Journal of Next Generation Information Technology
Volume 1, Number 2, August, 2010

Sub band Speech analysis using Gammatone Filter banks and optimal pitch extraction methods for each sub band using average magnitude difference function (AMDF) for LPC Speech Coders in Noisy Environments

Suma S.A.; Dr. K.S.Gurumurthy
University Visvesvaraya College of Engineering, K.R. Circle Bangalore 01, India.
Email: suma8055@yahoo.com;drksgurumurthy@gmail.com

doi: 10.4156/jnit.vol1.issue2.2

Abstract

Modern speech processing applications require operation on signal of interest that is contaminated by high level of noise. These situations call for a greater robustness in estimation of the speech parameters for mismatch environment and low environmental SNR level. In this paper the speech is analyzed with a Gammatone filter bank. This splits the full band speech signal s(n) into frequency bands(sub bands).and for each sub band speech signal pitch is extracted. We determine the Signal to Noise Ratio for each Sub band speech signal. Then the average of pitch periods of the highest SNR sub bands is used to obtain a optimal pitch value. This paper describes a computationally simple Pitch extraction algorithms using Average Magnitude Difference Function (AMDF) which is a new approach using weighted Autocorrelation [2] and very useful for accurate Pitch Period extraction. Both these algorithms can be software implemented and performance evaluated. Both of them uses center clipping for time domain processing. This paper also in general Compares the effectiveness of the new AMDF using weighted Autocorrelation and the existing Autocorrelation method and how it is possible to utilize this further in Speech Enhancement Systems using the proposed new algorithms for its implementation

Keywords: *Speech, Pitch extraction, Linear predictive coding (*LPC*), Noisy Environments, Average Magnitude Difference Function(AMDF), Weighted Autocorrelation, Gammatone filter banks*

1. Introduction

Many principles have been proposed for the modeling of human pitch perception and for practical pitch determination of speech signals [1]–[3]. For regular signals with harmonic structure, such as clean speech of a single speaker, the problem is solved quite reliably. When the complexity increases further, e.g., when harmonic complexes of sounds or voices are mixed in a single signal channel, the determination of pitches is generally a difficult problem that has not been solved satisfactorily. The concept of pitch refers to auditory perception and has a complex relationship to physical properties of a signal. Thus, it is natural to distinguish it from the estimation of fundamental frequency and to apply methods that simulate human perception. Many such approaches have been proposed and they generally follow one of two paradigms: place (or frequency) theory and timing (or periodicity) theory. Neither of these in pure form has been proven to show full compatibility with human pitch perception and it is probable that a combination of the two approaches is needed. Also modern speech processing applications require operation on signal of interest that is contaminated by high level of noise. These situations call for a greater robustness in estimation of the speech parameters for mismatch environment and low environmental SNR level. In this paper the speech sound is filtered by Gammatone filter and for each sub band speech signal pitch is extracted. We determine the Signal to Noise Ratio for each Sub band speech signal. Then the average of pitch periods of the highest SNR sub bands is used to obtain a optimal pitch value. PITCH Period (i.e., fundamental frequency fo and period, To – 1/fo) is an important parameter of speech signal, which is used in speech analysis, synthesis and recognition. For speech recognition applications, the pitch extraction algorithm provides the basis for voiced/unvoiced classification decision. Other than voicing information, the pitch extraction algorithm provides prosodic information such as stress and intonation. The accuracy of pitch extraction is directly related to the quality of speech. Thus, we need to extract the pitch of speech signals in practical noisy environments for most of the applications. Pitch extraction methods are classified into the following three categories; (a) waveform processing, (b) Spectral processing and (c) correlation processing known to be comparatively robust against noise. Auto Correlation function method of category (c), is one of the conventional methods being used for Pitch determination, which provides the best Performance in noisy environments. Since the Average Magnitude Difference Function (AMDF) has similar characteristics with the Auto Correlation function, a new pitch extraction method, which uses an Auto Correlation function weighted by the inverse of an Average Magnitude Difference Function (AMDF) [2], can be implemented in Linear Predictive Coding Schemes. The characteristics of the AMDF are very similar with those of the autocorrelation function. The Average Magnitude Difference Function (AMDF) produces a notch, while the autocorrelation function produces a peak. However, both functions essentially have the same periodicity. The new AMDF method using weighted Autocorrelation [2] utilizes the feature that in a noisy environment, the noise components included in the autocorrelation function and AMDF behave independently (and are uncorrelated each other). By such uncorrelated properties, the peak of the autocorrelation function is emphasized in a noisy environment when the autocorrelation function is combined with the inversed AMDF. As a result, it is expected that the accuracy of pitch extraction for the AUTOC is improved. This paper describes two computationally simple pitch extraction algorithms using the new AMDF for pitch determination. For ease of presentation, these algorithms will be identified throughout this paper as #1 and #2. Algorithm #1 uses center clipping and infinite peak dipping for time domain preprocessing before computing AMDF while Algorithm #2 nonlinearly distorts the speech signal before center clipping and AMDF computation. In fact Matlab results shows that #2 provides a better pitch detection estimate than #1. The results obtained by comparing the average gross pitch error rate suggest that #2 is better than #1. Both the methods are computationally simple and more reliable in noisy environments. With the growth of wireless DSP Processors where hardware is well molded for specific applications such Computationally simple algorithms becomes simple to implement and leads to cycle count gain ensuring reliability in noisy environments. The organization of the paper is as follows. Section 2, we describe the LPC-10 Speech coder. Section 3, we describe the gammatone filter bank used for sub band speech analysis. Section 4, we describe the subband speech processing. Section 5, describes the principle of the new weighted autocorrelation method by inverse of AMDF. Section 6, describes the proposed new computationally simple algorithms for the weighted autocorrelation method by inverse of AMDF. Section 7, describes Experiments and Results. Finally, we conclude this paper in Section 8.

