wavelets offer a promising and flexible tool, as early attempts indicate [2], [25], [22].

ACKNOWLEDGMENT

Useful discussions with N. Gache, E. Payot, O. Rioul, G. Ruckebush, and within the Rennes Working Group "Analyse multirésolution de signaux aléatoires" (M. Basseville and A. Benveniste) are gratefully acknowledged.

REFERENCES

[1] D. W. Allan, "Statistics of atomic frequency clocks," *Proc. IEEE*, vol. 54, no. 2, pp. 221-230, 1966.
[2] E. Bacry, A. Arnéodo, U. Frisch, Y. Gagne, and E. Hopfinger, "Wavelet analysis of fully developed turbulence data and measurement of scaling exponents," in *Turbulence and Coherent Structures*, O. Métais and M. Lesieur, Eds. New York: Kluwer, 1991, pp. 203-215.
[3] L. F. Burlaga and L. W. Klein, "Fractal structure of the interplanetary magnetic field," *J. Geophys. Res.*, vol. 91, no. A1, pp. 347-350, 1986.
[4] A. Cohen, "Ondelettes, Analyses Multirésolutions et Traitement Numérique du Signal," thése de Doctorat, Univ. Paris IX, Dauphine, 1990.
[5] J. M. Combes, A. Grossmann and Ph. Tchamitchian, Eds., Wavelets. New York: Springer-Verlag, 1989.
[6] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Comm. Pure Appl. Math.*, vol. XLI, no. 7, pp. 909-996, 1988.
[7] K. Falconer, *Fractal Geometry*, Chichester: J. Wiley and Sons, 1990.
[8] P. Flandrin, "On the spectrum of fractional Brownian motions," *IEEE Trans. Inform. Theory*, vol. 35, pp. 197-199, Jan. 1989.
[9] P. Flandrin, "Some aspects of nonstationary signal processing with emphasis on time-frequency and time-scale methods," in [5], pp. 68-98, 1989.
[10] P. Flandrin, "Fractional Brownian motion and wavelets," to appear in *Wavelets, Fractals and Fourier Transforms - New Developments and New Applications*, M. Farge, J. C. R. Hunt and J. C. Vassilicos, Eds. Oxford: Oxford Univ. Press.
[11] N. Gache, P. Flandrin, and D. Garreau, "Fractal dimension estimators for fractional Brownian motions," in *IEEE Int. Conf. Acoust., Speech and Signal Processing, ICASSP-91*, Toronto, pp. 3557-3560, 1991.
[12] A. Grossmann and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape," *SIAM J. Math. Anal.*, vol. 15, no. 4, pp. 723-736, 1984.
[13] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D*, vol. 31, pp. 277-283, 1988.
[14] M. S. Keshner, "$1/f$ noise," *Proc. IEEE*, vol. 70, pp. 212-218, 1982.
[15] M. Kim and A. H. Tewfik, "Multiscale signal detection in fractional Brownian motion," in *Advanced Signal Processing Algorithms, Architectures and Implementations*, F. T. Luk, Ed., also in *SPIE*, vol. 1348, pp. 462-470, 1990.
[16] P. Lévy, "Le Mouvement Brownien," *Mém. Sc. Math.*, fasc. 126, pp. 1-81, 1954.
[17] T. Lundahl, W. J. Ohley, S. M. Kay, and R. Siffert, "Fractional Brownian motion: A maximum likelihood estimator and its application to image texture," *IEEE Trans. Med. Imaging*, vol. MI-5, no. 3, pp. 152-161, 1986.
[18] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-11, no. 7, pp. 674-693, 1989.
[19] B. Mandelbrot, *The Fractal Geometry of Nature*. Freeman: San Francisco, 1982.
[20] B. B. Mandelbrot and J. W. van Ness, "Fractional Brownian motions, fractional noises and applications," *SIAM Rev.*, vol. 10, no. 4, pp. 422-437, 1968.
[21] Y. Meyer, "Orthonormal wavelets," in *Wavelets*, J. M. Combes, A. Grossmann, and Ph. Tchamitchian, Eds. New York: Springer-Verlag, 1989, pp. 21-37.
[22] J. F. Muzy, E. Bacry, and A. Arnéodo, "Wavelets and multifractal formalism for singular signals: Application to turbulence data," preprint, 1991.
[23] H. O. Peitgen and D. Saupe, Eds., *The Science of Fractal Images*. New York: Springer-Verlag, 1988.
[24] A. H. Tewfik and M. Kim, "Correlation structure of the discrete wavelet coefficients of fractional Brownian motions," to appear in *IEEE Trans. on Inform. Theory*.
[25] M. Vergassola and U. Frisch, "Wavelet transforms of self-similar processes," to appear in *Physica D*.
[26] N. Wiener, *Nonlinear Problems in Random Theory*. Cambridge, MA: MIT Press, 1958.
[27] G. W. Wornell, "A Karhunen-Loève-like expansion for $1/f$ processes via wavelets," *IEEE Trans. Inform. Theory*, vol. 36, pp. 859-861, July 1990.
[28] G. W. Wornell and A. V. Oppenheim, "Estimation of fractal systems from noisy measurements using wavelets," to appear in *IEEE Trans. Signal Proc.*
[29] A. M. Yaglom, *Correlation Theory of Stationary and Related Random Functions*. New York: Springer-Verlag, 1986.

Application of the Wavelet Transform for Pitch Detection of Speech Signals

Shubha Kadambe and G. Faye Boudreaux-Bartels

Abstract—**An event detection pitch detector based on the dyadic wavelet transform is described. The proposed pitch detector is suitable for both low-pitched and high-pitched speakers and is robust to noise. Examples are provided that demonstrate the superior performance of this event based pitch detector in comparison with classical pitch detectors that use the autocorrelation and the cepstrum methods to estimate the pitch period.**

Index Terms—**Glottal closure, event, dyadic wavelet transform, pitch detection, local maxima.**

I. INTRODUCTION

The pitch period is an important parameter in the analysis and synthesis of speech signals. Pitch period information is used in various applications such as 1) speaker identification and verification, 2) pitch synchronous speech analysis and synthesis, 3) linguistic and phonetic knowledge acquisition and 4) voice disease diagnostics. The task of estimating the pitch period is very difficult since a) the human vocal tract is very flexible and its characteristics vary from person to person, b) the pitch period can vary from 1.25 ms to 40 ms, c) the pitch period of the same speaker can vary depending upon the emotional state of the speaker and d) the pitch period can be influenced by the way the word is pronounced (accent). Therefore, no one algorithm that has been developed so far performs perfectly for 1) different speakers (male, female, children and people with different native languages), 2) different applications and 3) different environmental conditions [1].

