to vibrate and a semi-periodic waveform corresponding with Glottis opening is produced. For constants such as /s/ and /f/ the path of vocal cord are open and the source contains a semi- noise spectrum.

Sound frequency is determined with variations in vocal cords' length and protraction. Sound quality is related to sound resonators abode the Glottis (throat and mouth). Sound quality is also controlled by muscles of velum, tongue, cheeks, lips and jaws.

Filter- like specifications which is the responsibility of mouth and throat canals do not have rapid changes and we can estimate speech parameters within a short range of it (10-40 ms). Whenever experiments are based on short- time estimation, speech waveform shows different specifications. As an example by vibrating vocal cords, vowels are produced. Waveform behavior in the case of unvoiced is such that we can estimate it with white Gaussian noise.

3-Segmentation of Voiced/Unvoiced Frames:

For speech signal analysis , special specifications of this signal should be mentioned .In order to achieve this goal ,different segments of speech signal should be classified which is the base of speech signal analysis .Speech signals ,according to their specifications ,are classified into different segments and each segment would be analyzed separately .

Binary classification of Voiced/Unvoiced (V/UV) is a very common method .In this technique, each frame is identified as Voiced or Unvoiced .The main factor of this division is the periodicity of a frame .Voiced frames expose periodic characteristics, while Unvoiced frames are more similar to a random noise.

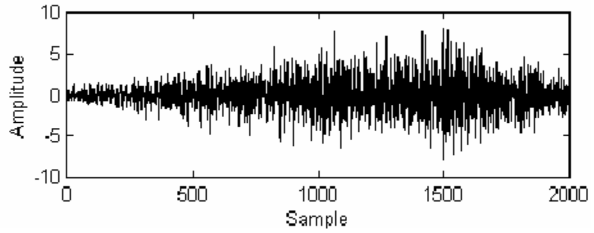

Figure 3-1: Speech signal in UnVoiced segment.

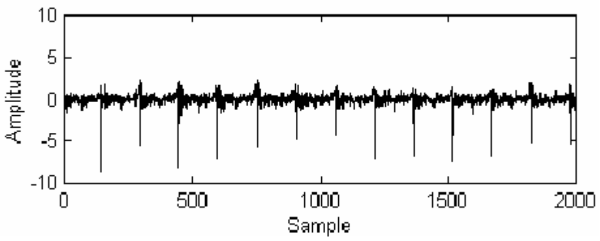

Figure 3-2: Speech signal in Voiced segment.

3-1: Problems Occurred While V/UV Segmentation:

In binary V/UV segmentation, according to the specifications of each frame, the class or group of the frame, as Voiced or Unvoiced, would be determined .Such decision has two difficulties which occurs in (a) transient frames (Voiced to Unvoiced and Unvoiced to Voiced) and (b) frames in which both periodic and noisy parameters are visible (example: /v/ and /z/) .In such cases, binary V/UV decision usually causes unnatural states in process.

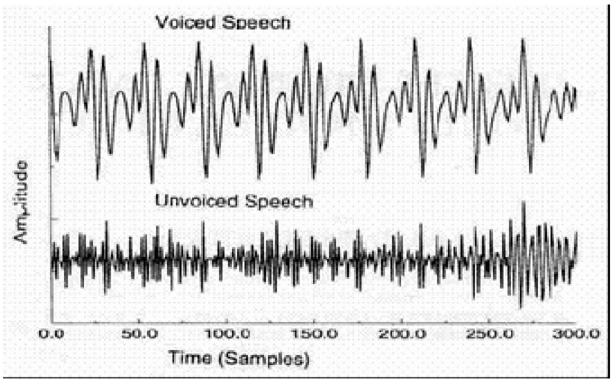

Figure 3-3: Comparison between Voived/UnVoiced speech.

3 – 2 - Main characteristics for V/UV classification:

The main method for segmentation of Voiced speech is using its periodic nature, but also because of specific characteristics of speech signal, other specifications could be used .Most important specifications used in V/UV are listed below:

A – Periodicity: periodicity is the most prominent specification of Voiced speech and could be evaluated in various ways .For example short-time and long-time estimation gain for Voiced speech is greater .periodic signals have strong short-time contacts which could be evaluated by linear estimation coefficients (which are greater) and Voiced signal spectrum has also apparent harmonic structures .