2. LPC-10 Speech Coder

The LPC-10 speech coder is the US standard for linear predictive coding of speech at 2400 bits per second. It used the analysis-by-synthesis technique, which based on the 10th order lattice filter, to create the prediction parameters. It employs linear predictive coding (LPC) shown in Figure 1, that models the short-term spectral information as an all-pole filter which captures the Power spectral density (PSD) of the speech signal. The speech output from the LPC model is not acceptable for many applications because it does not provide sound like human speech. Usually it is applied in military applications, which do not require high quality speech but need low bit rate. However most of the modern speech coder operating principle is derived from the LPC model with modifications to improve quality and coding effectiveness. The LPC model is inspired by observations of the basic properties of speech signals and represents an attempt to mimic the human speech production mechanism which is shown in Figure 1. The combined spectral contributions of the glottal flow, the vocal tract, and the radiation of the lips are represented by the synthesis filter. The driving input of the filter or excitation signal is modeled as either an impulse train (voiced speech) or random noise (unvoiced speech). Therefore, depending on the voiced or unvoiced state of the signal, the switch is set to the proper location so that the appropriate input is selected. Energy level of the output is controlled by the gain parameter. At the analysis stage (encoder), the four important parameters pitch period, power, voicing information and LP coefficients are computed and only these parameters are appropriately coded and transmitted. With these parameters it is possible to reconstruct the speech signal at the decoder to reproduce the speech signal. Estimating the parameters precisely is the responsibility of the encoder. The decoder takes the estimated parameters and uses the speech production model to synthesize speech. Though the synthesized speech waveform is slightly different from the original wave, but since the power spectral density (PSD) of the original speech is captured by the synthesis filter, PSD of the synthetic speech is close to the original speech.

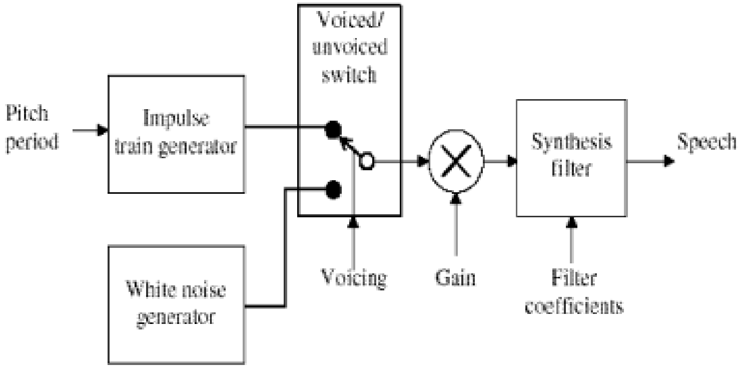

Figure 1. LPC model that synthesize the speech signal with four inputs: Pitch Period, voicing, Gain and Prediction Coefficients.

3. Gammatone Filter Bank

Gammatone filters can be implemented using FIR or IIR filters or frequency domain techniques. In this research, FIR filters can be employed in order to implement linear phase filters with identical delay in each critical band. The analysis filters had a length of 2N-1 coefficients, and were obtained by convolving a sampled gammatone impulse response g(n) of length N = 100 with its time reverse, where

$$g(n) = a(nT)^{N-1} e^{-2\pi bERB(f_c)nT} \cos(2\pi f_c nT + \varphi) \tag{1}$$

fc is the centre frequency, T is the sampling period, n is the discrete time sample index, a, b, p, f are constants, and ERB (fc) is the equivalent rectangular bandwidth of an filter. At a moderate power level, ERB(fc) = 24.7 + 0.108 fc .Examples of the impulse responses of these filters are shown in Figure. 2.