The pitch detectors that have been developed so far, can be broadly classified into either event detection pitch detectors or

Manuscript received February 15, 1991; revised September 1, 1991. This work was supported in part by ONR Grant #N00014-89-J-1812. This work was presented at the International Conference on Spoken Language Processing, Kobe, Japan, November 18, 1990, and at the Twenty-Fourth Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, November 5-7, 1990.

S. Kadambe is with the Applied Science and Engineering Laboratories, A. I. duPont Institute, Wilmington, DE 19899.

G. F. Boudreaux-Bartels is with the Department of Electrical Engineering, University of Rhode Island, Kingston, RI 02881.

IEEE Log Number 9104511.

nonevent detection pitch detectors. For a detailed review, refer to [1].

Event detection pitch detectors estimate the pitch period by locating the instant at which the glottis closes (called an event) and then measuring the time interval between two such events. Only a few event based pitch detectors [2]–[5] have been developed recently. In [2], the instant at which the glottis closes is determined by locating the instant at which the determinant of the autocovariance matrix of a given signal is maximum. The advantage of this autocovariance method is that it can estimate the pitch period very accurately in the case of certain vowels that are produced by vigorous vocal cord vibrations with sharp glottal closure. However, the disadvantages of this autocovariance based method are the following: a) it is unsuitable for all vowels and for nonstationary pitch periods and b) it is computationally complex. The pitch detection techniques in [4] and [5] use the occurrence of discontinuities in the derivatives of glottal airflow to detect the Glottal closure instant (GCI). These two methods can detect precisely the instant at which the vocal tract is excited. However, these two epoch extraction methods are applicable for only "clean" data and certain vowels. Finally, the maximum likelihood epoch determination technique is applied in [3] to detect the GCI. This method gives accurate estimation of pitch period in the case of synthetic signals (all vowels), noisy signals (up to 0 dB signal-to-noise ratio) and phase distorted signals. However, this method is not suitable for high pitched speakers since the data length available for the linear predictor could be very small. This method is also computationally intensive.

Classical pitch detectors estimate the pitch period by a direct approach and hence, they are referred to here as nonevent based pitch detectors. Some of the nonevent based pitch detectors developed so far [6]–[10] estimate the average pitch period over a segment of a speech signal that they obtain by using a window whose length is fixed. For each segment the average pitch is estimated using one of the following methods: 1) compute the autocorrelation of the infinitely and centrally clipped signal [6], 2) compute the cepstrum of a given segment of a signal [7], 3) evaluate the autocorrelation of an inverse filtered signal [8], 4) compute the average magnitude difference function of a given signal [9] or 5) calculate the autocorrelation of a given signal and look for the value of the pitch period which maximizes the sum of the autocorrelation functions [10]. These nonevent based pitch detectors are computationally simple; however, they assume that the pitch period is *stationary* within each segment and each segment contains at least two full pitch periods. Hence, the disadvantages of these pitch detectors are that: they are a) insensitive to nonstationary variations in the pitch period over the segment length and b) unsuitable for both low pitched and high pitched speakers.

In this correspondence, we describe an event detection pitch detector which is robust to noise and is suitable for a wide range of pitch periods and for different speakers. We apply a time-scale representation known as the dyadic wavelet transform (D_yWT) to the task of locating glottal closure. This correspondence is organized as follows: in Section II, we review the D_yWT and show how it is suitable for the detection of the instant at which the glottis closes. Section III is devoted to the description of the event detection pitch detector based on the D_yWT. In Sections IV and V, we provide examples and discuss the applicability of this method to 1) a wide range of pitch periods, 2) nonstationary pitch periods, and 3) noisy signals. We also compare the performance of the D_yWT pitch detector with the classical pitch detectors. In Section VI, we conclude with a summary of the merits and the demerits of the D_yWT pitch detector.

II. Dyadic Wavelet Transform

The D_yWT of a signal $x(t)$, [11]

$$D_yWT_x(b,2^j) = \frac{1}{2^j}\int_{-\infty}^{\infty} x(t)\, g^{\star}\left(\frac{t-b}{2^j}\right) dt = x(t) \otimes g_{2^j}^{\star}(t), \quad (1)$$

computes the wavelet transform using a scale parameter, $a = 2^j$ that is discretized along the dyadic sequence. Here, $g^{\star}(t)$ is the complex conjugate of a wavelet function $g(t)$ that satisfies the conditions mentioned in [11], $g_{2^j}(t) = \frac{1}{2^j} g\left(\frac{t}{2^j}\right)$, and $\otimes$ represents the convolution operator. From a signal processing point of view, the D_yWT can be considered as the output of a bank of constant-Q, octave band, band pass filters whose impulse responses are $\frac{1}{2^j} g\left(\frac{t}{2^j}\right)$. The bandwidth and the center frequency of each such filters are proportional to $\frac{1}{2^j}$.

The following are some of the interesting properties which make the D_yWT a useful tool for the analysis of speech signals.

- The D_yWT is linear and shift invariant [11], which are useful properties for speech signals since they are often modeled as a linear combination of shifted and damped sinusoids.
- If a signal $x(t)$ or its derivatives have discontinuities, then the modulus of the D_yWT of $x(t)$, $|D_yWT_x(b,2^j)|$, exhibits local maxima around the points of discontinuity [11]. This property will prove useful for D_yWT pitch detector since glottal closure causes sharp changes in the derivative of the air flow in the glottis and transients in the speech signal.

