A Method Utilizing Window Function Frequency Characteristics for Noise-Robust Spectral Pitch Estimation

Iman Haji Abolhassani[1], Douglas O'Shaughnessy[1], and Sid-Ahmed Selouani[2]

[1] INRS-Energie-Matériaux-Télécommunications, Université du Québec, Montréal QC H5A 1K6, Canada
[2] Université de Moncton, Campus de Shippagan NB E8S 1P6, Canada
imanhaji@emt.inrs.ca, dougo@emt.inrs.ca, sid-ahmed.selouani@umcs.ca

ABSTRACT

A novel method for spectral-domain fundamental frequency (F0) estimation is proposed. The basis of this method is estimating F0 using the power spectrum of a windowed speech segment. For this purpose, a new transform is introduced. The prominent feature of this transform is that it estimates F0 from the speech segment power spectrum by exploiting the window function power spectrum. As a result, this transform is named the Window-Based transform. By comparison between the proposed method and the autocorrelation and the cepstral pitch estimation methods, the superiority of the proposed method under noisy environments is demonstrated.

1. INTRODUCTION

The fundamental frequency *F0* is a primary acoustic cue to intonation and stress in speech, and is crucial to phoneme identification in tone languages. Most low-rate voice coders require accurate *F0* estimation for good reconstructed speech, and some medium-rate coders use *F0* to reduce transmission rate while preserving high-quality speech [1]. As a result, many fundamental frequency, or pitch, estimation methods have been proposed each having their own advantages and drawbacks [2]. For instance, the cepstral method for pitch estimation is known for its acceptable performance in high signal-to-noise (SNR) ratios [3], whereas the time-domain autocorrelation method is considered to be one of the most efficient methods in noisy environments [3,4]. In this paper, a novel frequency-domain pitch estimator has been introduced, and it is experimentally demonstrated that it is very efficient both in high and low SNR ratios.

The Fourier transform of a voiced speech signal is the Fourier transform of the glottal excitation *X(f)* multiplied by the Fourier transform of the vocal tract filter *V(f)*. By windowing the time-domain speech signal, in the frequency domain, the Fourier transform of the window function *W(f)* gets convolved with the Fourier transform of the speech signal. So the overall Fourier transform of the windowed speech segment *S(f)* can be written as:

$$S(f) = (X(f).V(f)) * W(f). \quad (1)$$

For windows with big enough window lengths (e.g., at least two times the fundamental period of the speech segment), by assuming the glottal excitation to be an ideal impulse train in the time domain we can rewrite Equation 1 for the magnitude and power spectra as:

$$|S(f)| = (|X(f)|.|V(f)|) * |W(f)| \quad (2)$$

$$|S(f)|^2 = (|X(f)|^2.|V(f)|^2) * |W(f)|^2. \quad (3)$$

Since we assumed the glottal excitation to be an ideal impulse train in the time domain, its magnitude spectrum |*X(f)*| also becomes an impulse train in the frequency domain. The multiplication of |*X(f)*| by the vocal tract filter |*V(f)*| scales the impulses in |*X(f)*| according to |*V(f)*|. So basically, (|*X(f)*|.|*V(f)*|) is still an impulse train, but with scaled impulses. Thus, the convolution of (|*X(f)*|.|*V(f)*|) and |*W(f)*| results in shifted scaled instances of |*W(f)*| at the locations of these impulses. These shifted scaled instances of |*W(f)*| are called "harmonics". The same idea is valid for the power spectrum of the speech segment $|S(f)|^2$ where the shifted scaled instances of $|W(f)|^2$ represent the "power-spectrum harmonics".

Estimation of *F0* in the frequency domain is the estimation of the periodicity of the harmonics. Since we decide which window function to use for windowing the time-domain speech signal, we have a good knowledge of the shape of the harmonics in the power spectrum of the windowed speech segment. This is because, as mentioned earlier, the power spectrum harmonics are shifted scaled versions of the power spectrum of the window function $|W(f)|^2$. The proposed method exploits this knowledge to come up with a more accurate *F0* estimate. Some attempts have been made earlier to estimate the shape of the harmonics in the frequency domain and then use that harmonic shape estimate to calculate the fundamental frequency in the frequency domain (e.g., the harmonics sieve method [5]). However, using the frequency characteristics of the window function for this purpose, which is suggested in this paper, is novel. In order to estimate *F0* from the power spectrum of the windowed segment, a new transform is introduced. The transform is named the Window-Based Transform (T_{WB}).