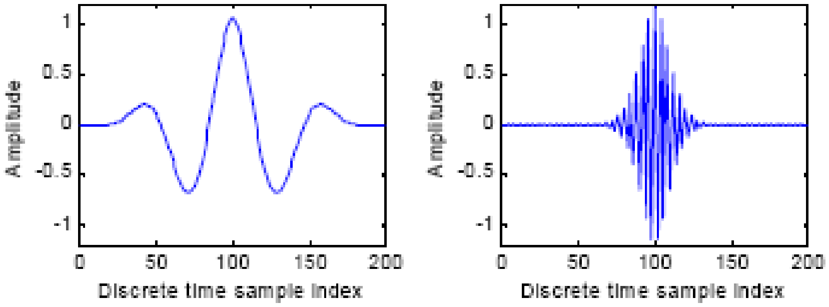

Figure 2. Impulse responses of the (a) 3rd (centre frequency 250 Hz) and (b) 18th (centre frequency 4 kHz) critical band linear phase gammatone filters.

The gammatone filter bank employed in this approach contains gammatone filters Hi(z) whose centre frequencies and bandwidths match those of the critical bands. Thus, for an 8 kHz signal bandwidth, 21 filters were used.

4. Subband Speech Processing

The Speech signal s(t) is applied to the gammatone filter bank described in the previous section. The sub band speech $s_1(t)$ to $s_{21}(t)$ are the outputs of the Gammatone filter bank as shown in Figure.3.

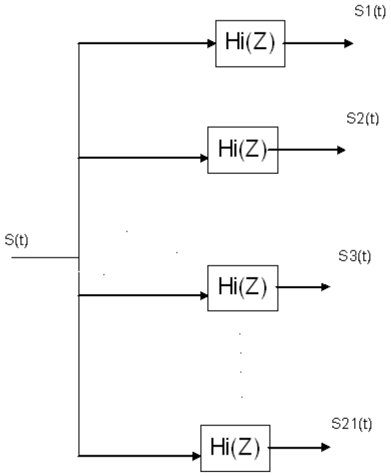

Figure 3. Subband speech processing.

The SNR estimation is used for weighing the sub-bands. Noise Power is estimated as an average of the non speech frames in each utterance. Whether a frame is speech or non-speech is determined by comparing the current frame energy with the average frame energy of the first 10 frames in input test speech. The full band SNR of frame can be simply computes as:

$$SNR_t^{Full} = 10 \log_{10} \left[\frac{\sum_{k=1}^{K} |s_t(k)|^2}{\sum_{k=1}^{K} \left| \bar{N}(k) \right|^2} \right] \tag{2}$$

$$|s_t(k)| = \max \left\{ |x_t(k)| - \alpha \left| \bar{N} \right|, \beta \left| \bar{N} \right| \right\} \tag{3}$$

Where k, $|x_t(k)|$, $|s_t(k)|$, and $\bar{N}$ are *the frequency Index, the magnitude spectrum of noisy speech, that of estimated clean speech, and the average magnitude spectrum of noise, respectively. In order to compute*
the SNR, the magnitude spectrum of clean speech has to be estimated. We use the spectral subtraction method, the overestimating factor a *subtracts an overestimate of the noise power spectrum from the noisy speech power spectrum* in order to minimize the presence of residual noise, and the spectral flooring factor β prevents the spectral components of estimated clean speech from falling below the lower value, $\beta \left| \bar{N}(k) \right|$. The values of overestimating and spectral flooring factors are set to 1.1 and 0.001 empirically. From Eqs. (2) and (3), the sub-band SNR can be easily obtained as

$$SNR_t^i = 10\log_{10}\left[\frac{\sum_{k\in Sub-bandi}\left|S_t(k)\right|^2}{\sum_{k\in Sub-bandi}\left|\bar{N}(k)\right|^2}\right] \tag{4}$$

Where i is the sub-band index. From the SNR obtained, the full-band or sub-band weight coefficient, ρ_t^{Full} or ρ_t^i, is calculated by applying a sigmoid function to full-band or sub-band SNR as

$$\rho_t^{Full} = \frac{1}{1+\exp[-0.5(SNR_t^{Full}-\eta)]} \tag{5}$$

$$\rho_t^i = \frac{1}{1+\exp[-0.5(SNR_t^i-\eta)]} \tag{6}$$

Figure 4. shows the plots of weight from Eqs. (5) and (6) depending on η .