In [11], Mallat has shown that if we choose a wavelet function $g(t)$ that is the first derivative of a smoothing function (a smoothing function is a function whose Fourier transform has energy concentrated in the low-frequency region) $\theta(t)$, then the local maxima of the D_yWT indicate the sharp variations in the signal whereas the local minima indicate the slow variations. Hence, the local maxima of the D_yWT using a wavelet which is the first derivative of a smoothing function should be useful for detecting the abrupt changes or transients in a speech signal caused by the glottal closure. Further, Mallat has demonstrated that really sharp changes in a signal at time $t = t_0$ exhibit local maximum in the D_yWT at $t = t_0$ across several consecutive dyadic scales. He has developed efficient image coding algorithms that utilize the correlation of D_yWT local maxima across two to three successive dyadic scales. We also will check for a correlation of D_yWT local maxima across two successive scales.

III. Event-Based Pitch Detector Using the Dyadic Wavelet Transform

In this section, we describe the algorithm that was used to estimate the pitch period using the D_yWT. We use the D_yWT, since the D_yWT is a very good tool for the analysis of transients as mentioned above. In addition, the analysis of speech signals using the D_yWT is similar to the analysis done by the human ear and therefore, adapted to the auditory perception [12].

- The D_yWT event detection pitch detector algorithm proceeds as follows. The D_yWT of a segment of a speech signal of length L ms is computed at the scale $a = 2^i, i = i_l, i_l + 1, \cdots, i_u$. The starting index i_l and ending index i_u are determined by physical constraints to be explained later. For each scale 2^i, locate the local maxima with respect to b of the

$D_yWT(b, 2^i)$ which exceed a given threshold. In this correspondence, the threshold equals 80% of the global maximum of the D_yWT of the given segment of a speech signal. Compare the locations of the local maxima across consecutive scales as in [13]. If the locations of the thresholded local maxima agree across two scales, we assume that the locations of these maxima correspond to the time of transients caused by glottal closure. We estimate the pitch period by measuring the time interval between two such local maxima.

- Generally, the D_yWT is computed at scales $a = 2^i$ for, theoretically, all i. However, we can limit the number of scale parameters that are needed for the computation of the D_yWT, based on the nature of speech signals which can be broadly classified into either voiced or unvoiced sounds. First, speech generally spans only 10 octaves. Second, the pitch or fundamental frequency of voiced signals is a low frequency (30–500 Hz) phenomena. Third, unvoiced sounds are random in nature and contain high frequency information. Consequently, we have found that computing the D_yWT at three dyadic scales is sufficient for the purposes of estimating the pitch period. The required three dyadic scales are chosen as follows. Given a wavelet with input center frequency f_{c_i} and input bandwidth Δf_i, one can choose the scale parameter 'a' corresponding to the required output center frequency f_{c_o} using the following equation:

$$a = \frac{f_{c_i}}{f_{c_o}}. \tag{2}$$

In this study, we choose the input bandwidth of the wavelet $\Delta f_i = 2 \times f_{c_i}$ and the output bandwidth $\Delta f_o = 2 \times f_{c_o}$. If (f_{c_i}/f_{c_o}) in equation (2) is not equal to some power of two then it is rounded off to nearest power. Using an approach similar to that mentioned in [11], we generate a cubic spline wavelet with input center frequency $f_{c_i} = 8000\ Hz$ and input band width $\Delta f_i = 16\,000\ Hz$. First, we set a lower bound $a = 2^{i_l}$ on the scale parameter by choosing the required output frequency $f_{c_o} = 1000\ Hz$ and bandwidth $\Delta f_{c_o} = 2000\ Hz$. Second, we set an upper bound $a = 2^{i_u}$ on the scale parameter by choosing the required output center frequency $f_{c_o} = 250\ Hz$ and bandwidth $\Delta f_{c_o} = 500\ Hz$ to ensure that the frequencies corresponding to the range of fundamental frequency i.e., 30–500 Hz are never filtered out. We then compute the D_yWT starting at the lowest scale and continue doubling the scale parameter until we reach the highest scale parameter. In this study, the lower and the upper bound on the scale parameter corresponds to $a = 2^3$ and $a = 2^5$, respectively. Hence, we compute the D_yWT at only at three scales; this has the advantage of significantly reducing the computational complexity of the D_yWT.

The flow chart of the D_yWT event based pitch detector algorithm is shown in Fig. 1. The algorithm exploits the fact that the frequency content of unvoiced speech signal is much higher than the range of fundamental frequency; hence, at the higher D_yWT scales unvoiced speech gets filtered out, producing relatively low amplitude D_yWT. Therefore, we can classify the given segment of the speech signal into voiced or unvoiced by comparing the maximum amplitude of the D_yWT with some threshold level T in addition to checking whether the local maxima of the D_yWT correlates across two scales as shown in the flow chart.

IV. Results and Discussion

In this section, we provide synthetic examples illustrating the applicability of the event-based pitch detector described in the previous section for a wide range of pitch periods and for nonstationary pitch periods.

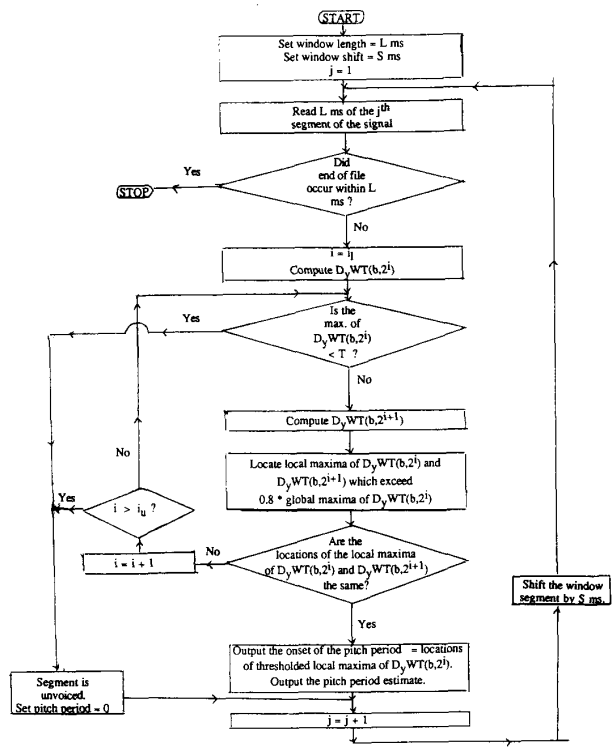

Fig. 1. Flowchart of the D_yWT-based pitch detector.

