Fundamental Frequency Estimation Based on the Joint Time-Frequency Analysis of Harmonic Spectral Structure

Der-Jenq Liu and Chin-Teng Lin, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a new scheme to analyze the spectral structure of speech signals for fundamental frequency estimation. First, we propose a *pitch measure* to detect the harmonic characteristics of voiced sounds on the spectrum of a speech signal. This measure utilizes the properties that there are distinct impulses located at the positions of fundamental frequency and its harmonics, and the energy of voiced sound is dominated by the energy of these distinct harmonic impulses. The spectrum can be obtained by the fast Fourier transform (FFT); however, it may be destroyed when the speech is interfered with by additive noise. To enhance the robustness of the proposed scheme in noisy environments, we apply the joint time-frequency analysis (JTFA) technique to obtain the adaptive representation of the spectrum of speech signals. The adaptive representation can accurately extract important harmonic structure of noisy speech signals at the expense of high computation cost. To solve this problem, we further propose a fast adaptive representation (FAR) algorithm, which reduces the computation complexity of the original algorithm by 50%. The performance of the proposed fundamental-frequency estimation scheme is evaluated on a large database with or without additive noise. The performance is compared to that of other approaches on the same database. The experimental results show that the proposed scheme performs well on clean speech and is robust in noisy environments.**

***Index Terms*—Adaptive representation, harmonic structure, partial FFT, pitch contour, pitch measure, spectrum analysis.**

I. Introduction

THE estimation of fundamental frequency is an essential component in a variety of speech processing systems such as the speech analysis-synthesis system and speech coding system [1], [2]. The contour of fundamental-frequency (i.e., pitch contour) also plays an important role in language communication [3]–[6]. There are some difficulties in the estimation of fundamental frequency, although it can be observed by eye inspection. First, the voiced speech is not a perfectly periodic waveform because of the variation of fundamental frequency and the movement of vocal tract. Second, it is difficult to estimate the fundamental frequency of low-level voiced speech at its beginning and ending. Third, the performance of estimation will degrade when the speech signal is corrupted by noise.

Manuscript received February 28, 2000; revised April 17, 2001. This work was supported by the National Research Council, R.O.C., under Grant NSC 89-2218-E-009-040. The associate editor coordinating the review of this paper and approving it for publication was Dr. Philip C. Loizou.

The authors are with the Department of Electrical and Control Engineering, National Chiao-Tung University, Hsinchu, Taiwan, R.O.C. (e-mail: william@falcon3.cn.nctu.edu.tw; ctlin@fnn.cn.nctu.edu.tw).

Publisher Item Identifier S 1063-6676(01)07431-4.

Several algorithms for the estimation of fundamental frequency, which may utilize the properties of speech signals in either time-domain or frequency-domain, or in both, have been proposed in [7]. Time-domain estimators operate directly on the speech waveform to estimate the pitch period. The measurements used include peak and valley measurement, zero-crossing and energy measurement, and auto correlation measurement. The class of frequency-domain estimators uses the property that if the signal is periodic in the time domain, then its spectrum will consist of a series of impulses at the fundamental frequency and its harmonics. The measurement for detecting the impulses is made on the spectrum of the signal. The class of hybrid estimators incorporates features of both the time-domain and frequency-domain approaches for pitch detection [8], [9]. The performance of these algorithms is good on clean speech, but degrades rapidly in noisy conditions.

In this paper, we propose a new scheme to analyze the spectral structure of speech signals for fundamental-frequency estimation. First, we propose a new measure, called *pitch measure*, to detect the harmonic characteristic of voiced sound on the spectrum of speech signals. It is proved that this measure will not be trapped by the pitch-doubling or pitch-halving problems. The spectrum for analysis can be obtained by the fast Fourier transform (FFT); however, it may be destroyed when the speech signal is interfered with by additive noise. This will degrade the performance of our scheme based on the FFT-spectrum. To enhance the robustness of the proposed scheme in noisy environments, we apply the joint time-frequency analysis (JTFA) [10], [11] technique to find the adaptive representation of the spectrum of a speech signal. Adaptive representation [12], [13] flexibly decomposes any signal into a linear expansion of waveforms which are selected from a redundant dictionary of functions. It selects the best matching elementary function in some optimal sense to approximate the signal we want. The inspection of the JTFA of a Gaussian-type function reveals that it is localized in time and frequency domains simultaneously such that the problem of cross-term interference [11] is reduced. Hence, we adopt the Gaussian-type functions as the dictionary to characterize the speech signal's time-varying nature in adaptive representation. Since only important factors are used to represent the speech signal, the adaptive representation can accurately extract important harmonic structure from noisy speech signals. However, this is achieved at the expense of high computation cost. To attack this problem, we further propose a fast adaptive representation (FAR) algorithm, which performs partial FFT and reduces the computation complexity of the original algorithm by 50%.

The performance of the proposed fundamental-frequency estimation scheme is evaluated on a large database with or without additive noise. It is compared to that of other approaches on the same database. The comparison results show that the proposed scheme performs well on clean speech and is robust in noisy environments.

The organization of this paper is as follows. In Section II, we propose the pitch measure and study its properties on the speech spectrum. A pitch-tracking algorithm is also proposed in this section to identify continuous pitch contours and make voiced/unvoiced decisions. In Section III, we propose the FAR algorithm to obtain the spectrum of speech signals. The pitch measure is then applied to the FAR-spectrum to form a robust fundamental-frequency estimation scheme. In Section IV, six meaningful objective error measurements to evaluate the performance of a fundamental-frequency estimator are defined, based on which the performance of the proposed and compared schemes is evaluated. Finally, conclusions are made in Section V.

II. Detection of Harmonic Spectral Structure

A. Spectral Analysis

The production of voiced speech can be described by a linear system mathematically [14], [15]. We use $\Theta(\omega)$ to denote the Fourier transform of the impulse response of the vocal tract model $\theta(n)$. Because the excitation source $e(n)$ for voiced speech is essentially a quasi-periodic train of pulses, its Fourier transform can be described as $E(\omega) = (2\pi/P)\sum_{k=-\infty}^{\infty}\delta_a(\omega - k(2\pi/P))$, by the Poisson sum formula, where P is the period of the pulse, or $1/P$ is the fundamental frequency, and the delta function $\delta_a(\omega)$ is the unit impulse function. The voiced speech signal $s(n)$ is modeled in the time domain as the convolution of $e(n)$ and $\theta(n)$. That is, $s(n) = \theta(n) * e(n)$, where $*$ is the convolution operator. Using the convolution property of Fourier transform, we have

$$S(\omega) = \Theta(\omega)E(\omega) = \frac{2\pi}{P}\sum_{k=-\infty}^{\infty}\Theta\left(k\frac{2\pi}{P}\right)\delta_a\left(\omega - k\frac{2\pi}{P}\right) \tag{1}$$

where $S(\omega)$ is the Fourier transform of $s(n)$.