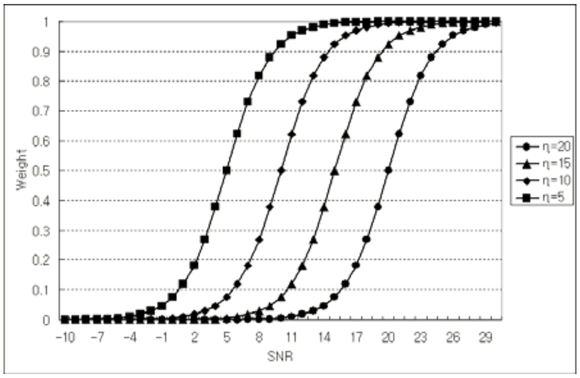

Figure 4. Plots of weights depending on h

Sub band SNR is determined for all the sub bands $s_1(t)$ to $s_{21}(t)$ and the average of pitch periods of the highest SNR subbands is used to obtain a optimal pitch value.

5. Weighted Autocorrelation Method by Inverse of AMDF

5.1. Principle

The autocorrelation function of a periodic signal is given by

$$\phi(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+k) \tag{7}$$

where
x(n) is the speech signal;
k is the lag number;
n is the time for the discrete signal;

The autocorrelation function representation of the signal is a convenient way of displaying certain properties of the Signal. If the signal is periodic with period P samples, then

$$\phi(k) = \phi(k+P) \qquad \text{for } 0, \pm P, \pm 2P, ... \tag{8}$$

Also it is an even function i.e.

$$\phi(k) = \phi(-k) \tag{9}$$

It attains a maximum value at k=0; i.e.,

$$|\phi(k)| \le \phi(0) \quad \text{for all k.} \tag{10}$$

Considering the above properties of autocorrelation for periodic signals, we can see that the function attains a maximum at samples *0, P, 2P* where P is the pitch period. Let us assume that x(n) is a noisy Speech signal given by x(n)=s(n)+w(n) (11)

where s(n) is a clean speech signal and w(n) is a white Gaussian noise.

$$\phi(k) = \frac{1}{N}\sum_{n=0}^{N-1}[s(n)+w(n)][s(n+k)+w(n+k)]$$

$$\phi(k) = \frac{1}{N}\sum_{n=0}^{N-1}\begin{matrix}[s(n)s(n+k)+s(n)w(n+k)+ \\ w(n)s(n+k)+w(n)w(n+k)]\end{matrix} \tag{12}$$

In equation (12) we can see that Auto Correlation and Cross-Correlation of s(n) and w(n) is done. For large N, if s(n) is not cross-correlated with w(n) and w(n) is not self-correlated except for k=0

then $\phi(k) = \frac{1}{N}\sum_{n=0}^{N-1}[s(n)s(n+k)]$ for $k \neq 0$

and $$\phi(k) = \frac{1}{N}\sum_{n=0}^{N-1}[s(n)s(n+k)+w(n)w(n+k)] \quad \text{for k=0} \tag{13}$$

Thus Auto Correlation function provides robust performance against noise. Now let us come to AMDF. It is described by

$$\psi(k) = \frac{1}{N}\sum_{n=0}^{N-1}\left|x(n)-x(n+k)\right| \tag{14}$$

Now considering equations (7) to (13) we can see that the maximum peak is located at k=P except for cases of k=0. But in some cases the peak located at k=2P becomes larger than that at k=P as shown in Figure (5). Then a half pitch error occurs. Also there is a peak at k<P. This situation in some cases leads to a double pitch error. Thus the accuracy of pitch extraction using Autocorrelation becomes higher if unnecessary peaks are suppressed. In case of AMDF as in equation (14) $\psi(k)$ becomes smaller when $x(n)$ is similar to $x(n+k)$ i.e. if $x(n)$ has a period P,

$y(k)$ produces a notch at k=P as shown in Figure (5) i.e.

$1/y(k)$ makes a peak at k=P. Now if we substitute

equation (11) in (14) it reduces to

to $$\psi(k) = \frac{1}{N}\sum_{n=0}^{N-1}\left|s(n)+w(n)-s(n+k)-w(n+k)\right| \tag{15}$$

i.e. $$\psi(k) = \frac{1}{N}\sum_{n=0}^{N-1}\left|s(n)-s(n+k)\right| + \frac{1}{N}\sum_{n=0}^{N-1}\left|w(n)-w(n+k)\right| \tag{16}$$

i.e. there is an AMDF for speech signal and an AMDF for w(n). We see that the noise component is obviously independent as compared to the Autocorrelation function seen in (13). Hence, using the Autocorrelation function weighted by $1/y(k)$, it is expected that true peak is emphasized, and as a result the errors of pitch extraction are decreased. So we can define a new function which is given by

$$\eta(k) = \phi(k) / (\psi(k) + \tau) \tag{17}$$

where τ is a fixed number (τ >0). The AMDF in equation (14) at k=0 is $\psi(0) = 0$ and therefore the denominator is stabilized in equation (17) by adding the number τ .