A. A Synthesized Labiodental Voiced Fricative /v/

First, we consider a synthesized signal /v/ (see Fig. 2(a)) whose pitch period is constant and equal to 10 ms. From Fig. 2(a), it is clear that the onset of the pitch period can *not* be located very easily. In Fig. 2(b)–(f), we plot the D_yWT of the signal /v/ computed at the dilation scales $a = 2^1, 2^2, \cdots, 2^5$, respectively. It can be seen that the D_yWT exhibits local maxima across all these scale parameters at the instant of the onset of each pitch period. However, from Fig. 2(b)–(f), we can also see that the high-frequency information is filtered out as we increase the scale parameter (see Fig. 2(e)–(f)). Hence, in order to estimate the pitch period accurately, we need to choose the D_yWT computed at the scale of 2^4 or 2^5. Our algorithm chose the D_yWT computed at the scale $a = 2^4$, since the locations of the thresholded local maxima of the D_yWT's computed at the scales $a = 2^4$ and $a = 2^5$ matched. For this example, we obtain a 2% relative error of the pitch period estimate using the D_yWT at the scale $a = 2^4$, where the relative error is defined as

$$\text{relative error} = |\text{true pitch period} - \text{estimated pitch period}|\,/\,(\text{true pitch period}).$$

Other synthetic examples demonstrating similar performance are given in [14]. The accuracy of the pitch period estimation depends upon the choice of the wavelet function [15]. From the preliminary studies [14], [15], we have found that the cubic spline wavelet

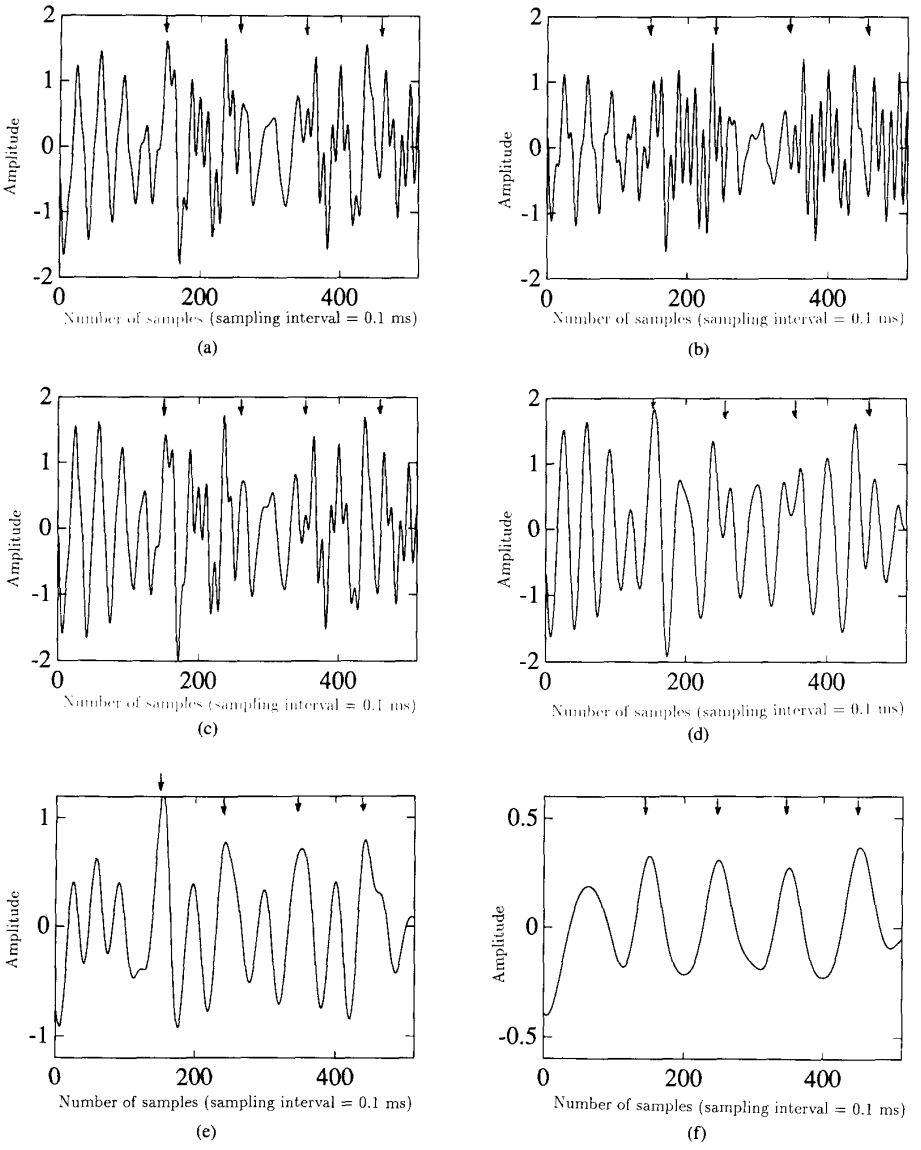

Fig. 2. (a) Synthesized signal /v/ (the arrows indicate the onset of the true pitch period). (b) D_yWT of /v/ computed at the scale $a = 2^1$ (the arrows indicate the onset of the true pitch period). (c) D_yWT of /v/ computed at the scale $a = 2^2$ (the arrows indicate the onset of the true pitch period). (d) D_yWT of /v/ computed at the scale $a = 2^3$ (the arrows indicate the onset of the true pitch period). (e) D_yWT of /v/ computed at the scale $a = 2^4$ (the arrows indicate the onset of the true pitch period). (f) D_yWT of /v/ computed at the scale $a = 2^5$ (the arrows indicate the onset of the true pitch period).

provides the best estimates of the pitch period as compared to the Gaussian, the Haar and the minimum phase wavelets.

B. Comparison of Performance

In this section, we compare the performance of the D_yWT event based pitch detector with standard nonevent based pitch detectors based on the cepstrum and the autocorrelation of a given signal.