Equation (1) gives an important insight into the spectral structure of voiced sounds; it is a linear combination of the impulses located at harmonics of fundamental frequency. If the harmonic spectral structure can be identified, the corresponding fundamental frequency can also be obtained. The point to do this is to detect the distinct impulses at fundamental frequency and its harmonics. To detect a distinct impulse, we apply two windows, inner window and outer window, on an impulse, where the centers of both windows are located at the center of the impulse. A distinct impulse as well as the two windows are illustrated in Fig. 1. The widths of the two windows in our study are determined experimentally, as described in Section II-C. Based on these two windows, we define three basic indexes on an individual impulse:

1) inner energy, $h_{in}(\omega_c) = \int_{\omega_c - w_{\text{in}}/2}^{\omega_c + w_{\text{in}}/2} S(\omega)\,d\omega$, the area under the curve of spectrum bounded by the inner window;

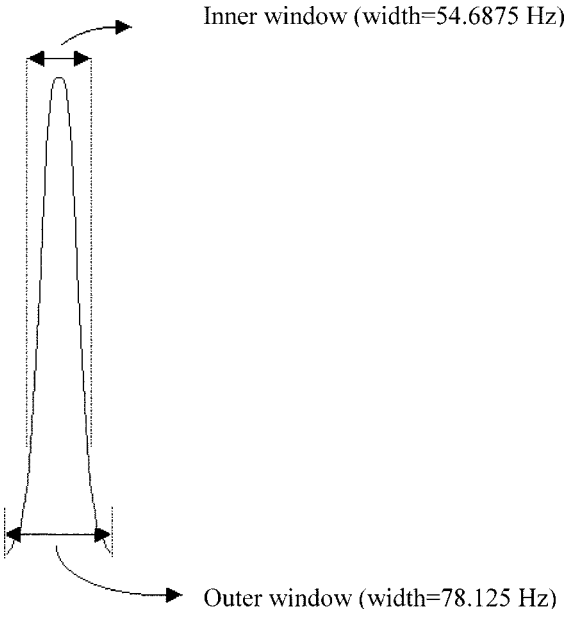

Fig. 1. Obvious impulse with inner and outer windows.

2) outer energy, $h_{\text{out}}(\omega_c) = \int_{\omega_c - w_{\text{out}}/2}^{\omega_c + w_{\text{out}}/2} S(\omega)\,d\omega$, the area under the curve of spectrum bounded by the outer window;
3) total energy: $E = \int_0^{\infty} S(\omega)\,d\omega$, the total area under the curve of spectrum.

If there is a distinct impulse located at frequency ω_c, the values of $h_{\text{in}}(\omega_c)$ and $h_{\text{out}}(\omega_c)$ will be very large.

Based on the above three indexes, we define three measures to identify the harmonic spectral structure of speech signals in the following.

- Energy Measure: The energy measure of a fundamental-frequency candidate, ω_f, is defined as

$$R_E(\omega_f) = \frac{\sum_{n=1}^{K(\omega_f)} h_{\text{in}}(n\omega_f)}{E} \tag{2}$$

subject to the constraint

$$\frac{h_{\text{in}}(n\omega_f)}{h_{\text{out}}(n\omega_f)} \geq \theta_I \quad \text{for each} \quad n \tag{3}$$

where $K(\omega_f)$ is the number of the harmonics of fundamental frequency ω_f, and θ_I is a preset threshold. If a distinct impulse is located at some harmonic $n\omega_f$, the value $h_{\text{in}}(n\omega_f)/h_{\text{out}}(n\omega_f)$ will be large. The constraint in (3) means that only the harmonics with distinct impulses are considered in the calculation of the energy measure. In other words, (2) measures the total energy concentrated on the harmonics with distinct impulses of the fundamental-frequency candidate ω_f. The value of θ_I is set as 0.85 in this study as described in Section II-C. If $\hat{\omega}_f$ is a true fundamental frequency, the value of $R_E(\hat{\omega}_f)$ will be quite large since the voiced-sound energy is dominated by the energies of distinct harmonic impulses. One good property of the energy measure is that it exists no pitch-halving problem; i.e., we always have $R_E(\hat{\omega}_f) >$

$R_E(2\hat{\omega}_f)$. However, the energy measure could lead to the confusion between $\hat{\omega}_f$ and $\hat{\omega}_f/2$, i.e., the pitch-doubling problem. In other words, it could happen that $R_E(\hat{\omega}_f) < R_E(\hat{\omega}_f/2)$. The proof of these properties can be found in Appendix A.

- Impulse Measure: The impulse measure of a fundamental-frequency candidate ω_f is defined as

$$R_I(\omega_f) = \frac{\sum_{n=1}^{K(\omega_f)} h_{\mathbf{in}}(n\omega_f)}{\sum_{n=1}^{K(\omega_f)} h_{\mathrm{out}}(n\omega_f)}. \tag{4}$$

Equation (4) measures if there always exist distinct impulses on the harmonic positions of the fundamental-frequency candidate ω_f. If $\hat{\omega}_f$ is a true fundamental frequency, the value $R_I(\hat{\omega}_f)$ will be close to 1 since a distinct impulse is always located on each harmonic. This situation does not exist on the frequencies other than the fundamental frequency and its harmonics in normal speech signals. In other words, a large impulse-measure value can indicate the fundamental frequency $\hat{\omega}_f$ or its multiples $n\hat{\omega}_f$. Hence, the impulse measure exists no pitch-doubling problem; i.e., we always have $R_I(\hat{\omega}_f) > R_I(\hat{\omega}_f/2)$. However, $R_I(\hat{\omega}_f)$ might not be the maximum over $R_I(\hat{\omega}_f)$, $R_I(2\hat{\omega}_f)$, ... Hence, there could exist confusion between $\hat{\omega}_f$ and $2\hat{\omega}_f$ in the impulse measure (i.e., the pitch-halving problem); it might happen that $R_I(2\hat{\omega}_f) > R_I(\hat{\omega}_f)$. The proof for these properties is given in Appendix B.

The energy measure $R_E(\omega_f)$ and the impulse measure $R_I(\omega_f)$, respectively, capture the two major characteristics of the harmonic spectral structure of voiced speech; there are distinct impulses at the harmonics of fundamental frequency, and the total energy is dominated by these distinct harmonic impulses. Since both of these two measures have large values at the true fundamental frequency simultaneously, we take the product of these two measures to form the final form of our measure for detecting the true fundamental frequency. This measure, called pitch measure, is defined as follows.

- Pitch Measure: The pitch measure at the fundamental frequency candidate, ω_f, is defined as

$$R_P(\omega_f) = R_E(\omega_f) R_I(\omega_f). \tag{5}$$

The equation to estimate the fundamental frequency is

$$\hat{\omega}_f = \arg\max_{\omega_f} R_P(\omega_f). \tag{6}$$

It can be shown that the pitch measure does not have the pitch-doubling and pitch-halving problems (see Appendix C); in other words, we always have $R_P(\hat{\omega}_f) > R_P(\hat{\omega}_f/2)$ and $R_P(\hat{\omega}_f) > R_P(2\hat{\omega}_f)$ for the true fundamental frequency $\hat{\omega}_f$.

One example to illustrate the behavior of the above measures for fundamental frequency estimation on a speech segment is shown in Fig. 2. The fundamental frequency of the speech segment shown in Fig. 2(a) is $\hat{\omega}_f = 233.5$ Hz. We observe that $R_E(\hat{\omega}_f) > R_E(2\hat{\omega}_f)$, $R_E(\hat{\omega}_f) \approx R_E(\hat{\omega}_f/2)$ in Fig. 2(b), and $R_I(\hat{\omega}_f) > R_I(\hat{\omega}_f/2)$, $R_I(\hat{\omega}_f) \approx R_I(2\hat{\omega}_f)$ in Fig. 2(c). Hence, we have $R_P(\hat{\omega}_f) > R_P(2\hat{\omega}_f)$, $R_P(\hat{\omega}_f) > R_P(\hat{\omega}_f/2)$ in Fig. 2(d), and $\hat{\omega}_f = 233.5$ Hz is determined to be a true fundamental frequency.