B – Energy Content: energy is also of most important specification which could be used in V/UV segmentation of frames. Generally, energy content of Voiced segments are much more than Unvoiced segments.

Speech signals have low-pass nature. Consequently ,in Voiced segments ,main energy content is hidden in low harmonics .However, this specification for noise-like Unvoiced signal is not considerable .Therefore ,using low frequency to high frequency bands rate could be mentioned as a proper method for Voiced/Unvoiced frames classification .

C – Zero–Crossing: Because of natural limits of base frequency and high energy content of low frequency harmonics of Voiced speech signals, these signals have lower zero-crossing rate in a frame in comparison with Unvoiced speech signals.

D – continuity :Length of Voiced and Unvoiced speech is usually more than length of a frame which is especially obvious for Voiced segments .Therefore ,using this specification and with comparison between current ,previous and next frames ,we can improve the revenue .Rate of changes in period in Voiced segments is also limited .Amount of permitted changes of period in a frame ,therefore ,could be a criterion for Voiced/Unvoiced frames classification .

3–3–Advantages and Disadvantages of Represented Methods:

Although ,all the specifications listed ,can be used in V/UV decision ,effectiveness of each is severely dependant to speech signal's specifications .In other word ,a specification could act much better than others in a frame ,while within a few frames ,other specifications becomes more applicable .As an example ,in a frame ,average of energy ,in general ,is a specification which could be used effectively .But presence of Glottal pulses such as /p/ ,makes using this specification alone difficult .On the other hand ,zero-crossing and high and low bands rate would be effective in many cases .In the case of low energy Voiced signals which are mixed with a little noise ,however ,this two specifications greatly lose their effectiveness .

3 – 4 – Correlation Method for V/UV classification:

Chosen method in this letter for V/UV classification and pitch estimation is based on comparison between a threshold value T (t) and the current value.

Regulations of the threshold value should be low enough to detect the differences in Voiced segments (especially at the beginning of the speech) .And it should be high enough to detect Unvoiced even when random correlation occurs.

In most of the cases , accurate determination of threshold value is very difficult .A threshold value should be chosen which can cope with changes in correlation value for different sounds ,noise and other effective factors .A good strategy is better adaptation in any time which is determined instantly with relation to current pitch periods for current Voiced segments .Two boundary values $T_{low}(t)$ for Voiced and $T_{high}(t)$ for Unvoiced segments are defined .$T_{low}(t)$ is always varying during the algorithm computation until it reaches the maximum of equation below :

$T_{low}(t) = \text{Max}\{T_{min}, T_{max}\}$

Where T_{min} is a constant value and used for general low band and T_{max} is a value relative to maximum of cross-correlation coefficients which is extracted from the current Voiced.

Note that T_{low} (t) never deducts T_{min} and minimum value for T_{low} (t) is equal to general threshold value in T_{min}, but threshold value in Voiced segments increases by correlation values and follows it.

3 – 5 – Practical Results for Correlation Method:

From the computation of floating point , practically ,at sampling rate of 8KHz for proper efficiency ,T_{min}=0.8 ,T_{max}=0.85 and minimum value for threshold in Voiced segment (Max T_{max}) =0.87 ;was chosen .note that ,in order to determine the threshold value ,except accuracy of computations ,sampling rate ,also ,affect the exactness.

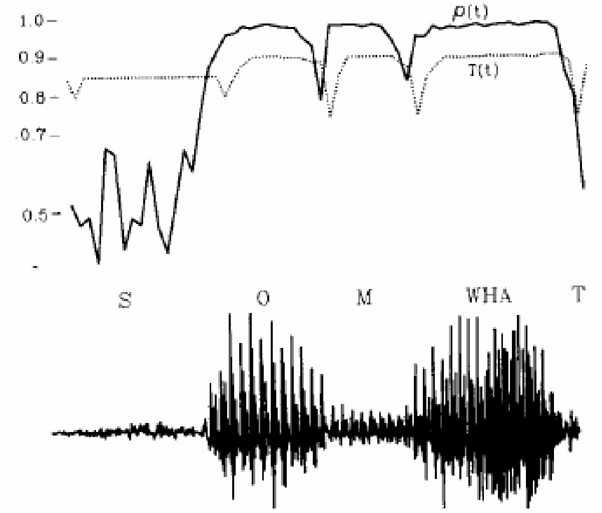

Figure 3-5: Study of adaptable threshold behavior against cross-correlation.