Noll [7] has used the cepstrum of a speech signal to estimate the pitch period, since the cepstrum of a periodic signal exhibits the same periodicity as the signal under consideration. The cepstrum of a signal $x(t)$ is defined as

$$C_x(t) = \left[\int_0^\infty \log|X(\omega)|^2 \cos(\omega t)\, d\omega\right]^2, \qquad (3)$$

where $X(\omega)$ is the Fourier transform (FT) of $x(t)$. Dubnowski, Schafer and Rabiner in [16], have used the autocorrelation function of a speech signal to estimate the pitch period. The autocorrelation

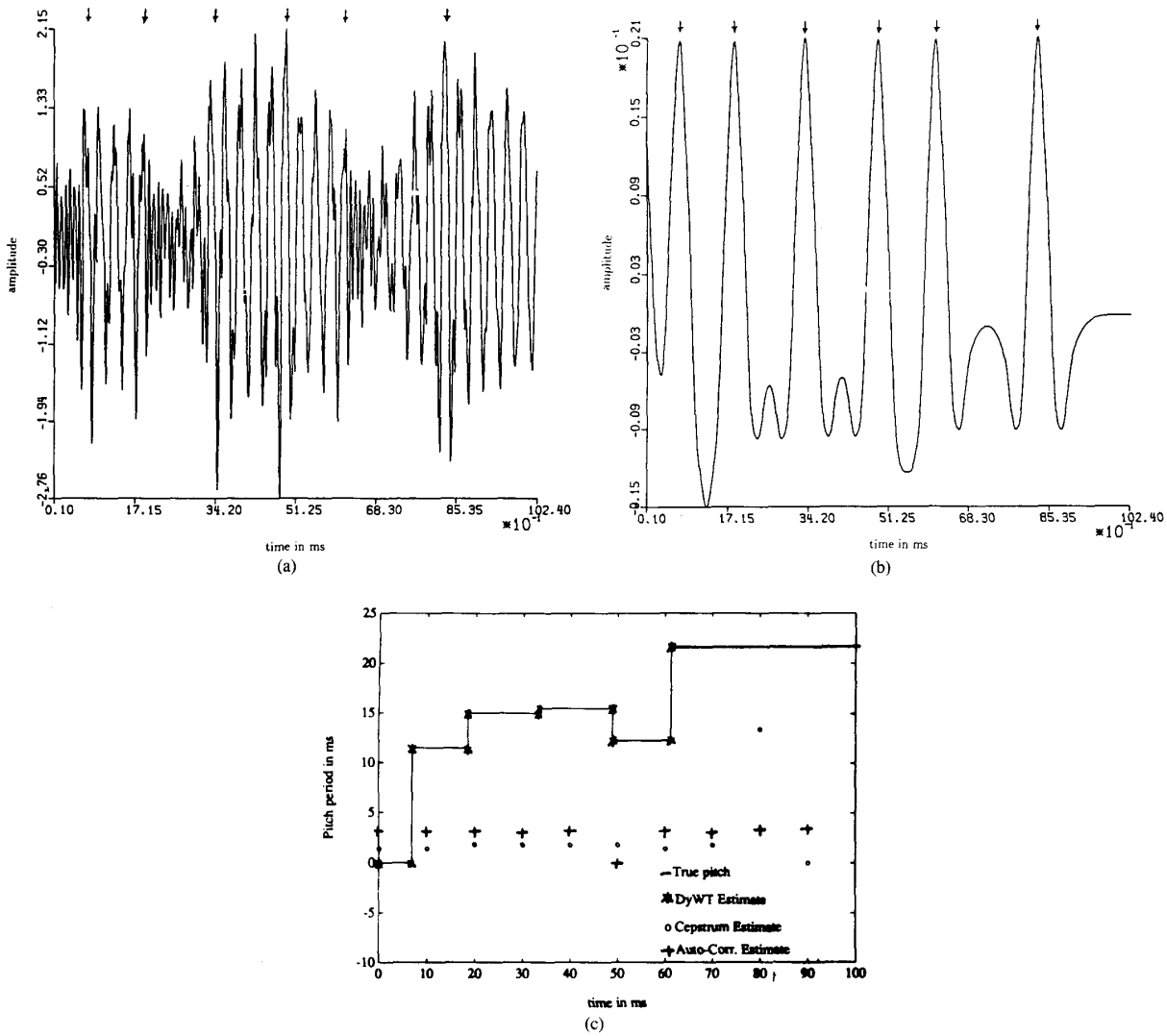

Fig. 3. (a) Synthesized nasal signal /n/ (the arrows indicate the onset of the true pitch period). (b) D_yWT of a synthesized signal /n/, computed at the scale $a = 2^4$ (the arrows indicate the onset of the true pitch period). (c) Comparison of the performance of the D_yWT, the cepstrum, and the autocorrelation based pitch detectors for a synthesized signal /n/.

function $R_x(\tau)$ of a signal $x(t)$ is defined as [17]

$$R_x(\tau) = \int_{-\infty}^{\infty} x^*(t)\,x(t+\tau)\,dt. \tag{4}$$

The autocorrelation function of a periodic signal also exhibits periodicity equal to that of the signal [17]. Dubnowski *et al.* make use of this property of the autocorrelation function to estimate the pitch period of a speech signal in [16]. They compute the autocorrelation of both centrally and infinitely clipped signal.

We have implemented these two classical pitch detectors and compared their performance with the performance of the event based D_yWT pitch detector. The comparison focused upon the following: 1) the accuracy with which the pitch periods can be estimated, 2) the robustness of the algorithms to noise, 3) the computational complexity of the algorithms, and 4) the problems associated with the segmentation of a speech signal.

Accuracy: Here, we use a synthesized nasal sound /n/ to compare the accuracy with which the pitch period can be estimated by the D_yWT, the cepstrum and the autocorrelation based pitch detectors. Other examples are given in [14]. The synthesized signal and its D_yWT are plotted in Fig. 3(a) and (b), respectively. In this example, we deliberately vary the length of the pitch period, from 11.5 ms to 21.7 ms, to make the signal nonstationary. From Fig. 3(b), we can see that the thresholded local maxima of the D_yWT computed at the scale $a = 2^4$ corresponds to the onset of the true pitch period. In Fig. 3(c), we plot the true pitch period versus the pitch period estimated by all three methods. From this figure, it is clear that the D_yWT pitch detector exhibits superior performance and estimates the pitch period with 100% accuracy. This is a worst case scenario for both the cepstrum and the autocorrelation methods since they assume stationarity within the analysis window and

measure only the average pitch period, and hence, are unable to detect the nonstationarities within a signal segment. For stationary, low-pitched signals all three algorithms give similar results [14].