B. Continuous Pitch-Tracking Algorithm With Voiced/Unvoiced Decision

Applying the pitch measure in (5) and (6) on each frame of a speech signal, we can obtain the estimated fundamental frequency for each frame; whether it is voiced or unvoiced. A pitch-tracking algorithm is then utilized to obtain the continuous pitch contours and make the voiced/unvoiced decision. The algorithm utilizes the property that the pitch curve of voiced sound is continuous in local region. The steps of the proposed pitch-tracking algorithm are as follows.

Step 1) Pitch Detection: Apply the pitch measure in (5) and (6) to find the fundamental frequency of each frame of the input speech signal.

Step 2) Pitch Contour Search: For every two adjacent frames, check if the difference of their fundamental frequencies estimated in Step 1 is less than 12% of either one of these two frequencies, and check if the impulse-measure values of them are both greater than the threshold $\theta_V = 0.8$. If every two adjacent frames pass this checking, they form a portion of one pitch contour with the pitch of each frame being the reciprocal of the fundamental frequency estimated in Step 1. This step will produce a set of piecewise-continuous pitch contours.

Step 3) Continuity Detection: Check if the length of each pitch contour formed in Step 2 is greater than eight frames. If yes, it is recognized to be a continuous pitch contour; otherwise it is discarded.

Step 4) Pitch Doubling/Halving Checking: Track each continuous pitch contour recognized in Step 3 by extending it forward from its beginning and backward from its ending along the frames axis to see if there are pitch doubling or halving errors. If yes, the extended frame with error is added to the current continuous pitch contour, and its fundamental frequency is corrected by multiplying (for pitch doubling) or dividing (for pitch halving) the one estimated in Step 1 by two.

Step 5) V/UV Decision: If a frame is on a continuous pitch contour finally formed in Step 4, it is considered to be voiced; otherwise it is unvoiced.

In the above algorithm, the first checking in Step 2 is to make sure the piecewise continuity of a pitch contour, where the difference "12%" is set by experience. The second checking for impulse-measure values is to make sure that the two adjacent frames are both voiced sound, where the threshold θ_V is set as 0.8 in our study, as described in Section II-C. The pitch contour with short length in frames may not be a voiced contour as detected in Step 3, where the minimum length of eight frames is set by trial and error. Although the fundamental frequencies of voiced frames can always be estimated by the pitch measure, there still exists the possibility of pitch doubling or halving errors, especially at the transition of voiced and unvoiced frames.

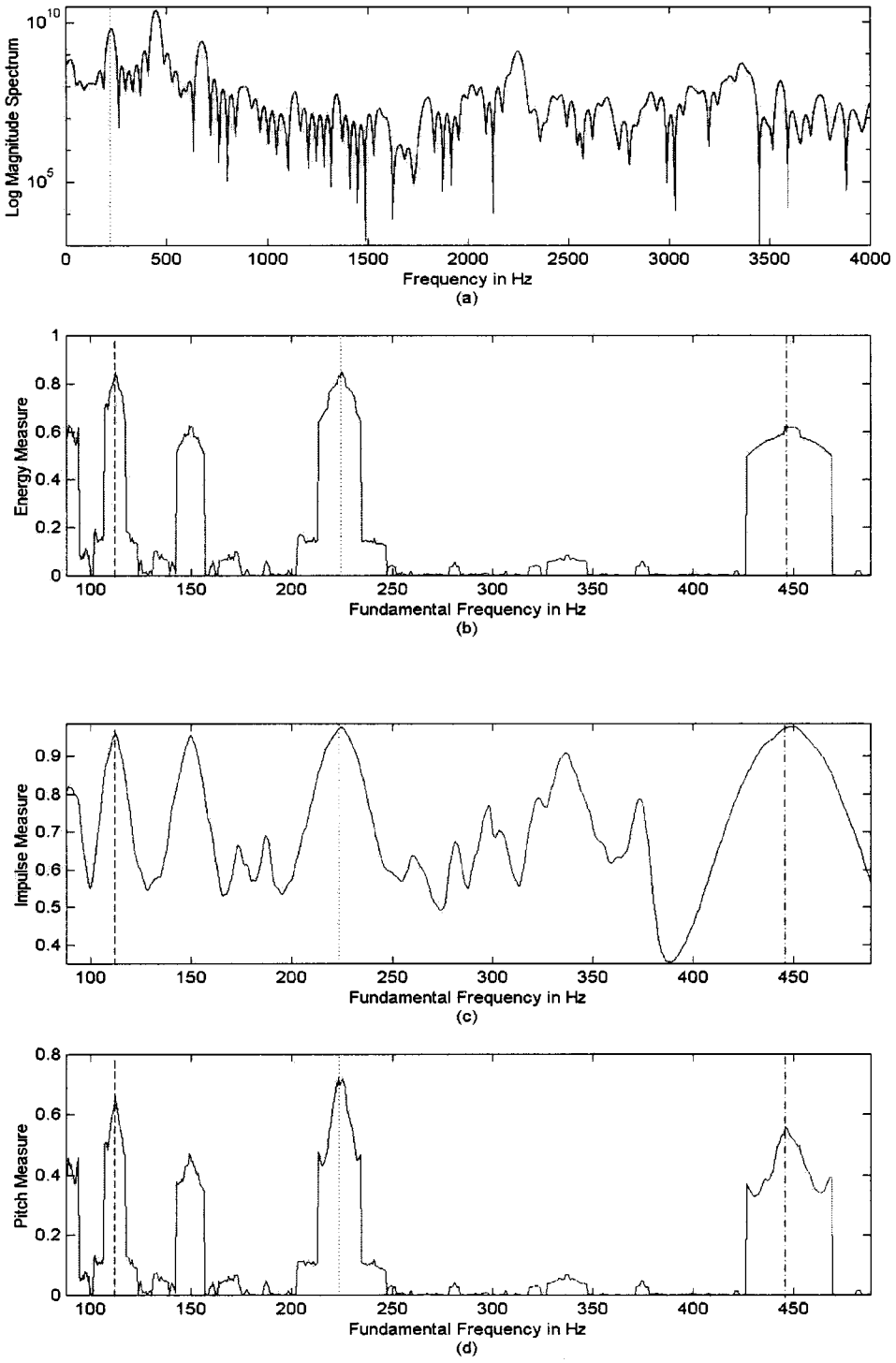

Fig. 2. Illustrations of the proposed measures on one speech frame. (a) Spectrum of the speech signal, where the fundamental frequency ω_f is labeled by the dotted line. (b) Energy measure $R_E(\omega)$ on (a). (c) Impulse measure $R_I(\omega)$ on (a). (d) Pitch measure $R_P(\omega)$ on (a). The frequencies, $\omega_f/2$, ω_f, and $2\omega_f$ are labeled by dashed, dotted, and dash-dotted lines, respectively, in (b), (c), and (d).

Hence, the checking and correction of such errors are done in Step 4 to reinforce the smoothness of the obtained pitch contours.