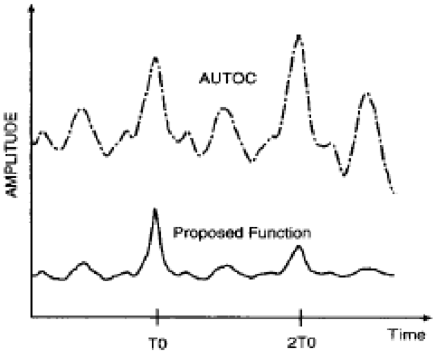

Figure 5. Autocorrelation function and proposed function. T0 corresponds to the pitch period.

Figure. 5 shows the autocorrelation and proposed functions obtained for a speech signal corrupted by noise. In this case, by picking the maximum amplitude of each function, the proposed function leads to the true pitch, while the autocorrelation function does an erroneous one. Now the new computationally simple algorithms #1 and #2 proposed to implement this are as follows.

6. Proposed New Algorithms of the New Weighted Autocorrelation Method by Inverse of AMDF

6.1 Algorithm #1

The technique of removing the formant structure for reliable pitch detection by center clipping was shown by Sondhi while retaining periodicity(pitch period information). Figure 6 shows the block diagram of the pitch extraction algorithm.

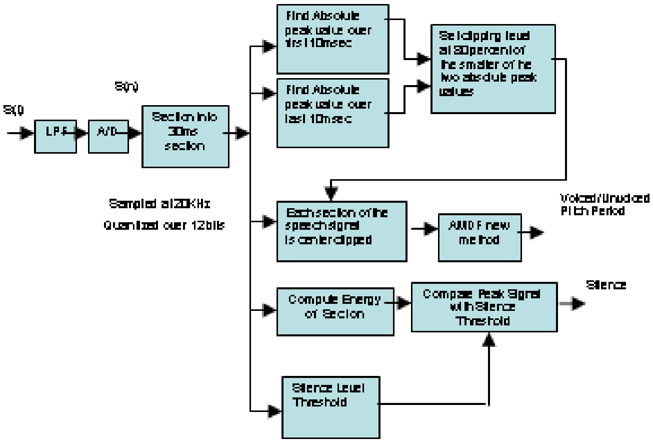
S(t)
LPF
A/D
S(n)
Section into 30ms section
Sampled at 20KHz
Quantized over 12bits
Find Absolute peak value over first 10msec
Find Absolute peak value over last 10msec
Set clipping level at 30percent of the smaller of the two absolute peak values
Each section of the speech signal is center clipped
AMDF new method
Voiced/Unvoiced Pitch Period
Compute Energy of Section
Compare Peak Signal with Silence Threshold
Silence
Silence Level Threshold

Figure 6. Block diagram of the pitch extraction Algorithm #1

The Speech signal s(n) is sectioned in to Overlapping frames of 20-30 ms. Duration with 50 percent overlap between adjacent frames. There is always a potential loss of information during voicing transition and voiced/unvoiced boundary and hence the signal needs to be clipped accurately. For this, the maximum absolute peak levels for the first and the last 10ms sections of the speech are determined and the clipping level is set at 80 percent of the smaller of the two absolute peak values. Then, each section of the speech signal is center clipped which is given by

$$s_c(n) - \begin{cases} s(n)+c_t, & s(n)\leq -c_t \\ 0, & -c_t\leq s(n)\leq +c_t \\ s(n)-c_t, & s(n)\geq +c_t \end{cases}$$