Robustness to Noise: To compare the robustness of the three pitch detectors under noisy situations, we add white Gaussian noise to a synthesized signal /i/. We have chosen this signal since it is a relatively complicated speech signal to analyze. We consider signal-to-noise ratios (SNR) ranging from 0 dB down to −18 dB. In Table I, we tabulate the percentage of relative error obtained at different SNR's by each of the three pitch detection methods. From this table and other examples in [14], we have found that the D_yWT method is generally the most robust to noise as compared to the autocorrelation and the cepstrum based pitch detectors. In general, the D_yWT event based pitch detector gives relative errors $\leq 2\%$ for signals contaminated by noise with SNR's down to −18 dB. However, the performance of the cepstrum method starts to degrade for signals contaminated by noise with SNR less than 0 dB, and the autocorrelation method is very susceptible to noise.

TABLE I
A COMPARISON OF THE ROBUSTNESS OF THE D_yWT, THE CEPSTRUM, AND THE AUTOCORRELATION-BASED PITCH DETECTORS TO NOISE BY ADDING WHITE GAUSSIAN NOISE TO A SYNTHESIZED SIGNAL /i/ WITH SNR'S RANGING FROM 0 dB DOWN TO −18 dB

	Percentage of Relative Error			
SNR	0 dB	−6 dB	−12 dB	−18 dB
D_yWT-based Pitch Detector	.44%	.44%	0.92%	1.1%
Cepstrum-based Pitch Detector	17%	20%	43%	76.5%
Autocorrelator-based Pitch Detector	52%	54%	60.5%	80.3%

C. Computational Complexity

In this subsection, we compare the computational complexity of the cepstrum, the autocorrelation and the D_yWT pitch detectors. The pitch detector based on the cepstrum of a signal is computationally complex as compared to the other two methods when the signal segment length L is long, since the cepstrum method involves computing the FT, the logarithm of the power spectrum and its inverse FT. The computation of both the sampled autocorrelation and the D_yWT methods involve only a summation of products. In Table II, we list the approximate number of additions and multiplications required to compute the cepstrum, the autocorrelation, and the D_yWT of a signal of length L samples. In general, the number of correlation coefficients p is less than the length of the wavelet, M, that is used to compute the D_yWT, making the autocorrelation method the fastest with the D_yWT method a close second.

D. Segmentation

Here, we discuss the problems that are associated with the segmentation or windowing of a speech signal in the case of all three pitch detectors under consideration. For the cepstrum and the autocorrelation methods, the choice of the segment length is very important. Both methods estimate the *average* pitch period of a length L signal segment and hence they need at least two pitch periods within a chosen segment. If the speech segment is too short, then the algorithms will not be able to estimate the pitch period accurately; if the segment is too long, then these algorithms will not be able to detect the nonstationary variations in the length of the pitch period from period to period. However, in the case of the D_yWT, the choice of the segment length is not very crucial, since we estimate the pitch period by locating the instant at which the glottis closes. For example, in Table III, we have tabulated the results obtained by the D_yWT based event pitch detector after segmenting a synthesized voiced fricative /v/. We have chosen segment lengths of 40 ms, 25.6 ms and 12.8 ms, respectively. Note that in the case of the segment length of 12.8 ms each segment could contain at the most only one complete pitch period since the pitch period of the synthesized signal under consideration is 10 ms. From Table III, it is clear that for all the three different segment lengths, we obtain similar levels of accuracy in the results. The increase in relative error from 2%–6% for decreasing segment lengths is primarily due to edge effects. That is, for a noncausal wavelet, the algorithm does not have enough data to accurately compute the D_yWT when it is less than half the length of the wavelet, L_w, away from either edge of the signal segment. This can be compensated for by partially overlapping successive signal segments by at least $(L_w/2)$.

V. CONVERSATIONAL CONTINUOUS SPEECH

In this section, we provide examples of the D_yWT pitch detector applied to real speech signals. We consider conversational continuous speech "*the seat is weeping the boat*" spoken by an English speaking native male American and "*when the sun light*" spoken by a female American. These two speech signals are sampled at 16 kHz. The pitch period was estimated by segmenting the speech signal using a rectangular window of length $L = 32$ ms and applying the event detection pitch detector based on the D_yWT. In order to take care of the end effects of convolution, we only computed the D_yWT for the block output time between $\left(\frac{L_w}{2}, L - \frac{L_w}{2}\right)$ where L_w is the length of the wavelet. The next segment of the signal is obtained by shifting the window by $S = \frac{L_w}{2}$. In Fig. 4(a), we plot the speech signal "*the seat*" a portion of the sentence *the seat is weeping the boat*, which consists of both voiced and unvoiced segments. In Fig. 4(b), we plot the track of the pitch period estimated by the D_yWT based pitch detector. The D_yWT pitch detector classifies the given segment as voiced or unvoiced, estimates the pitch period during the voiced portion and sets the pitch estimate to zero whenever the algorithm judged the speech to be unvoiced. In Fig. 5(a), we plot the speech signal "*when the sun light*" spoken by a female speaker. This signal also consists of both voiced and unvoiced speech sounds. In Fig. 5(b)–(d), we plot the D_yWT computed at three scales $a = 2^3$, $a = 2^4$, and $a = 2^5$ which corresponds to center frequencies 1000 Hz, 500 Hz, and 250 Hz, respectively. Fig. 5(e) is the plot of the pitch tracked by the D_yWT pitch detector.