C. Determination of Window Widths and Threshold Values

In applying the pitch measure and pitch-tracking algorithm, the widths of inner and outer windows (w_{in} and w_{out}), as well as the thresholds θ_I in (3) and θ_V in Step 2 of the pitch-tracking algorithm need to be determined in advance. The widths w_{in} should be chosen such that the energies of all distinct impulses on speech spectrum are included in the numerator of (2) while computing the energy measure at true fundamental frequency. The width w_{out} should be greater than w_{in} to the extend that the impulse measure approaches one while computing the impulse measure at true fundamental frequency. To achieve these goals,

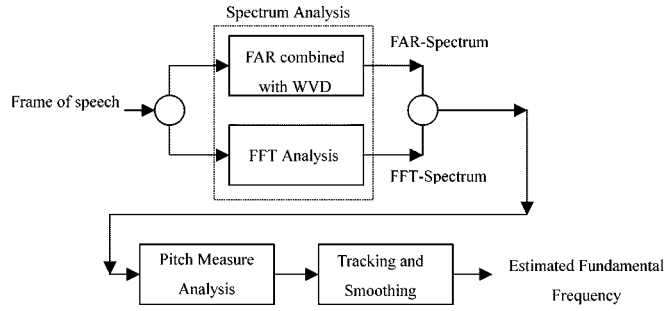

Fig. 3. Flowchart of using the proposed pitch measure for estimating the fundamental frequency based on FFT-spectrum or FAR-spectrum.

we observe the 2048-point FFT-spectrum of a 16- kHz-sampled voiced speech signal from the prepared database. The initial search value for w_{in} is set as three points [equivalent to $(3/2048) \times 16$ k $= 23.4375$ Hz] by visual inspection, since the widths $w_{\text{in}} <$ three points are smaller than the widths of main lobes of most distinct impulses on the spectrum. Starting from $w_{\text{in}} = 3$ points and setting $\theta_I = 0$, we can calculate an average energy-measure value [denoted by $E_{R,\,avg}(w_{\text{in}})$] by averaging the energy measures at true fundamental frequencies of all voiced frames for each $w_{\text{in}} = 3, 4, \ldots$ points. It is observed that the average energy-measure value increases as w_{in} increasing, and then saturates when w_{in} is about 13 points. Hence, the search region for w_{in} is from 3 to 13 points. To reduce the search complexity, we set the ratio $w_{\text{in}}/w_{\text{out}}$ as $5/7$ according to visual inspection, since the width w_{out} satisfying this ratio can cover most side lobes of a distinct impulse and exclude the side lobes of its neighboring impulses. In the w_{in} search region (3, 13) and with the ratio $(w_{\text{in}}/w_{\text{out}}) = 5/7$, we search for the w_{in} and w_{out} values such that the average impulse-measure value of the same prepared voiced frames is above 0.9. In this way, we obtained $w_{\text{in}} = 7$ points $\equiv 54.6875$ Hz, and $w_{\text{out}} = 10$ points $\equiv 78.125$ Hz.

With these w_{in} and w_{out} values, the thresholds θ_I and θ_V are determined according to the average energy-measure value $E_{R,\,avg}(\cdot)$, and the variance of energy measure $E_{R,\,var}(\cdot)$ of the prepared voiced frames. They are

$$\theta_I = \frac{E_{R,\,avg}(w_{\text{in}}) - a_{\theta_I} E_{R,\,var}(w_{\text{in}})}{E_{R,\,avg}(w_{\text{out}}) - a_{\theta_I} E_{R,\,var}(w_{\text{in}})}$$

and

$$\theta_V = \frac{E_{R,\,avg}(w_{\text{in}}) - a_{\theta_V} E_{R,\,var}(w_{\text{in}})}{E_{R,\,avg}(w_{\text{out}}) - a_{\theta_V} E_{R,\,var}(w_{\text{in}})}$$

where a_{θ_I} and a_{θ_V} are the parameters allowing us to adjust the values of θ_I and θ_V to obtain a good result. It is better to have $\theta_I > \theta_V$, so we set $a_{\theta_I} < a_{\theta_V}$. When all the parameters, w_{in}, w_{out}, θ_I, and θ_V are determined, they are fixed and used in all the experiments in the rest of this paper.

D. Experiments

We shall now apply the pitch measure to estimate the fundamental frequency on the speech spectrum obtained by FFT, called FFT-spectrum. The flowchart of the proposed estimation scheme based on FFT-spectrum is shown in Fig. 3. In the experiments, the speech signal, sampled at 16 kHz, is blocked into frames of $N_{FL} = 400$ samples using a rectangular window, with adjacent frames being separated by $M_{FL} = 100$ samples. Then we use 2048-point FFT to obtain the spectrum of each frame. The N_{FL}-sample frame are zero-padded to 2048 samples. Since the sampling rate of speech signals is 16 kHz and the 2048-point FFT is used, the resolution of the estimated fundamental frequency is only 4 Hz. To achieve a better resolution in fundamental frequency, the auto correlation of the periods around the estimated period (the reciprocal of the estimated fundamental frequency) is calculated and the period with the maximum auto correlation value is adopted as pitch period and the corresponding fundamental frequency is calculated as the final estimated result. The performance of the proposed scheme on clean speech of a female is shown in Fig. 4(a), which shows the clean speech waveform and the estimated pitch contour. The proposed scheme is also evaluated on noisy speech. A Gaussian noise was added to the clean speech at SNR value of 4 dB. The estimated pitch contour as well as the corresponding noisy speech waveform are shown in Fig. 4(b). We observe that the performance degrades greatly when the speech is interfered with by additive noise. To enhance the robustness of the proposed scheme in noisy condition, we shall propose a FAR algorithm to obtain the speech spectrum for robust fundamental frequency estimation in the following section.

III. Adaptive Representation of Speech Spectrum

In this section, we shall give the details of adaptive representation, propose a fast algorithm to realize it, and then integrate this algorithm with the pitch-measure-based scheme developed in Section II to form a robust fundamental-frequency estimator.

A. Adaptive Representation

The *adaptive representation* is to find the most important factors that characterize the signals in which we are interested [12], [13]. Adaptive representation flexibly decomposes a signal, $s(t)$, into a linear expansion of waveforms selected from a redundant dictionary of elementary functions, $D = \{h_p(t)\}$

$$s(t) = \sum_p B_p h_p(t) \tag{7}$$

where B_p is a proper coefficient. Adaptive representation allows us to select a set of appropriate elementary functions to best match the structure of a target function for both time and frequency localization. Because of capturing only important factors of speech signals, adaptive representation can provide useful information in noisy environments.

The Gaussian-type function, which is defined as

$$h_p(t) = \left(\frac{\alpha_p}{\pi}\right)^{1/4} \exp\left\{-\frac{\alpha_p}{2}(t - T_p)^2\right\} \exp\{j\omega_p t\} \tag{8}$$

is a natural selection to form the set of elementary functions D for adaptive representation according to the lower bound of the uncertainty principle [11] and the fact that any function can be decomposed into a linear combination of Gaussian-type functions [12], [13]. To see this, we take the Wigner–Ville distribution (WVD) [16] of a Gaussian-type function. The WVD is a tool to study the time-frequency characteristic of a signal; it