So, the speech section is infinite peak clipped, resulting in a signal, which takes three possible values; -1 if it falls below the negative clipping level ($-c_t$), +1 if the sample exceeds the positive clipping level ($+c_t$) and 0 otherwise. The samples are thus reduced to three levels so that the computational and hardware complexity is reduced. Next the AMDF is performed for each section and amplitude normalized. From equation (17) the AMDF is given by

$$\eta(k) = \frac{\frac{1}{N}\sum_{n=0}^{N-1} s_c(n)s_c(n+k)}{\left(\frac{1}{N}\sum_{n=0}^{N-1} |\, s_c(n) - s_c(n+k)\,| + \tau\right)}$$

k=0,1,………, M (18)

where N is the number of samples in the speech section and k is the lag number. The normalized AMDF is given by

$$\hat{\eta}(k) = \eta(k) / \eta(0) \tag{19}$$

Since each individual product and difference term can have only 3 values +1, -1, or 0, a simple up/down counter and a differential combinational logic circuit is only desired to perform the computation in equation (18). The use of such a weighted function for picking the maximum amplitude as the peak and the corresponding position of this peak gives the pitch period. Usually in speech processing the signal is windowed (rectangular or hamming window) so that the peaks are gradually tapered to zero with the peak at the maximum level at the fundamental frequency and reduced amplitude levels at the harmonics. Other than locating the pitch period, each section can be classified as voiced/unvoiced by comparing the correlation peak value at the pitch period to a predetermined threshold value. If the value exceeds the threshold then the section is classified as voiced otherwise as unvoiced. Based on peak signal levels, a silence level threshold is chosen in all LPC Coders. By computing the energy for each section given by

$$E = \sum_{n=0}^{N} s^2(n) \tag{20}$$

By comparing this to the silence threshold, each section is classified as background noise/speech. The accuration of this pitch determination in noisy environments can be further enhanced by doing interpolation on 3 points around the detected peak. The interpolation method used in this paper was

Lagrange's method. The infinite peak clipping of the speech signal is done in the range of 50 Hz to 400 Hz which corresponds to the region of fundamental frequencies for most men and women.

6.2 Algorithm #2

The setting of the clipping level threshold in the previous algorithm is sensitive to pitch detection and based on extensive computer simulation study. So a non-linear distortion of the speech sections eliminates the need to adjust critically the clipping level.

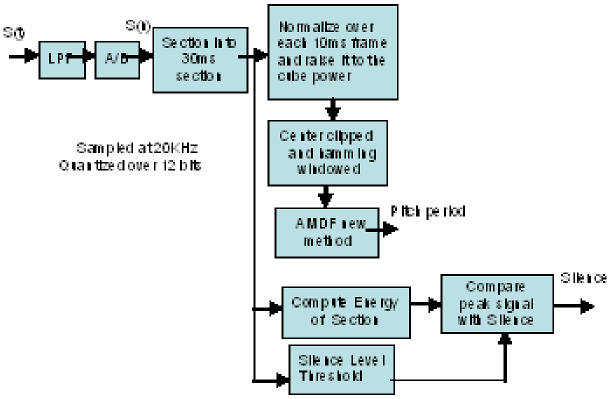
S(t)
LPF
A/D
S(t)
Section into 30ms section
Normalize over each 10ms frame and raise it to the cube power
Center clipped and hamming windowed
AMDF new method
Pitch period
Sampled at 20KHz
Quantized over 12 bits
Compute Energy of Section
Compare peak signal with Silence
Silence
Silence Level Threshold

Figure 7. Block diagram of the pitch extraction Algorithm #2

The clipping level is set at 50 percent of the peak absolute value for each frame. The speech signal is again sectioned in to overlapping Frames of 30 ms duration with 50 percent overlap between adjacent frames. Figure 7 shows a general block structure of this algorithm. The amplitude is normalized to unity for each 10 ms section and then the signal is nonlinearly distorted by raising its cube power. By raising the signal to some high power before applying weighted autocorrelation method highlights the high amplitude peaks while suppressing low amplitude peaks .

Each signal section is then center clipped as in algorithm #1 and then windowed using hamming window before performing weighted autocorrelation. As in algorithm #1, this also distinguishes voiced/unvoiced speech section and also background noise from speech section other than pitch detection.

7. Experiments and Results

For our experiments, 100 male and 100 female speakers was selected from TIMIT database. A universal background model (UBM) is also trained from other 50 male and 50 female speakers in theTIMIT database. For noise data, we down-sampled TIMIT database from 16kHz to 8kHz and artificially added noise to clean test speech with various SNRs. The speech analysis frame rate is set to 20ms with 10ms interval. The UBM and speaker models contain 160 Gaussian components respectively. The performances of speaker identification according to sampling rates in clean condition are provided in Table 1.