In the next section, the proposed transform will be defined. Section 3 focuses on the implementation of the method. In section 4 experiments are carried out to confirm the efficiency of the proposed method in noisy environments and section 5 is the conclusion of this paper.

2. PROPOSED METHOD

The essence of the proposed pitch estimation method is the Window-Based Transform (T_{WB}). The component functions of this transform are named the Window-Based Functions (F_{WB}) and are defined as:

$$F_{WB}(f,F,|W(f)|^2) = \left((-1)^{\left\lfloor \frac{2f}{F} \right\rfloor} \times \Delta_{\frac{F}{2}}(f)\right) * \frac{|W(f)|^2}{Max(|W(f)|^2)}. \quad (4)$$

f is the frequency (Hz) and *F* is the period of F_{WB}. *F* is measured in Hz since F_{WB} is going to be used in the frequency domain. $\Delta_F(f)$ is the impulse train function, which is defined by:

$$\Delta_F(f) = \sum_{k=-\infty}^{\infty} \delta(f - kF). \quad (5)$$

The $\left((-1)^{\left\lfloor \frac{2f}{F} \right\rfloor} \times \Delta_{\frac{F}{2}}(f)\right)$ part of the F_{WB} represents an impulse train

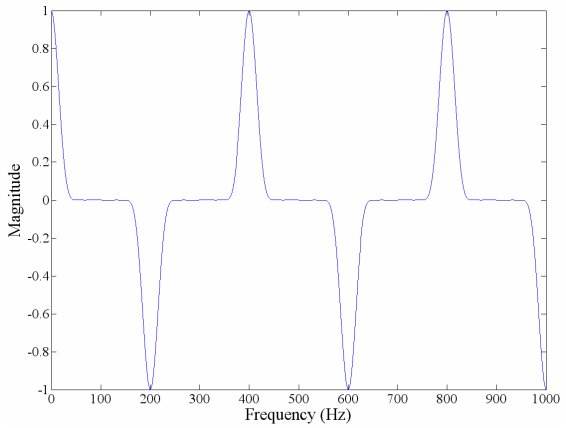

Figure 1 - F_{WB} with *F*=400 Hz and Bartlet-Hann window with the window length of 40 msec plotted for 0 to 1 kHz.

with inverted odd impulses with the overall period of *F*. $|W(f)|^2$ is the power spectrum of the window function and the division by $Max(|W(f)|^2)$ limits the range of F_{WB} to: $-1 \le F_{WB}(f, F, |W(f)|^2) \le +1$. In Equation (4) we can see the dependency of the Window-Based Function on the power spectrum of the window function. An example of F_{WB} is shown in Figure 1.

Based on the F_{WB}'s, we define the T_{WB} as:

$$T_{WB}(F) = \int_{f=f1}^{f2} |S(f)|^2 \cdot F_{WB}(f, F, |W(f)|^2) df \,. \qquad (6)$$

In Equation 6, $|S(f)|^2$ is the power spectrum of the windowed speech segment. The boundaries of the integral depend on the frequency region we choose to analyze, e.g., for the experiments in this paper they were chosen as 0 to 1 kHz.

The value of *F* for which $T_{WB}(F)$ is maximum is the fundamental frequency. The reason is that $T_{WB}(F)$ will have a maximum at *F0* if the positive components (shifted instances of $|W(f)|^2$) of the corresponding F_{WB}, which is $F_{WB}(f, F0, |W(f)|^2)$, occur at the same frequencies as the harmonics of the speech segment power spectrum. This is illustrated in Figure 2. As can be seen from the figure, by exploiting the power spectrum of the window function, the Window-Based Function adapts its components to the harmonics of the speech segment power spectrum. The importance of this adaptation reveals itself in noisy environments. Since we are matching the components of F_{WB} with the harmonics of $|S(f)|^2$, we are decreasing the effect of noise in tracking the harmonic peaks. The idea is similar to the concept of matched filtering in digital communication systems [6].