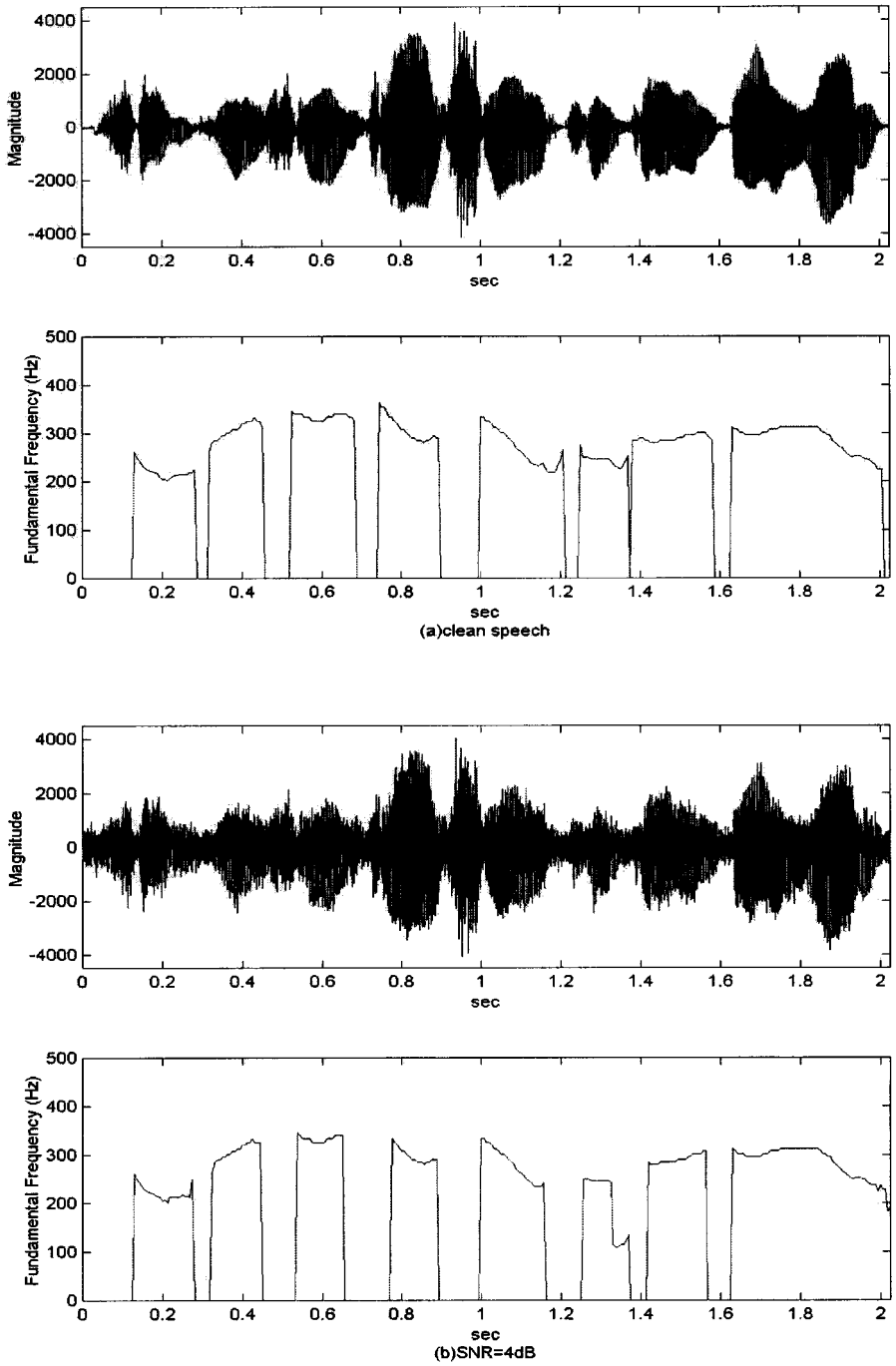

Fig. 4. Performance of the proposed pitch-measure-based fundamental frequency estimation scheme based on the FFT-spectrum of a female's speech. (a) Clean speech signal and the estimated pitch contour. (b) Speech signal with adding noise at SNR $= 4$ dB, and the estimated pitch contour.

calculates the time-dependent power spectrum of a signal. The WVD of a Gaussian-type function $h_p(t)$ is defined by

$$\begin{aligned}\mathrm{WVD}_{h_p}(t,\omega) &= \int_{-\infty}^{\infty} h_p\left(t+\frac{\tau}{2}\right) h_p^*\left(t-\frac{\tau}{2}\right) \exp\{-j\omega\tau\}\, d\tau \\ &= 2\exp\left\{-\left[\alpha_p(t-T_p)^2 + \frac{(\omega-\omega_p)^2}{\alpha_p}\right]\right\}. \quad (9)\end{aligned}$$

This equation indicates that the Gaussian-type function is localized in both time and frequency domains with the time-frequency center located at (T_p, ω_p).

Based on the above analysis, the dictionary we choose in our scheme is the family of Gaussian-type functions defined by $D = \{h_p(t)\}$, where $h_p(t)$ is defined in (8). However, in the three parameters, $\{\alpha_p, T_p, \omega_p\}$ of $h_p(t)$, we fix the value of α_p for all p, since the length of each speech frame is fixed, and set the parameter T_p located at the center of a speech frame. Then the frequency ω_p is the only parameter to be determined in choosing the best $h_p(t)$ [or denoted as $h_p(t;\ \omega_p)$] from D. Since the dictionary D is redundant, there is no unique solution for (7). We need an iterative procedure to select $h_p(t)$ from D successively to best match the structure of the speech signal

$s(t)$. This is done by successive approximations of $s(t)$ with orthogonal projections on elements of D; that is, the coefficients in (7) are determined by

$$B_p = \langle s_p,\, h_p \rangle = \int_t s_p(t) h_p(t)\, dt \tag{10}$$

which reflects the similarity between $s_p(t)$ and $h_p(t)$, where $s_p(t)$ is the residual after the pth iteration of approximating signal $s(t)$ in the direction of $h_p(t)$. The coefficient B_p reflects the signal's local behavior over $h_p(t)$.

Let us start with $p = 0$ and $s_0(t) = s(t)$, which is the original speech signal. The signal $s_p(t)$ can be decomposed into

$$s_p(t) = B_p h_p(t) + s_{p+1}(t) \tag{11}$$

in the sense of

$$\begin{aligned} B_p &= \max_{\omega} \langle s_p,\, h(t;\omega) \rangle \\ &= \max_{\omega} \left(\frac{\alpha}{\pi}\right)^{1/4} \int_t s_p(t) \exp\left\{-\frac{\alpha}{2}(t-T)^2\right\} \exp\{-j\omega t\}\, dt \end{aligned} \tag{12}$$

and

$$\omega_p = \arg\max_{\omega} \langle s_p,\, h(t;\omega) \rangle \tag{13}$$

for $p = 0$. Repeat this process to sub-decompose the residual $s_p(t)$, for $p \geq 1$, by projecting it on an elementary function $h_p(t)$ from D, which has the best match with $s_p(t)$. Finally, we can obtain a set of elementary functions, $\{h_p(t)\}$, selected from D, which most resemble the structure of $s(t)$.

According to the above equation, to find the best-matching Gaussian-type function at the pth iteration, we must take Fourier transform of $s_p(t) \exp\{-(\alpha/2)(t-T)^2\}$ and search for the frequency ω such that the determined Gaussian-type function has the maximum similarity with the pth residual of a speech signal. To gain the best accuracy at lower computation cost, we choose the eight most important Gaussian-type functions to expand a frame of a speech signal. This needs eight times of Fourier transform, where the number "8" is determined by experience.

As was done in the experiments of Section II, we use 2048-point FFT to implement the adaptive representation to obtain the spectrum of a speech signal. At the pth iteration, the pth residual of a N_{FL}-sample frame is zero-padded to 2048 samples. The estimated fundamental frequency for each frame is also finely tuned by the auto correlation method as done in Section II. In the following, we aim at reducing the computation complexity of adaptive representation.