Table 1. Performances of speaker identification according to sampling rates in clean condition

Sampling Rate	8 kHz	16 kHz
Accuracy (%)	93.9	99.6

The tests suggested that using gammatione filtering for the sub-banding resulted in an improvement in the output SNR of up to 3dB compared with the linear case, when speech was used as the signal. This improvement was observed using both white and speech shaped noise as the corrupting noise signal. In this test, six corrupting noise signals were used; white and speech shaped noise at high, medium and low SNRs. Then pitch for each sub band was determined using weighted Autocorrelation by inverse of AMDF .The speech data was clipped after every 30ms with 50 percent overlap between adjacent frames and then the amplitude was normalized and new AMDF method was applied. The test results on simulation using both the algorithms was able to detect the pitch peaks for different sections of frames as the high amplitude peaks had suppressed low amplitude peaks. Results are given in percentage gross pitch error (%GPE). If any estimated pitch is not within 1 ms of the reference pitch, then it is termed as gross error. %GPE is provided for both male and female speech. Tables 1 and 2 show %GPE of the proposed pitch detection Algorithms for both female and male speech, respectively, at SNR = 20 dB, 15 dB, 10 dB, 5 dB and 0 dB. A Conventional Pitch Extraction algorithm using Autocorrelation was used for comparison with the new proposed algorithms. From the tables we can see that the proposed algorithms outperforms the conventional pitch extraction algorithm with Autocorrelation. Also we can see that Algorithm #2 offers improved % GPE compared to Algorithm #1. For example, in SNR = 0 dB, the proposed algorithms improves %GPE from 43.75%, obtained by the conventional algorithm, to 25.22%(Algorithm #1) and 18.34%(Algorithm #2) for female speech, and from 40.74% to 22.22%(Algorithm#1) and 16.53%(Algorithm #2) for male speech.

Table 2. Performance comparison of the proposed algorithms and conventional algorithm using Autocorrelation in terms of global pitch error (%GPE) for female speech.

SNR	20 dB	15 dB	10 dB	5 dB	0 dB
Proposed Algorithm #1	6.25	12.51	16.22	18.75	25.22
Proposed Algorithm #2	5.23	9.21	12.45	15.31	18.34
conventional Algorithm using Autocorrelation	12.50	25.32	31.25	38.01	43.75

Table 3. Performance comparison of the proposed algorithms and conventional algorithm using Autocorrelation in terms of global pitch error (%GPE) for male speech.

SNR	20 dB	15 dB	10 db	5 dB	0 dB
Proposed Algorithm #1	7.40	12.01	14.81	18.51	22.22
Proposed Algorithm #2	6.17	8.98	10.02	14.23	16.53
conventional Algorithm using Autocorrelation	11.11	14.81	18.51	25.92	40.74

An example test simulation carried out for female speech at SNR =0 dB is as shown in Figures (8) (9), (10), (11) and (12)

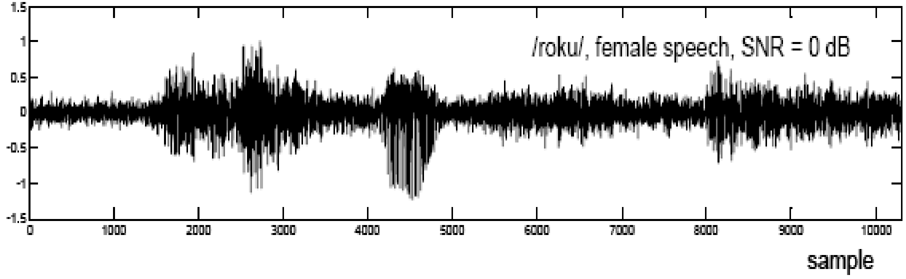

Figure 8. Speech data captured

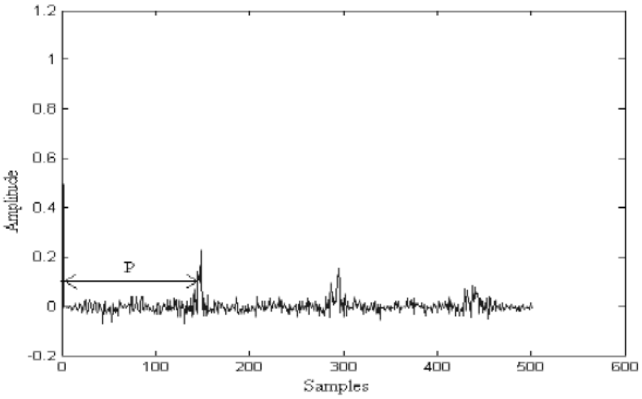

Figure 9. Residual Signal (Applying AMDF discussed)

‘P’ refers to the pitch period. The Speech data was captured and then noise was added. It was then clipped after every 30ms with 50 percent overlap between adjacent frames and then the amplitude was normalized and AMDF was applied.