The role of the positive components in the Window-Based Function has been explained. Negative components have also been added to the function because of the following reasons:

- They make the mean of the F_{WB} equal to zero. If the negative components were not added, the F_{WB}'s with lower periods would have higher means and thus would tend to return higher values for the corresponding T_{WB}. This would make the comparison for choosing the maximum for T_{WB} unfair.
- The negative components have been positioned between the positive ones. This characteristic minimizes pitch doubling errors. The reason is that for the F_{WB} with the period equal to the double of the fundamental frequency, the negative components of the F_{WB} will be placed on harmonics of the power spectrum and this will lead to a major decrease in the value of T_{WB} for the corresponding period *F*=2*F0*.

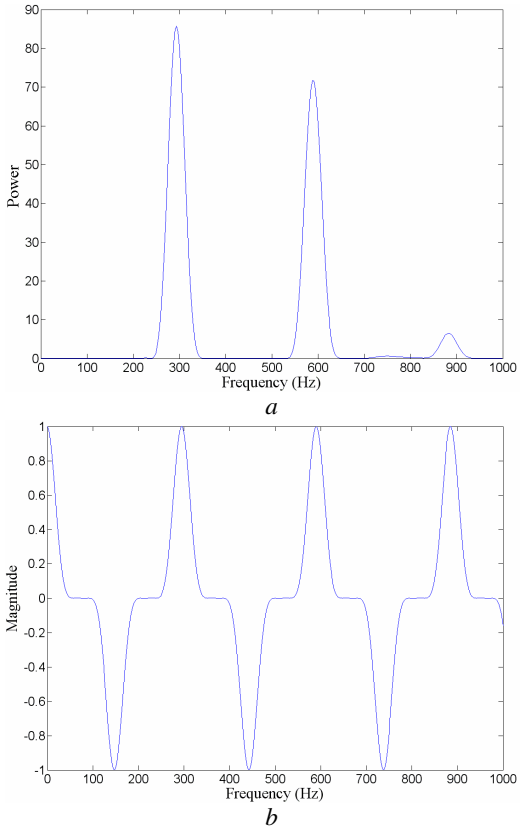

Figure 2 - a) The power spectrum of a real speech segment for 0 to 1 kHz with *F0*=295 Hz, windowed by a 35 msec Hanning window; b) The F_{WB} using the same window and *F*=295 Hz.

Another issue that should be mentioned is that, although the glottal excitation is not an ideal impulse train, the experimental results in the following sections will confirm that considering it as an ideal impulse train is not a bad approximation for this method.

Comparison between T_{WB} and the Fourier transform

Since the power spectrum of a sample is the frequency representation of its time-domain autocorrelation, applying the Fourier transform on the power spectrum of a sample is identical to its autocorrelation in the time domain. Thus, a comparison between T_{WB} and the Fourier transform can be regarded as a comparison between our pitch-estimation method and the autocorrelation method. This theoretical comparison is done in this section. In the Experiments section, our method and the time-domain autocorrelation method are experimentally compared as well.