Figure 3-5 studies the behavior of the adaptable threshold with cross-correlation .Threshold T(t) is shown in dotted line ,cross-correlation ρ(t) in solid lines and waveform of the word is given in time domain for comparison .In Unvoiced segment in /s/ ,correlation values increase up until correlation values exceed T_{high} and Voiced region is distinguished .At the same time ,threshold value is transferred to $T_{low}(t)$.$T_{low}(t)$ begins with $T_{min}(0.8)$ and rapidly and due to it's high correlation value ,exceeds from maximum of Voiced speech ρ(t) (=0.87) .Intensity of ρ(t) is not only used for V/UV classification but also for V/V segmentation .While V/V transients , ρ(t) slowly decreases relatively to the obtained value from Voiced .This is illustrated in figure 3-5 for three parts of /o/ ,/m/ and /wha/ from the word "*somewhat*" .This transient fall was followed by a rapid recovery and as soon as confirmation of new Voiced ,reaches to high values of correlation .

Segmentation between consecutive segments of a sound happens when curve of the correlation meets threshold curve .In any division, the threshold decreases, transiently, down to 0.75 to allow classification of sound segments.

If there is a transient between V/UV, then the correlation would be recovered to exceed T_{low} (t) and threshold increases .And when transient is V/V, threshold is set at T_{high} and makes Unvoiced to be detected and segmented .As shown in figure 3-6, in the last single sound of the word "*somewhat*" we have a transient from V to UV.

4-Pitch of Speech signal:

4-1-Pitch determination of speech signals:

Pitch determination is one the most difficult operations in speech processing .Many pitch determination algorithms (PDAs) have been represented ,in both time and frequency domains .

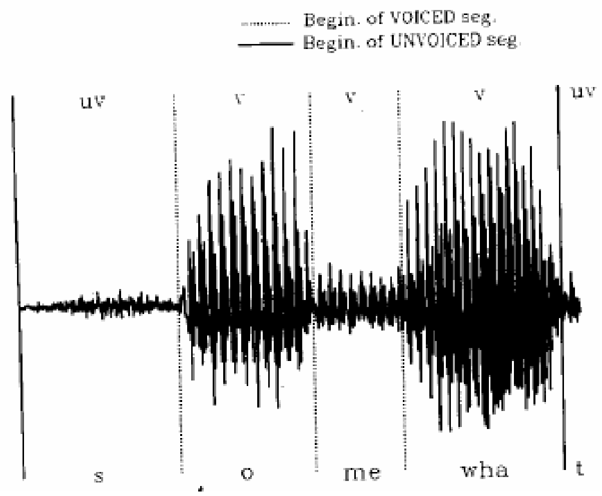

Figure 3-5: Study of adaptable threshold behavior against cross-correlation.

Pitch determination complexity is due to the irregularity and variability of speech signal .Because of reasons listed below, measuring pitch period in an accurate and reliable way is very difficult.

Figure 4-1: Example of a speech signal.

-Impulse waveform of Glottis opening is not a perfect train of periodic pulses .Although finding the periodicity factor of a periodic signal is very simple but measuring the period of speech waveform which is variable in both structure and period, could be very difficult.

-In some cases, vocal system's structure can affect the waveform of Glottis opening such that accurate pitch detection becomes very difficult.

-Accurate and reliable pitch measuring is limited with unseparatable problem in definition of beginning and end of the pitch from Voiced segment of speech.

-Another problem in pitch detection is separation between low-level Voiced and Unvoiced segments of speech .In some cases transients between low-level Voiced/Unvoiced is very fine and therefore scarring between them is very difficult .

Fundamental assumption in this project is :in a short segment (frame) of speech signal ,the value of pitch period is constant and attempts are concentrated on finding this constant value .

Note that the existing stability frequency in signal ,practically ,is limited to values of 50Hz and 400Hz .Therefore ,it is better to cause the speech signal to pass from a low-pass filter ,before applying to these two values .A low-pass filter with low cut-off frequency at 800Hzto 1KHz is satisfying .