B. Fast Adaptive Representation (FAR) Algorithm

High computation complexity is the major shortcoming of the adaptive representation scheme, especially for the requirement of high frequency resolution. For example, if we take 2048-point FFT, we need $2048/2 \times (\log_2(2048) - 1)$ times of complex multiplications to obtain the adaptive representation in each search iteration, p, for each frame, if the butterfly computation is used. Reducing the computation complexity for real-time applications becomes an important issue. In this section, we propose a fast algorithm to realize the adaptive representation with lower computation complexity. The basic concept of the fast algorithm is that, for each search iteration, we start the search from the frequency of the best-matching Gaussian-type function at lower frequency resolution on the full frequency range, and then increase the search resolution on more focused search region step by step to reach the final desired resolution. In other words, for each (e.g., the pth) search iteration in the FAR algorithm, we start from using smaller point-number FFT to find the raw candidate frequency region, in which the frequency of the Gaussian-type function that can best describe $s_p(t)$ is located. Then, in the next step, we focus our search only on this raw candidate frequency region using larger point-number FFT to obtain a finer candidate frequency region. Continuing such steps, we can finally find the best-matching Gaussian-type function $h_p(t;\, \omega_p)$, whose frequency lies within the desired resolution for the pth search iteration. This kind of "divide-and-conquer" approach reduces the computation complexity obviously.

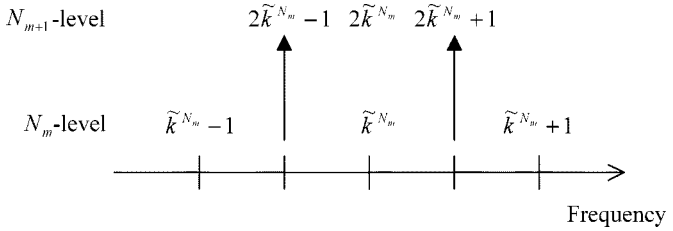

Fig. 5. Illustration of the proposed FAR algorithm; selection of $\hat{k}^{N_{m+1}}$ at N_{m+1}-*level* from the three candidates: $2\hat{k}^{N_m} - 1$, $2\hat{k}^{N_m}$, and $2\hat{k}^{N_m} + 1$ at N_m-*level*.

The proposed FAR algorithm is summarized as follows (see Fig. 5) and the details are given in Appendix D.

At the initial step: We set

$$\hat{k}^{N_0} = \arg\max_{k} \langle s_p,\, h(t;\omega) \rangle \tag{14}$$

at N_0-level, where $\omega = 2\pi(k/N_0)$, $N_0 = N/d$, and $0 \leq k < N_0$.

At the mth step: We set

$$\hat{k}^{N_m} = \arg\max_{k} \langle s_p,\, h(t;\omega) \rangle \tag{15}$$

at N_m-level, where $N_m = (2^m)(N/d)$, $1 \leq m \leq M$, and $k \in \{2\hat{k}^{N_{m-1}} - 1,\, 2\hat{k}^{N_{m-1}},\, 2\hat{k}^{N_{m-1}} + 1\}$.

To implement the procedure proposed above, we define an operation called "partial FFT," which computes only some FFT values we want, in contrast to the traditional FFT, which computes all the FFT values. The computation flow graph for computing 8-point (traditional) FFT values is shown in Fig. 6 [17], where $W_N^l = \exp\{-j2\pi(l/N)\}$. Assume C_0 and C_2 are the two values we want. Then, the solid lines and solid circles in the figure show the partial FFT for computing C_0 and C_2. The partial-FFT computation flow in Fig. 6 reveals that nodes A_0, A_2, B_0, and B_2, as well as nodes A_4, A_6, B_4, and B_6, form a flow of butterfly. The number of complex multiplications to obtain the FFT values of node B_0 and B_2 is 1, if the simplified butterfly computation is used. The number of complex multiplications to obtain the values of node C_0 and C_2 is $1 \times 2 + 2$. By induction, the number of complex multiplications to obtain the FFT values of index k and $k+2$ at N-level

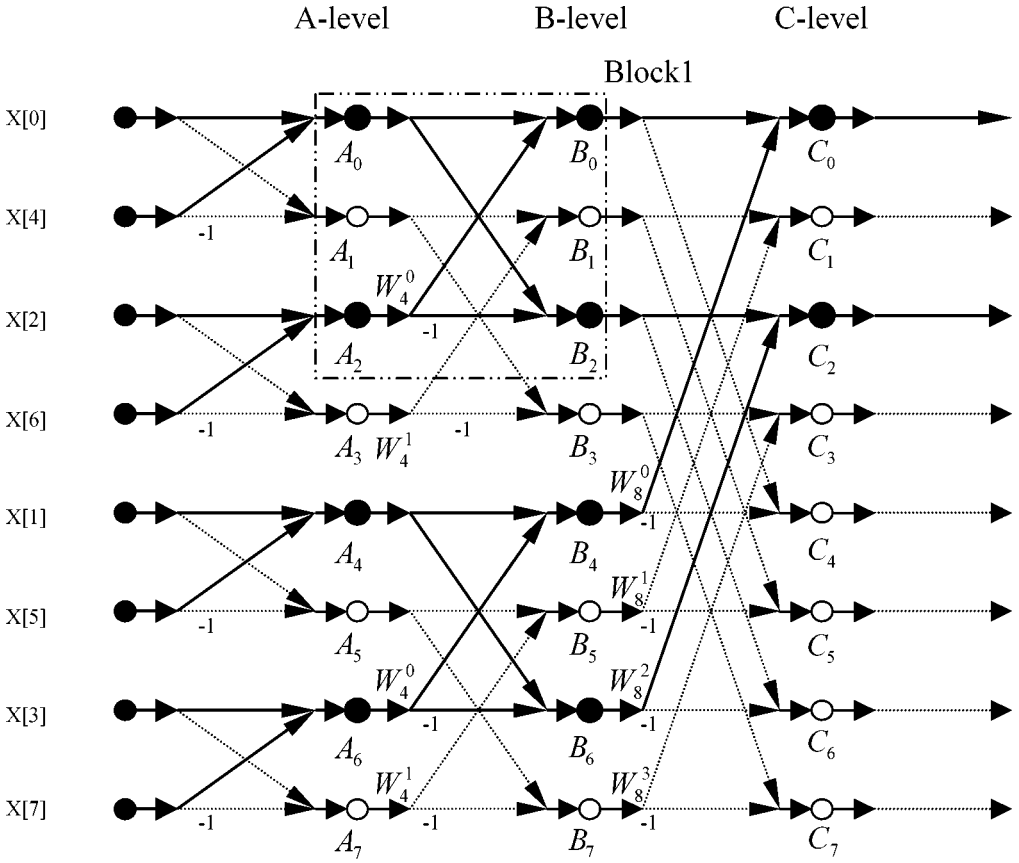

Fig. 6. Computation flow graph of traditional FFT and partial FFT.

is $2^{(\log_2(N)-2)} + \sum_{i=1}^{(\log_2(N)-2)} 2^i = 3 \times 2^{(\log_2(N)-2)} - 2$. In our algorithm, we perform the search steps from 512-level to 2048-level (i.e., $d = 4, M = 2$) to find the frequency ω_p for the best-matching Gaussian-type function $h_p(t;\ \omega_p)$. The total number of complex multiplications in our case is $512/2(\log_2(512) - 1) + \sum_{i=0}^{i=1}(3 \times 2^{\log_2(1024\times 2^i)-2} - 2) = 17 \times 2^8 - 4$, which is much less than 40×2^8, the total number of complex multiplications for 2048-point FFT. That is, by the proposed FAR algorithm, we reduce the number of complex multiplications of the original algorithm by about 50% in each search iteration, p, for each speech frame.