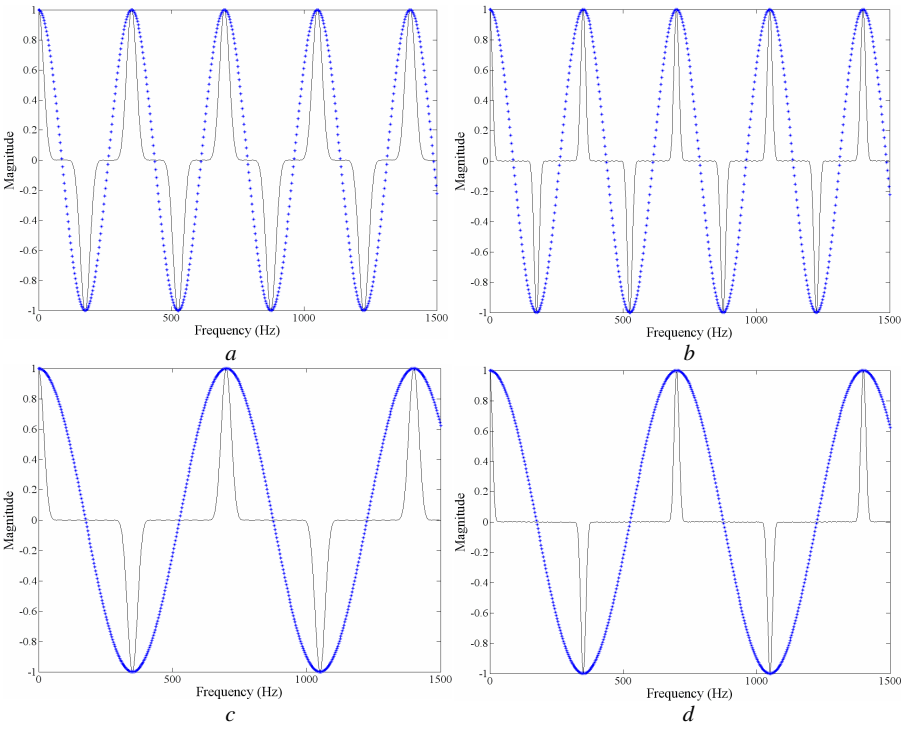

Figure 3 - In each of the figures a F_{WB} and a sinusoid having the same period (*F*) are plotted. The period is denoted by *F* and is measured in Hz since the plotted functions are used in the frequency domain. In a and b *F*=350 Hz and in c and d *F*=700 Hz. The window functions are all Hamming with the length of 30 msec in a and c, and 60 msec in b and d. The figures are plotted for the range of 0 to 1500 Hz.

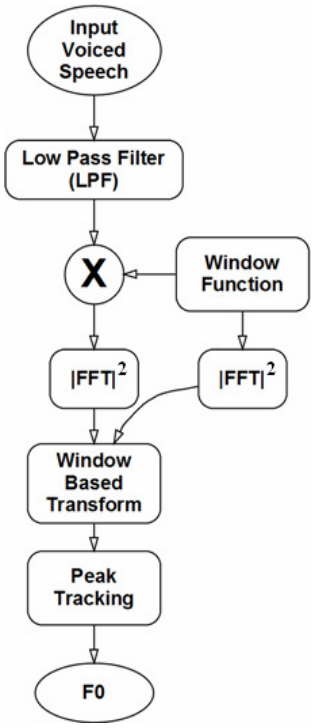

Figure 4 - Block diagram of the Implementation of the Window-Based Spectral-domain Pitch Estimation Method.

To compare the Fourier transform and the T_{WB} we need to compare their component functions, namely the sinusoids and the F_{WB}'s. In each of the Figures 3-a, 3-b, 3-c, and 3-d a sinusoid and a F_{WB} with the same period have been plotted together. These figures can be used to better understand the differences between sinusoids and F_{WB}'s. It should be noted that since we apply these functions on the speech segment power spectrum in the frequency domain, their periods are denoted by *F* and are measured in Hz:

- For different periods (*F*) and the same window functions, the bandwidth of the components of the F_{WB} stays unchanged while the bandwidth of the components of the sinusoid (the positive and negative half cycles) changes.
- For different windows (different in window type or window length) and the same period (*F*), the bandwidth of the components of the F_{WB} changes according the power spectrum of the window function to best match the harmonics of the speech segment power spectrum. However, the bandwidth of the components of the sinusoid (the half cycles) stays the same, since a sinusoid is uniquely defined by its period (*F*).

Thus, compared to the F_{WB}, the *FT* is more likely to make errors in estimating *F0* of the speech segment power spectrum because of two reasons:

- Since the components of the sinusoid (the half cycles) are not matched with the harmonics, in low SNR values they are more vulnerable to noise.
- Since the bandwidth of the components of a sinusoid (the half cycles) is not necessarily equal to the bandwidth of $|S(f)|^2$ harmonics, the *FT* might pick formants instead of harmonics because of this ambiguity in the bandwidth of the harmonics.