Next, we provide a comparison of the three pitch detectors on real speech signals. We plot the speech signal letter "*a*" spoken by a male speaker in Fig. 6(a) and the tracks of the pitch period estimated by the D_yWT, the cepstrum and the autocorrelation based pitch detectors in Fig. 6(b). A window of length $L = 32$ ms was used for all three methods. Fig. 7 represents the tracks of the pitch period estimated by all three methods for a continuous conversational speech signal plotted in Fig. 5(a), spoken by a female speaker. From Fig. 6(b), we can see that all of the three methods give similar estimates of the pitch period in the case of a male speaker. In the case of a female speaker (see Fig. 7), during voiced segments both the autocorrelation method and the D_yWT method give similar estimates; the cepstral estimates have more variation. Both the cepstrum and the autocorrelation methods exhibit errors in classifying the unvoiced signal around 200 ms–220 ms and 260 ms–300 ms. The poor performance of the cepstrum method is due to the small number of harmonics present in the spectra of a

TABLE II
A COMPARISON OF THE COMPUTATIONAL COMPLEXITY OF COMPUTING THE D_yWT, THE CEPSTRUM AND THE AUTOCORRELATION FUNCTION OF A SEGMENT OF THE SIGNAL OF LENGTH L SAMPLES, M = LENGTH OF THE WAVELET, AND P = NUMBER OF CORRELATION COEFFICIENTS

	Computational Complexity		
	Number of adds.	Number of mults.	Number of log ops.
D_yWT Pitch Detector	$2\min(M-1, L-1)L$	$2\min(M, L)L$	0
Cepstrum Pitch Detector	$0(2L\log_2 L)$	$2(O(L\log_2 L)+L)$	L
Autocorrelation Pitch Detector	$(L-1)\times p$	$L\times p$	0

*$0(x)$ means the order of x.

TABLE III
A COMPARISON OF THE RELATIVE ERROR OBTAINED IN THE CASE OF THE D_yWT-BASED PITCH DETECTOR BY SEGMENTING A SYNTHESIZED VOICED FRICATIVE /v/ INTO SEGMENTS OF LENGTH 40 ms, 25.6 ms, AND 12.8 ms

Segment Length	Percentage of Relative Error
40 ms	$\leq 2\%$
25.6 ms	$\leq 2\%$
12.8 ms	$\leq 6\%$

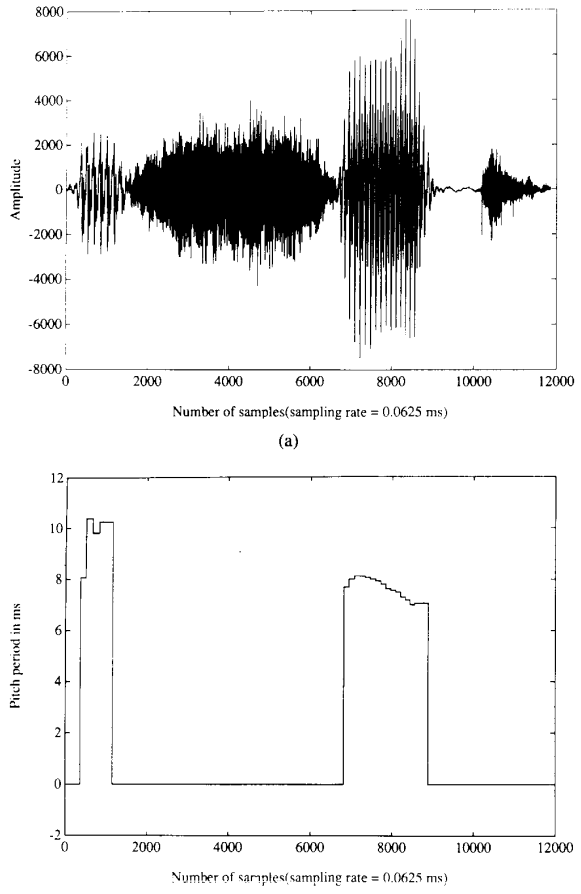

Fig. 4. (a) Continuous conversational speech, "the seat" spoken by an American male speaker. (b) Pitch period tracked by the D_yWT pitch detector.

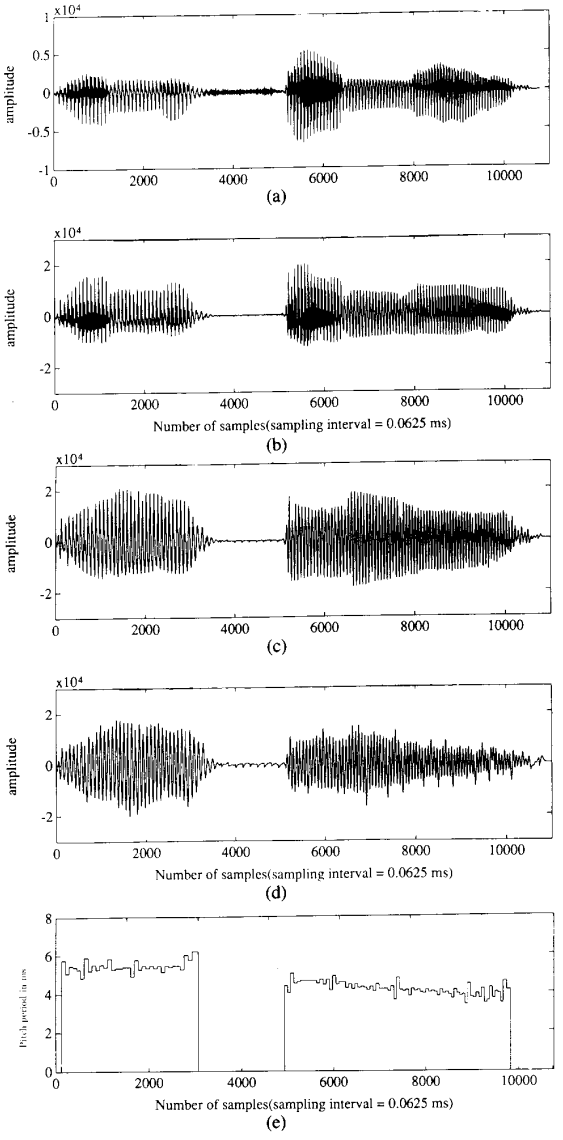

Fig. 5. (a) Continuous conversational speech, "when the sun light" spoken by an American female speaker. (b) D_yWT of "when the sun light" computed at the scale $a = 2^3$. (c) D_yWT of "when the sun light" computed at the scale $a = 2^4$. (d) D_yWT of "when the sun light" computed at the scale $a = 2^5$. (e) Pitch period tracked by the D_yWT pitch detector.