C. Fundamental Frequency Estimation Based on Adaptive Representation

By the FAR algorithm, we can obtain the adaptive representation of a speech frame as in (7). We then take the WVD of (7) and ignore the cross terms to obtain the speech spectrum, called FAR-spectrum, as

$$\mathrm{WVD}_{h_p}(t,\omega) = 2\sum_p |B_p|^2 \cdot \exp\left\{-\left[\alpha_p(t-T_p)^2 + \frac{(\omega-\omega_p)^2}{\alpha_p}\right]\right\}. \quad (16)$$

The reason for ignoring the cross-terms is that the term of double indefinite integral of cross-term over time and frequency is zero. It implies that the cross-term contains zero energy. More detailed information can be found in [11, ch. 8].

We can now use the proposed pitch measure in (5) and (6) to analyze the FAR-spectrum to estimate the fundamental frequency of speech signals. The flowchart of this scheme is shown in Fig. 3. The performance of the scheme based on FAR-spectrum for the clean speech of a female is shown in Fig. 7(a). The corresponding wave of the clean speech is shown in Fig. 4(a). The performance of this FAR-spectrum-based estimation scheme is also evaluated on noisy speech at SNR value of 4 dB shown in Fig. 4(b). The corresponding estimation results are shown in Fig. 7(b). Comparing the estimation results in Fig. 7 and Fig. 4, we observe that both of the two proposed schemes have good performance on clean speech. However, the performance of the FAR-spectrum-based scheme is better than that of the FFT-spectrum-based scheme in noisy condition.

IV. Experimental Results and Comparisons

In this section, we evaluate the performance of the proposed schemes on a large database according to six error measurements, and compare it to the performance of the simplified inverse filter tracking algorithm (SIFT) [9], cepstrum method [18]–[20], and a commercial fundamental-frequency estimation software, ESPS.

A. Testing Database

The prepared database for performance evaluation consists of 50 files of speech utterances spoken by 25 males and 25 females, where the sampling rate of the speech signals is 16 kHz. Every speaker provides one file to the database. These 50 files

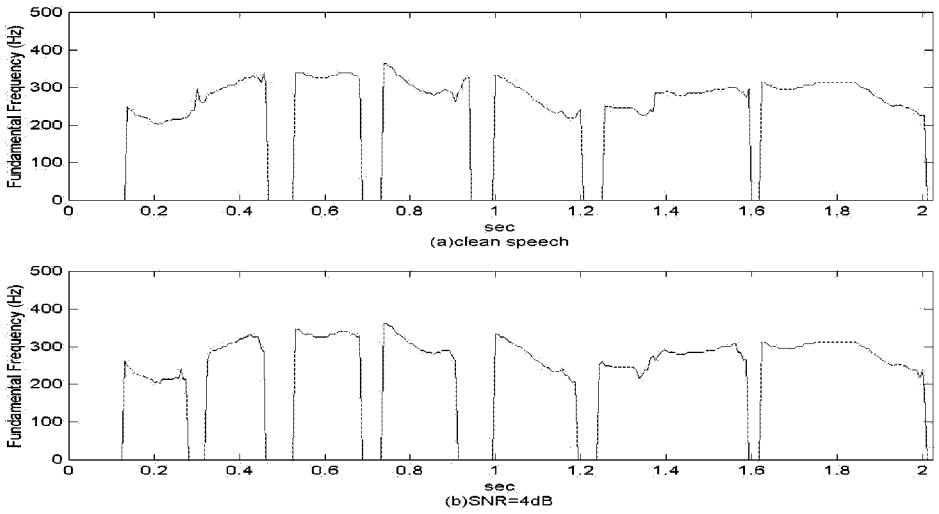

Fig. 7. Performance of the proposed pitch-measure-based fundamental-frequency estimation scheme based on the FAR-spectrum of a female's speech. (a) Estimated pitch contour on the clean speech signal shown in Fig. 4(a). (b) Estimated pitch contour on the speech signal with adding noise at SNR = 4 dB shown in Fig. 4(b).

are selected from the continuous speech database recorded by "Chunghwa Telecommunication Laboratories" in Taiwan. Each speech file is composed of the sentences from an article. The contents of the articles of all the files are different. Each speaker uttered one of the articles in continuous speech type to form a file. As a total, the whole database consists of 50 articles, about 500 sentences, 5000 Chinese characters, with length of 1500 s (240 000 frames).

We also provide a reference of pitch contour for each file in the database. It is obvious that a standard and perfectly labeled database does not exist. A labeled reference database of the pitch contours was generated by visual inspection of the original waveforms by authors. We recognized all the periods of the waveforms displayed on the monitor. This was done by labeling the positions of the beginning and ending of all periods on screen using the action of mouse clicking. At the beginning of a voiced sound, there are some valleys with maximum negative amplitude within the region of one pitch period. Since these valleys can be identified obviously and easily, we labeled these valleys to find the pitch of the waveform. Then we traced along the waveform to find the next valley one by one. It should be noted that the distance between any pair of two adjacent labeled valleys is indeed the pitch period of the speech signal in local region. After recognizing all the pitch periods on the waveform, we determined whether it is voiced or unvoiced sound and calculated the fundamental frequency if it is voiced for each frame. If a frame is at the middle of a voiced sound, it is full of periodic pulses and it is viewed as a voiced frame, and the corresponding pitch period is the average distance of all pairs of two adjacent labeled valleys within the frame. If a frame is at the beginning or ending of a voiced frame, it is viewed as a voiced frame if the length of periodic pulses is over 50% of the frame; otherwise it is viewed as an unvoiced one. The determination of pitch periods for the frames at the beginning or ending of a voiced sound is the same as that for the middle-frame of a voiced sound, i.e., taking the average pitch value. All the labeled positions were recorded and then the pitch-contour references were obtained.

B. Error Measurements

The performance measurements we use in the evaluation include voiced-to-unvoiced (V–UV) and unvoiced-to-voiced (UV–V) error rates, and the error of the estimated fundamental frequency. The first two measurements are used to indicate the accuracy in classifying voiced and unvoiced frames, respectively. The last measurement is to check the deviation of the estimated fundamental frequency from the reference. A V–UV error results from that a voiced frame in the reference is detected as an unvoiced one by the estimation algorithm, and an UV–V error results from that an unvoiced frame in the reference is detected as a voiced frame. These two measurements are defined as the ratio of the frame numbers of V–UV or UV–V errors to the total frame numbers in the database.

The weighted gross pitch error [21] is used to measure the difference between the estimated fundamental frequency and the reference. This measurement is defined as follows:

$$\text{GPE} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{E_k}{E_{\max}}\right)^{1/2}\left|\frac{f_k-\hat{f}_k}{\hat{f}_k}\right| \tag{17}$$

where

K	number of voiced frames in the reference;
E_k	short-time energy of the kth frame;
$E_{\max}$	maximum energy of the frames;
$\hat{f}_k$ and f_k	reference and estimated fundamental frequencies for the kth frame, respectively;
$E_{\max}$	is used for normalization.