Pitch detection algorithms are classified as below:

A–Pitch detectors using time - domain specification.

B–Pitch detectors using frequency - domain specifications.

C–Pitch detectors using both time-domain and frequency-domain specifications.

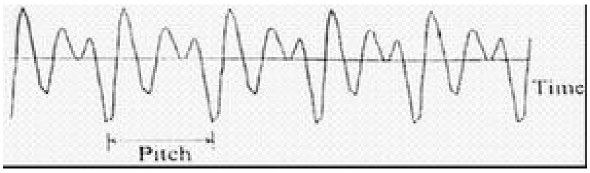

Figure 4-2: A sample of sound signal in Voiced segment.

4-2-Time-Domain Waveform Similarity Model:

One of the specifications of a periodic signal is the interval similarity of waveform in time domain. Fundamentals of PDAs based on waveform similarity are pitch determination using similarity comparison of original signal and its shifted sample. If the shifting interval was equal with pitch, two waveforms should have maximum similarity, which is the basis of most of existing PDAs.

Between these methods, the Auto-Correlation (ac) method and Amplitude Mean Difference Function (AMDF) are two popular cases. Basic idea of waveform similarity method based PDAs is the definition of similarity value. Direct interval measuring is the most current criterion which evaluates similarities between two waveforms and defined as:

$$E(\tau) = 1/N \sum_{n=0}^{N-1} \left[S(n) - S(n+\tau) \right]^{2} \qquad (4\text{-}1)$$

where N the frame length and τ is the shifting interval. Equation 4-1 is based on the assumption of constancy of signal level. This is not, however, true for the beginning of the speech. Therefore, we used Normalized Similarity criterion which takes account of non-stationary signals and defined as:

$$E(\tau) = 1/N \sum_{n=0}^{N-1} \left[S(n) - \text{ß}S(n+\tau) \right]^{2} \qquad (4\text{-}2)$$

where ß is the scaling factor or pitch gain and controls varieties in signal level. Figure 4-1 illustrates a sample of speech signal.

4-3-Auto-Correlation-Based PDAs:

By assuming the signal to be stationary, error criterion 4-1 can be defined as:

$$E(\tau) = \left[R(0) - R(\tau) \right] \qquad (4\text{-}3)$$

where

$$R(\tau) = 1/N \sum_{n=0}^{N-1} S(n).\, S(n+\tau) \qquad (4\text{-}4)$$

In fact, error minimizing, E (τ), in equation 4-1 is equal to maximizing auto-correlation, R (τ), where variable τ is called "lag". In this method function R (τ) is computed for different values of τ and then a value which maximizes R (τ), will be introduced as pitch.

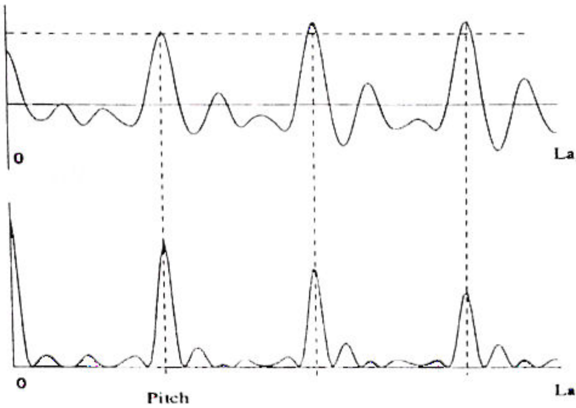

Figure 4-3: Comparing Direct and Normalized Auto-Correlation methods.

In practice, we use 8 KHz as sampling rate, during pitch search, to find out probable values of τ.

4-4-Advantages and Disadvantages of Auto-Correlation Method:

Although auto-correlation computations are consist of multiple multiplications but their implementation in real-time format, due to their regular form (multiplications addition) is very simple.

Now a day, by a single instruction in modern DSPs, multiplications addition is computed.

Another advantage of auto-correlation PDAs is their insensitivity to phase. Therefore, even if there is some degree phase distortion, pitch detection using this method satisfies requests.

Auto-correlation, as mentioned before, is always exposed to the problem of pitch multiple determination. This happens, especially, when speech signal has a sudden change in its energy content and adjacent cycles have considerable changes in their energy content. In this case, a wrong value which is a multiple of true pitch, is chosen as pitch. Figure 4-4 describes this case.