As a result, a better performance is expected from our method compared to the autocorrelation method.

3. IMPLEMENTATION

Figure 4 shows the block diagram of the implementation of the proposed method. The time-domain speech signal is first low-pass filtered to 1000 Hz. Then it is multiplied by the desired window. The frequency representation of the resulting frame is calculated by applying the Fast Fourier Transform (*FFT*). By multiplying this frequency representation by its complex conjugate, we obtain the power spectrum. The power spectrum of the window function is calculated in the same fashion. Using the power spectrum of the window function, the Window-Based Transform block synthesizes its F_{WB} s and, using them, calculates T_{WB}. The optimal implementation of the Window-Based Transform block is discussed in the coming subsection. After calculating T_{WB}, the period *F* that corresponds to the maximum of T_{WB} is picked as the fundamental frequency *F0*. In Figure 5, the power spectrum of an exemplary speech segment and T_{WB} result are shown.

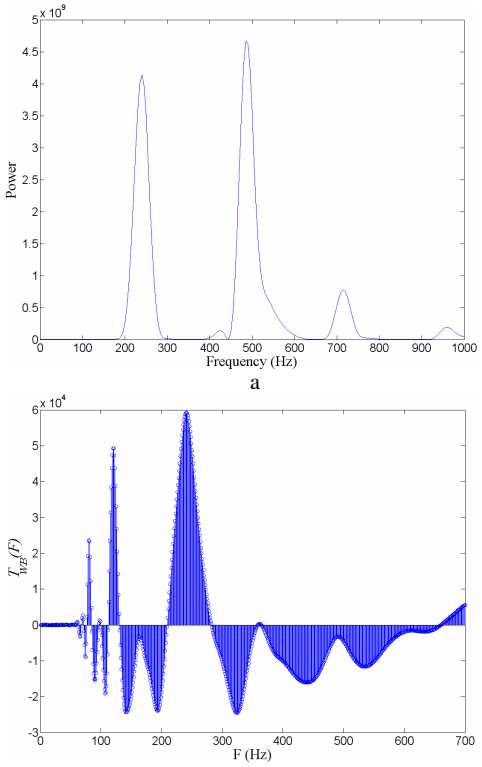

Figure 5 - a) Power spectrum of a speech segment windowed by a Hanning window with the length of 35 msec, b) T_{WB} of the same power spectrum has its maximum at *F*=241 Hz, which is the correct value of the pitch.

As mentioned earlier, the period value *F* for which $T_{WB}(F)$ is maximum is the fundamental frequency (*F=F0*). Sometimes, however, the maximum of $T_{WB}(F)$ occurs in period values equal to a division of the fundamental frequency (i.e., *F=F0/n, n*=2,3,...). This is because for these values of *F* the corresponding F_{WB} function, in addition to including all of the components of the F_{WB} at *F=F0*, includes additional positive and negative components. These additional components, although not occurring at the same frequency locations as the harmonics of $|S(f)|^2$, might make a small increase in the value of $T_{WB}(F)$ at *F=F0/n* compared to $T_{WB}(F0)$, especially at low SNR values. In this way, the maximum of $T_{WB}(F)$, instead of occurring at *F=F0* occurs at a division of *F0* (e.g., *F0*/2). However, since this increase is relatively small, we can solve this issue by using a threshold. We first find the maximum value of T_{WB}, then define a threshold based on that (e.g., 70% of that value) and finally among the peaks which exceed this threshold, we choose the one which points to the highest fundamental frequency. This is done in the peak-tracking module.

The last point to discuss in this section is that in order to enable our pitch estimation method to also accept unvoiced segments as input, we only need to add a Voiced/Unvoiced Detector (VUD) block in the beginning of the block diagram to bypass the unvoiced segments.