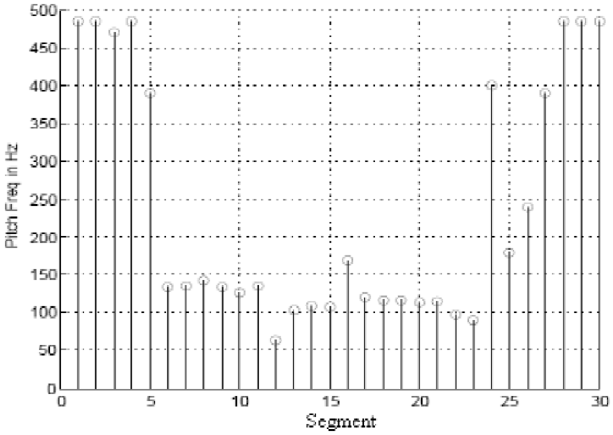

Figure 10. Pitch for a segment

It was observed that it provides a pitch for a segment Also it is assumed that the pitch is constant over a short segment of speech signal.

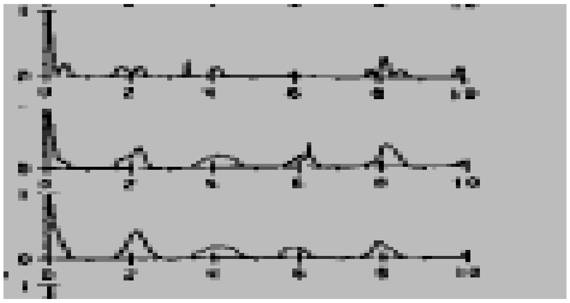

Figure 11. Algorithm #1 result for three sections of speech frames

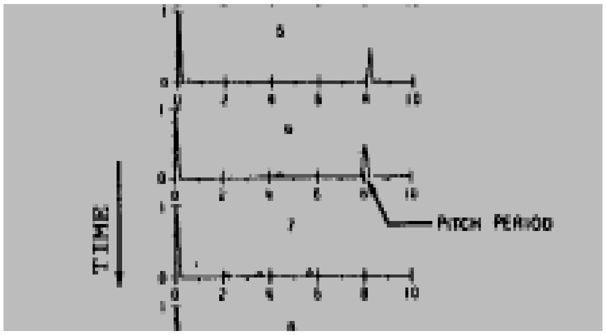

Figure 12. Algorithm #2 result for three sections of speech frames.

From these results, we conclude that the proposed pitch detection algorithms performs well even in very noisy condition and at different environmental noises. It may be noted that some of the data consists of babble noise, which is a major source of pitch error.

8. Conclusions

A Robust Pitch Extraction Using Gammatone Filter Banks for sub band Speech analysis and optimal pitch extraction for each sub band using weighted Autocorrelation is proposed. The proposed method showed better performance compared to conventional method using autocorrelation in both male and female speech corrupted with different colored noise.

9. References

[1] J.D.Gordy and R.A.Goubran, A combined LPC-based speech coder and filtered-X LMS algorithm for acoustic echo cancellation," *in Proc. IEEE ICASSP*, vol.4,pp.125-128, May, 2004.

[2] Tetsuya Shimamura and Hajirne Kobayashi “Weighted Autocorrelation for Pitch Extraction of Noisy Speech “, IEEE Transactions on Speech and Audio Processing, Vol. 9,No.7, October 2001.

[3] G. Muhammad, Noise robust pitch detection based on extended AMDF, Proc. 8th IEEE Int. Symp. on Signal Proessing and Information Technology,(Sarajevo,Bosnia & Herzegovnia,2008) pp. 133-138.

[4] R. G. Amado and J. V. Filho, Pitch detection algorithms based on zero-cross rate and autocorrelation function for musical notes,Proc. Int. Conf. on Audio, Language and Image Processing, (Shanghai, China, 2008) pp. 449-454.

[5] X-D. Mei, J. Pan and S-H. Sun, Efficient algorithms for speech pitch estimation, Proc.Int. Symp. on Intelligent Multimedia, Video and Speech Processing, (Hong Kong, 2001) pp. 421- 424.

[6] M. S. Rahman, H. Tanaka and T. Shimamura, Pitch determination using aligned AMDF, Proc. INTERSPEECH 2006 (Pittsburgh, USA, 2006) pp. 1714-1717

[7] W. Zhang, G. Xu and Y. Wang, Pitch estimation based on circular AMDF, Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing (Florida, USA, 2002) pp. 341- 344.

[8] H. Boril and P. Pollak, Direct Time Domain Fundamental Frequency Estimation of Speech in Noisy Conditions, Proc. European Signal Processing Conference, vol. 1 (Vienna, Austria, 2004) pp. 1003-1006.