4-5-AMDF PDAs:

AMDF is also a direct similarity criterion which is defined as:

$$E(\tau) = 1/N \sum_{n=0}^{N-1} \left| S(n) - S(n+\tau) \right| \qquad (4\text{-}5)$$

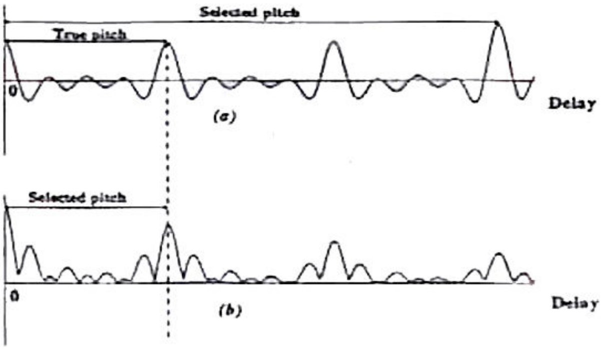

Figure 4-4: Prevail over pitch multiple selection problem using Normalized method.

AMDF, in comparison with auto-correlation function, which is the signal compromising criterion, measures differences. Consequently, it is known as anti-auto-correlation or unsimilarity measure. Figure 4-5 compares AC method with AMDF.

One advantage of AMDF is its computational simplicity. Because the structure of subtraction is very simple compared to the multiplications addition's structure in implementation in microprocessors without multiplier. This advantage has lost its efficiency by introduction of DSPs with integrated multiplier in middle of 1980's. In spite of this, the fact that AMDF computations need less integration is undeniable.

Another advantage of AMDF is its relatively smaller dynamic region narrower valley for stationary signals which pitch tracking methods to become more efficient.

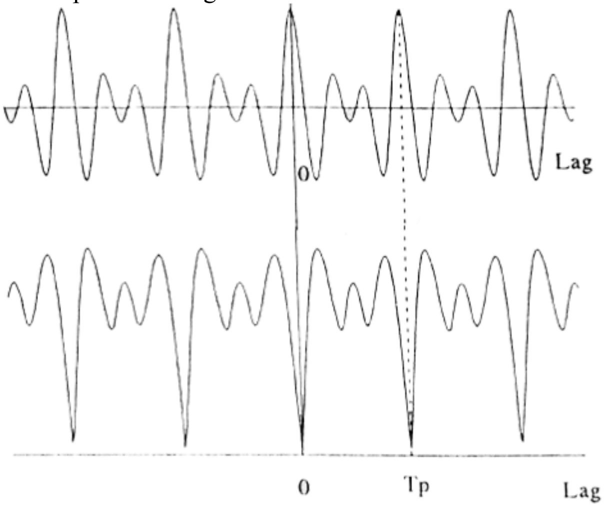

Figure 4-5: Comparison between AC and AMDF methods in pitch determination.

Direct similarity measure was generalized by Nguyen in 1977 as below:

$$E(\tau) = 1/N \sum_{n=0}^{N-1} \left[S(n) - S(n+\tau) \right]^{1/k} \qquad (4\text{-}6)$$

where k is a constant value. Although k could have any value but Nguyen prove that values of 1, 2 or 3 are suitable for k. In his practical experiences, Nguyen indicated that from values above, 2 is best for speech signal. Nevertheless, auto-correlation has preference over AMDF.

As shown in figure 4-1, in long sentences, speech is a non-stationary signal and direct similarity criterion may cause errors, denoting on the fact that in non-stationary signals, shifted signal with shift length of true pitch, has less similarity.

Figure 4-3(a) illustrates the direct auto-correlation function which is indicating more similarity over pitch period with increase in amplitude.

We have used Normalized Auto-Correlation Function to remove the problem of selection of a multiple of true pitch. This function is defined as:

$$R_n{}^2(\tau) = \frac{\sum_{n=1}^{N-1} S(n).\, S(n+\tau)}{\left[\sum_{n=0}^{N-1} S^2(n).\, \sum_{n=0}^{N-1} S^2(n+\tau) \right]^{1/2}} \qquad (4\text{-}7)$$

Where $R_n(\tau)$, is the normalized auto-correlation function. Figure 4-3(b) shows normalized auto-correlation function. It can be seen that, now, the maximum occurs in value of true pitch.