Implementation of the Window-Based Transform Block

In the Window-Based Transform block in Figure 4, the F_{WB}'s are synthesized using the power spectrum of the window function. Our base equation for extracting the F_{WB}'s is Equation 4 which is written in the frequency domain. However, in order to decrease the computational cost, practically we extract the F_{WB}'s in a different way:

As the first step, we rewrite Equation 4, but this time we separate the positive and negative components:

$$F_{WB}(f,F,|W(f)|^2)=(\Delta_F(f))*\frac{|W(f)|^2}{Max(|W(f)|^2)} -\left(\Delta_F(f-\frac{F}{2})\right)*\frac{|W(f)|^2}{Max(|W(f)|^2)} \quad (7)$$

By using Equation 5, we rewrite this equation as:

$$F_{WB}(f,F,|W(f)|^2)=\sum_{k=k1}^{k2}\frac{|W(f-kF)|^2}{Max(|W(f)|^2)} -\sum_{m=m1}^{m2}\frac{|W(f-mF-\frac{F}{2})|^2}{Max(|W(f)|^2)}. \quad (8)$$

The boundaries of the sums (*k1*, *k2*, *m1*, and *m2*) are matched with the boundaries of the $T_{WB}(F)$ integral in Equation 6, which define our analysis frequency region.

If the length of the window function is big enough (at least two times the pitch period), the bandwidth of the power spectrum harmonics, which are the shifted instances of the window function power spectrum $|W(f)|^2$, become small and as a result, we can assume the harmonics to be separate. Using this assumption, we can write Equations 9 and 10 as:

$$\sum_k\frac{|W(f-kF)|^2}{Max(|W(f)|^2)}\approx\left(\sum_k\frac{|W(f-kF)|}{Max(|W(f)|)}\right)^2 \quad (9)$$

$$\sum_m\frac{|W(f-mF-\frac{F}{2})|^2}{Max(|W(f)|^2)}\approx\left(\sum_m\frac{|W(f-mF-\frac{F}{2})|}{Max(|W(f)|)}\right)^2. \quad (10)$$

Using Equations 9 and 10, the base equation can be written as:

$$F_{WB}(f,F,|W(f)|^2)=\left(\sum_k\frac{|W(f-kF)|}{Max(|W(f)|)}\right)^2 -\left(\sum_m\frac{|W(f-mF-\frac{F}{2})|}{Max(|W(f)|)}\right)^2. \quad (11)$$

If we denote the time-domain window function as *w[n]* ($w[n]\xleftrightarrow{FFT}W(f)$), using the frequency shifting characteristic of the Fourier transform we can write:

$$w[n].e^{2\pi jf_0n}\xleftrightarrow{FFT}W(f-f_0). \quad (12)$$

By using the frequency shifting and linearity characteristics of the Fourier transform, we can finally write our base equation as:

$$F_{WB}(f,F,|W(f)|^2)=\left(Max(|W(f)|)^{-1}.\left|FFT\left(\sum_k w[n]e^{2\pi jkFn}\right)\right|\right)^2 -\left(Max(|W(f)|)^{-1}.\left|FFT\left(\sum_m w[n]e^{2\pi j(mF-\frac{F}{2})n}\right)\right|\right)^2. \quad (13)$$

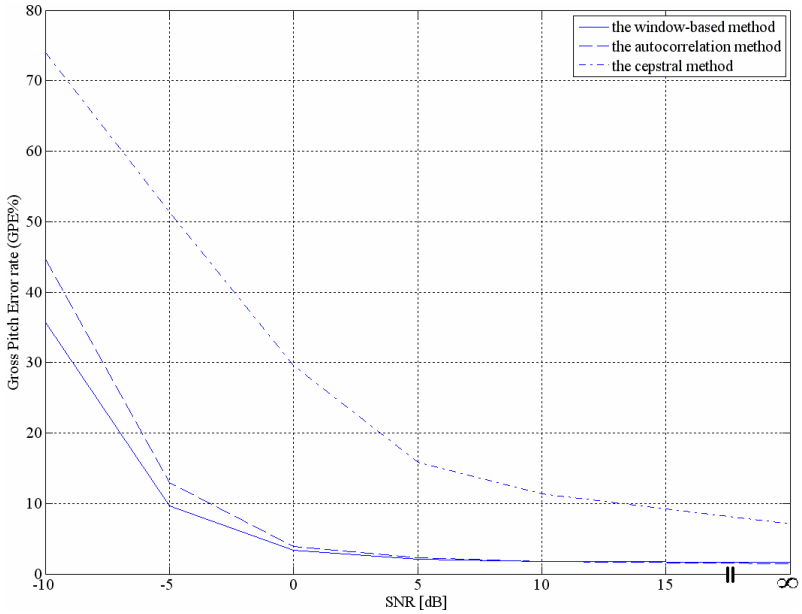

Figure 6 - Gross Pitch Error rate (GPE%) comparison between the window-based, autocorrelation, and cepstral pitch estimators.