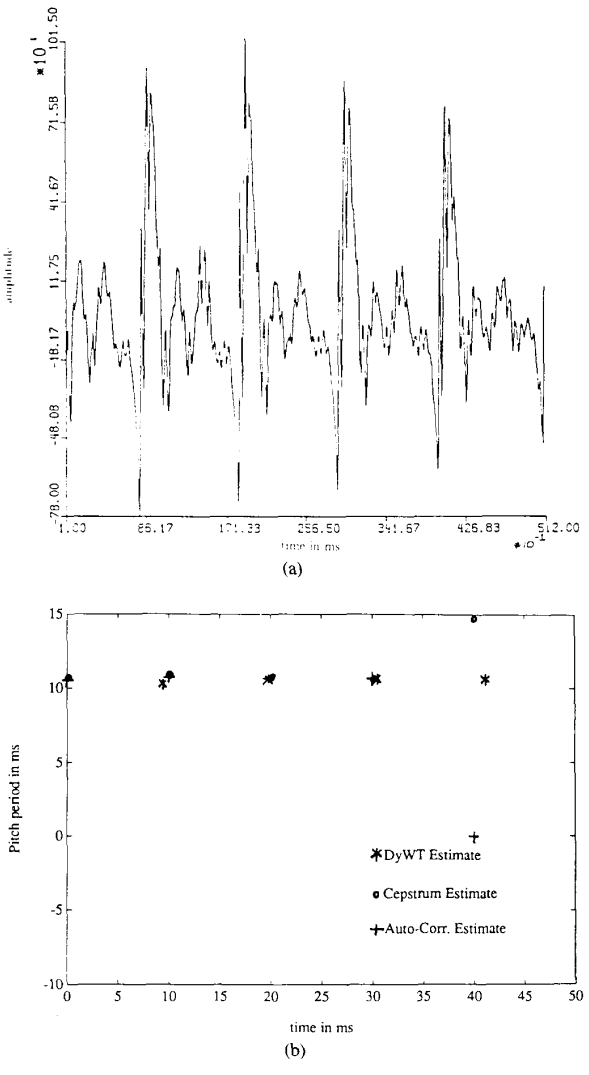

Fig. 6. (a) Real speech signal, "*a*" spoken by a male speaker. (b) Pitch period tracked by the D_yWT, the cepstrum and the autocorrelation pitch detectors for a real speech signal "*a*" spoken by a male speaker.

high-pitched (female) speaker, leading to difficulties in choosing the correct pitch.

Thus, the D_yWT provides accurate estimates of the pitch period for both low-pitched and high-pitched speakers. However, the autocorrelation method performs well for high-pitched speakers (female) and the cepstrum method performs well for low-pitched speakers (male) [1]. Both autocorrelation and the cepstrum methods are unable to detect the variations in the length of the pitch period from period to period and do not perform well in the case of certain sounds such as nasal [1], [14].

VI. Conclusion

In this correspondence, we have described an event based pitch detector using the D_yWT. We have compared its performance with classical pitch detectors with various examples and shown that it exhibits superior performance. The main advantages of the proposed D_yWT method in comparison with the existing pitch detectors are the following: it 1) does not assume stationarity or quasi-stationarity within the analysis window, 2) estimates the pitch period very

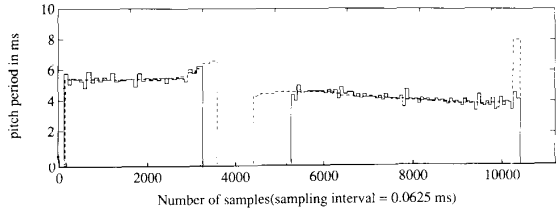

Fig. 7. Pitch period tracked by the D_yWT, the cepstrum and the autocorrelation pitch detectors for a continuous conversational speech signal "when the sun light" spoken by a female speaker. —— D_yWT estimate, ··· Cepstrum estimate, - - - Autocorrelation estimate.

accurately (maximum of 2% relative error for SNR ≥ -18 dB in simulated examples), 3) is suitable for a wide range of pitch periods, 4) can detect the beginning of a pitch period and the number of pitch periods present in a given segment of a speech signal and, hence, can be used for pitch or event synchronous modeling applications, 5) is computationally simple since we need to compute the D_yWT at only two or three scales and 6) exhibits superior performance as compared to the autocorrelation and the cesptrum-based pitch detectors.

References

[1] W. Hess, *Pitch determination of speech signals: algorithms and devices.* Berlin: Springer Verlag, 1983.

[2] H. W. Strube, "Determination of the instant of glottal closure from the speech wave," *J. Acoust. Soc. Amer.*, vol. 56, pp. 1625-1629, Nov. 1974.

[3] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1805-1815, Dec. 1989.

[4] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, pp. 562-570, Dec. 1975.

[5] ——, "Epoch extraction from linear prediction residual for identification and closed glottis interval," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 27, pp. 309-319, Aug. 1979.

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

[7] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293-309, Feb. 1967.

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

[9] M. J. Ross, H. L. Shafer, A. Cohen, R. Frendberg, and H. J. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 353-362, Oct. 1974.

[10] J. D. Wise, J. R. Caprio, and T. W. Parks, "Maximum likelihood pitch estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 418-423, Oct. 1976.

[11] S. G. Mallat and S. Zhong, "Complete signal representation with multiscale edges," tech. rep. RRT-483-RR-219, Courant Inst. of Math. Sci., Dec. 1989.

[12] X. Yang, K. Wang, and S. Shamma, "Auditory representations of acoustic signals," tech. rep. 91-16, Univ. of Maryland, College Park, MD, 1991.

[13] S. Zhong and S. G. Mallat, "Compact image representation from multiscale edges," *Proc. Third Int. Conf. Comput. Vision*, New York, NY, Dec. 1990.

[14] S. Kadambe, "The application of time-frequency and time-scale representations in speech analysis," Ph.D. thesis, Univ. of Rhode Island, Dept. of Elect. Eng., 1991.

[15] S. Kadambe and G. F. Boudreaux-Bartels, "A comparison of wavelet functions for pitch detection of speech signals," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Toronto, Canada, May 1991, pp. 449-452.

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

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