4-6-Frequency-Domain Method of Harmonics Peaks determination:

The mot direct way to period determination from frequency spectrum is to locate first harmonic. This can be performed by locating the lowest peak. But this is possible only when such harmonic exists in signal while this case does not always happen. A more reliable way is to determine all frequencies of peaks and determine the pitch frequency as the interval between adjacent peaks. In order to perform this, we can sample the spectrum for all possible pitch frequencies and add collected and so choose a value that gives the maximum value for true pitch frequency. For this reason, we can use a Comb function for sampling the spectrum. This function is defined as:

$$C(\omega, \omega_0) = \sum_{\beta=0}^{\Omega_0/\omega_0} \delta(\omega - k\omega_0) \quad ; k = 1, 2\ldots \Omega_0/\omega_0$$

where Ω_0 is the maximum existing frequency of spectrum. With this function coefficient in spectrum S (ω) and calculating the total, we can obtain a value for ω which maximizes the total. Figure 4-6 illustrates the state in which pitch frequency is determined by this method.

In this letter, however, we referred only to two common methods in time and frequency domains; we can say auto-correlation method is the most current method for pitch determination and the main reason is, in this method the basic used mathematical operations are multiplication and addition (a multiplication with an addition in each time) which is performed in a single cycle of DSP chips. While in frequency-domain pitch determination methods through Fourier transformation computation is still have more complexity from AC method, even using FFT algorithms.

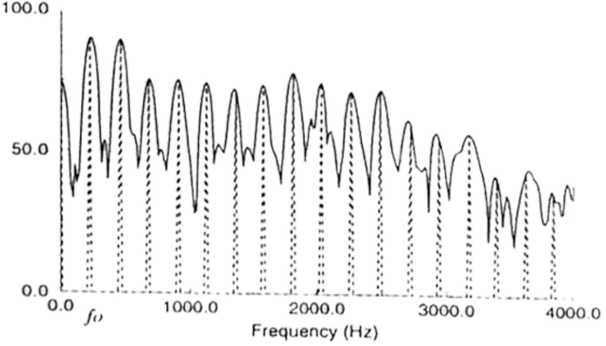

Figure 4-6: Frequency-domain harmonics peaks determination method.

Pitch determination resolution in auto-correlation method is dependant on the sampling frequency. For sampling frequencies of about 50 Hz the resolution varies between 2.5 to 3 percents. For higher resolution, sampling frequency should be increased by Upsampling method. Pitch resolution in frequency-domain methods is dependant to the method applied and accuracy of Discrete Fourier Transform computations.

5 – Pitch Changing Techniques:

5-1-Instumental Pitch Technique:

This algorithm permits instant-time pitch changing which has similar effects with new spectrum sampling. But in the new sampling it have time-domain expansions and contractions which in this case an upsampled speech has higher but shorter pitch and a down-sampled speech has lower but longer pitch. Due to invariability of speed in time domain it is obvious that resampling method can not be used in instant-time form. In instrumental pitch changing, we resample the spectrum in a way in which it does not affect the time axis. This state could be seen in figure 5-1.

In this algorithm, samples are written to a circular buffer and are read from the same buffer with a different sampling rate. Because of asynchronous operation of read and write pointers, therefore, by possible passes of pointer (read or write), a discontinuity may be caused in spectrum.

5-3-Modified Formant Pitch Shifting:

To conform to human speech, we have to change the pitch without changes in in formant frequencies. As can be seen in figure 5-3, harmonics' intervals (pitch) are increased but the spectral envelope is as original.

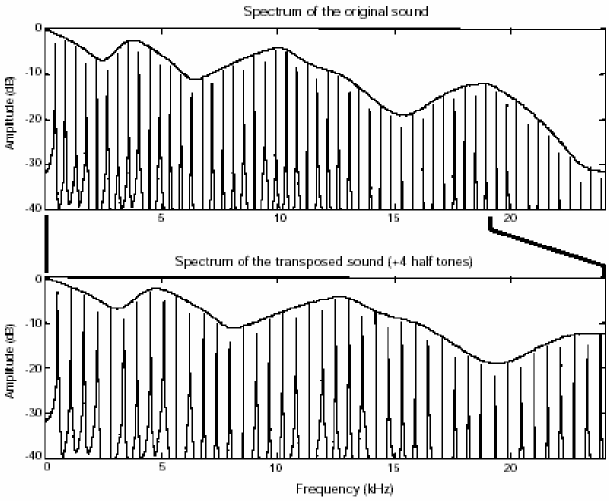

Figure 5-1: Spectrum expansion's effect on speech signal.