Implementing the synthesis of the F_{WB}'s using Equation 13 is much simpler and faster than doing the same task using Equation 4 directly. It should be noted that in this way, for synthesizing each F_{WB} we only need to perform two *FFT* operations (one for the positive and one for the negative components) and some simple multiplications which, altogether, make this method quite fast and practical.

4. EXPERIMENTS

4.1 Experimental Details

For evaluating the efficiency of the proposed method under noisy environments, we chose to compare it with the autocorrelation and the cepstral pitch estimation methods. The autocorrelation method was chosen since it is one of the best pitch estimators under noisy environments [3,4]. The cepstral method was also chosen because, like our method, it estimates *F0* using the speech segment frequency representation. The samples were taken out of the DARPA TIMIT acoustic-phonetic continuous speech corpus database. The sampling frequency was 16000 Hz. Speech samples were windowed using a 40 msec Hanning window and the time between frames was 10 ms. The samples were analyzed at the SNR values of infinity, 10, 5, 0, -5, and -10 dB by the window-based, autocorrelation, and cepstral pitch estimation methods, using additive white Gaussian noise. A total number of 3714 pitch estimation iterations were performed for each method. No post processing or smoothing was done on the data. Finally the number of the gross pitch estimation errors was counted. If P_e is the estimated pitch and P_r is the reference correct pitch, if $\left|\frac{P_e - P_r}{P_r}\right| > 0.1$, we regard this as a gross pitch estimation error.

4.2 Experimental Results

Figure 6 compares the performance of the three methods. It can be seen that, as expected, the proposed method shows a superior performance in low SNR values compared to both the autocorrelation and the cepstral pitch estimation methods. At the SNR value of 10 dB, the window-based and the autocorrelation pitch estimation methods show about the same performance and for higher SNR values, the autocorrelation method shows a slightly better performance. The performance of this method is also impressive compared to the other methods which are tested in [3,4].

5. CONCLUSION

A new method for noise-robust spectral pitch estimation is proposed that utilizes a new transform. Since the transform incorporates the frequency characteristics of the window function, it best matches its component functions to the harmonics of the speech segment power spectrum. In this way, the proposed method is efficient in minimizing the gross pitch estimation errors in noisy environments. The gross errors of this method, the autocorrelation method, and the cepstral method were compared and it was confirmed that the proposed method shows superior robustness to noise. We are continuing our efforts by assessing the performance of this scheme in speech recognition configuration in adverse conditions.

REFERENCES

[1] Douglas O'Shaughnessy, *Speech communications: Human and machine*, Second Edition, IEEE, 2000.

[2] L.R. Rabiner, M. Chen, "*A comparative performance study of several pitch detection algorithm*," *IEEE Trans*. Acoust. Speech, Signal Process. ASSP-24 (5) (October 1976) 56–60.

[3] Shimamura, T., and Kobayashi, H.: '*Weighted autocorrelation for pitch extraction of noisy speech*', *IEEE Trans*. Speech Audio Process., 2001, SAP-9, (7), pp. 727–730.

[4] K. A. Oh and C. K. Un, "*A performance comparison of pitch extraction algorithms for noisy speech*," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, *1984*, pp. 18B4.1–18B4.4.

[5] Duifhuis et al., 1982 H. Duifhuis, L.F. Willems and R.J. Sluyter, "*Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception,*" *J. Acoust. Soc. Amer.* 71 (1982) (6), pp. 1568–1580.

[6] B. Sklar, Digital Communications- *Fundamentals and Applications*, Second Edition, Prentice Hall, 2001, pp. 122-124.