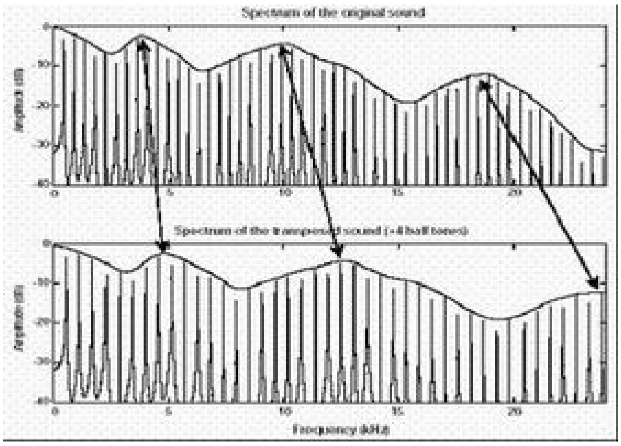

Figure 5-2: Pitch changing of speech signal using instrumental pitch changing.

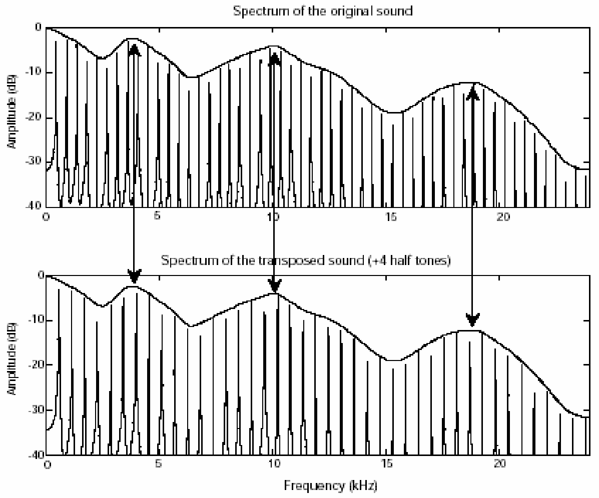

Figure 5-3: Modified formant pitch shifter.

References:

[1] Y. Medan, E. Yair and D.Chazan:” Super resolution Pitch Determination of Speech Signals”. IEEE Trans. Signal processing VOL. 39, NO. 1, PP. 39-48, JAN 1991.

[2] A. M. Kondoz: “Digital Speech” Suray -2001.

[3] W. Hess: “Pitch Determination of Speech Signal-Algorithms and Devices” Berlin-1983.

[4] L. P. Nguyen and S. Imai: “Vocal Pitch Detection Using Generalized Distance Function Associate With a Voiced/Unvoiced Logic”

[5] P. Bastein: “Pitch Shifting and Voice Transformation Techniques”

[6] Y. Medan and E. Yair:” Pitch Synchronous Spectral Analysis Scheme for Voiced Speech” IEEE Trans. Acoust. Speech, Signal Processing, Vol. 37, No. 9, PP. 1321-1328, Sept. 1989

[7] M. M. Sondhi: “New Method of Pitch Extraction" IEEE Trans. Audio Electroacoust, Vol. AU-16, PP. 262-266, June 1968

[8] W. Hess: "Pitch Determination of Speech Signals" New York, Springer, 1983

[9] Y. Medan and E. Yair: "Discrete Spectral Analysis of Periodic Time Function", in Proc., ICASSP-87, Apr.1987, PP. 1797-1800

[10] L. R. Rabiner and R. W. Schafer: "Digital Processing of Speech Signals", Englewood Cliffs, NJ: Prentice-Hall, 1978

[11] W. Dove and C. Myers: "Knowledge-Based Pitch Detection", Boston, ICASSP-83

[12] L. R. Rabiner: "on the use of autocorrelation analysis for pitch determination", IEEE Trans., Audio Electroacoust, Vol. AU-16, PP. 266-270, June 1968

•