The GPE measurement is applied to the voiced frames indicated by the reference database. A good fundamental-frequency estimation algorithm should have lower GPE. If we make an insight into the GPE measurement, we can see that the GPE is proportional to the frame energy and the term $|(f_k - \hat{f}_k/\hat{f}_k)|$. While computing the GPE for one given frame, the value $1/K(E_k/E_{\max})^{1/2}$ has been determined since E_k is fixed for the given frame. The exact GPE value contributed by one given frame depends on the accuracy of the estimated fundamental

frequency. The frame with higher estimation accuracy contributes less to the overall GPE. This means that the estimation accuracy of a frame with larger energy is more important than that with smaller energy. If a voiced frame in the reference is classified as an unvoiced one by the estimation algorithm, then the value of $1/K(E_k/E_{\max})^{1/2}$ is contributed to the overall GPE. That is, the maximum GPE value contributed by one frame is $1/K(E_k/E_{\max})^{1/2}$ when a V–UV error occurs. Thus, the GPE measurement indicates not only the difference between the estimated fundamental frequency and the reference, but also the V–UV error.

In addition to the above three performance measurements, Rabiner [7] suggested three other measurements, gross error count (GEC), fine pitch error—average value (FPEAV), fine pitch error—standard deviation (FPESD). These three measurements are also adopted to evaluate the performance of various fundamental-frequency estimation algorithms in this paper. A voiced frame results in a gross pitch period error if $e_k = |(1/f_k) - (1/\hat{f}_k)| > 1$ ms, where k represents the frame index. Gross error count is defined as the ratio of the frame numbers with gross pitch period error to the total frame numbers. A fine pitch error occurs when $e_k \leq 1$ ms. The average value of fine pitch errors is defined as $\overline{e} = (1/N_{\text{fine}}) \sum_{k=1}^{N_{\text{fine}}} e_k$, where N_{fine} is the number of fine pitch errors. The standard deviation of fine pitch errors is defined as

$$\sigma_e = \sqrt{\frac{1}{N_{\text{fine}}} \sum_{k=1}^{N_{\text{fine}}} e_k^2 - \overline{e}^2}.$$

C. Performance Evaluation and Discussion

Using the proposed algorithms, i.e., the pitch-measure-based estimation scheme based on FFT-spectrum or FAR-spectrum, we can obtain the estimated pitch contour of each file in the database. The performance of the proposed schemes is also evaluated in adding white noisy conditions with different SNR values 4 dB, 2 dB, and 0 dB. The evaluation results for male and female speakers are given in Tables I and II, respectively. Both tables include the estimation results of the proposed schemes based on FAR-spectrum and FFT-spectrum. The widths of inner and outer windows used in energy measure and impulse measure as well as the values of θ_I and θ_V are the same for all speakers. These results are also compared against those of SIFT [9], cepstrum method [18]–[20], and ESPS on the same database. In SIFT, a spectrally flattened time waveform is first obtained and the auto correlation measurement is made on the waveform to estimate the pitch period. In the cepstrum method, the detection of peaks on cepstrum is used to detect the pitch period. The ESPS is the Entropic Signal Processing System (or Software) designed by Entropic Research Laboratory. We used one function of this software, called Get-f0, which performs fundamental-frequency estimation using the normalized cross correlation function, dynamic programming, and a robust algorithm for pitch tracking (RAPT) [22]. The same database and references are used to evaluate the performance of SIFT, cepstrum method, and ESPS in clean and noisy conditions. The corresponding evaluation results are also shown in Tables I and II.

TABLE I
PERFORMANCE OF THE PROPOSED FUNDAMENTAL-FREQUENCY ESTIMATION SCHEMES AND SIFT, CEPSTRUM METHOD, AND ESPS IN DIFFERENT NOISE CONDITIONS FOR MALE SPEAKERS

Noise Condition	Error Measurements	FAR-Spectrum	FFT-Spectrum	SIFT	Cepstrum method	ESPS
clean	GPE(%)	0.76	0.86	1.34	1.93	0.91
	V-UV(%)	2.04	2.37	5.09	8.28	3.41
	UV-V(%)	1.78	4.56	2.56	2.15	1.29
	GEC(%)	2.66	3.22	5.41	8.61	3.91
	FPEAV(ms)	2.04e-2	1.60e-2	3.49e-2	6.33e-2	-1.83e-2
	FPESD(ms)	1.72e-1	1.74e-1	1.73e-1	1.96e-1	1.74e-1
4dB	GPE(%)	1.30	1.41	2.48	9.88	3.57
	V-UV(%)	3.82	4.09	8.08	30.16	14.53
	UV-V(%)	1.55	2.89	2.34	0.73	0
	GEC(%)	4.54	5.31	8.44	30.32	14.62
	FPEAV(ms)	2.04e-2	1.28e-2	3.36e-2	6.48e-2	-2.31e-2
	FPESD(ms)	1.86e-1	1.89e-1	1.82e-1	1.92e-1	1.36e-1
2dB	GPE(%)	2.12	2.35	4.66	15.98	8.45
	V-UV(%)	5.54	5.86	12.77	39.27	25.28
	UV-V(%)	1.47	2.52	2.09	0.38	0
	GEC(%)	6.29	7.51	13.08	39.40	25.33
	FPEAV(ms)	1.93e-2	1.09e-2	3.32e-2	6.06e-1	-2.41e-2
	FPESD(ms)	1.97e-1	2.00e-1	1.86e-1	1.83e-1	1.23e-1
0dB	GPE(%)	4.58	5.29	11.04	25.73	20.51
	V-UV(%)	9.49	10.56	22.93	48.90	42.21
	UV-V(%)	1.34	2.02	1.51	0.16	0
	GEC(%)	10.31	12.53	23.18	48.97	42.27
	FPEAV(ms)	2.00e-2	8.36e-3	3.08e-2	5.56e-2	-2.57e-2
	FPESD(ms)	2.12e-1	2.13e-1	1.86e-1	1.78e-1	1.05e-1

TABLE II
PERFORMANCE OF THE PROPOSED FUNDAMENTAL-FREQUENCY ESTIMATION SCHEMES AND SIFT, CEPSTRUM METHOD, AND ESPS IN DIFFERENT NOISE CONDITIONS FOR FEMALE SPEAKERS

Noise Condition	Error Measurements	FAR-Spectrum	FFT-Spectrum	SIFT	Cepstrum method	ESPS
clean	GPE(%)	0.82	1.31	1.19	1.69	1.07
	V-UV(%)	2.33	3.90	4.12	6.44	3.35
	UV-V(%)	2.63	3.70	5.60	3.67	1.94
	GEC(%)	3.03	4.69	4.42	6.76	4.44
	FPEAV(ms)	-3.97e-3	-4.98-3	7.99e-3	2.69e-2	-2.37e-2
	FPESD(ms)	1.02e-1	1.01e-1	1.11e-1	1.37e-1	1.05e-1
4dB	GPE(%)	1.26	3.89	2.52	6.20	3.55
	V-UV(%)	3.70	11.06	8.34	22.35	15.35
	UV-V(%)	2.08	2.38	4.22	1.51	0
	GEC(%)	4.19	16.39	8.56	22.49	15.86
	FPEAV(ms)	-9.85e-4	1.45e-4	1.40e-2	3.60e-2	-2.14e-2
	FPESD(ms)	1.22e-1	1.06e-1	1.68e-1	1.40e-1	7.9e-2
2dB	GPE(%)	1.98	7.60	4.65	11.04	7.96
	V-UV(%)	5.30	16.84	13.47	31.60	25.87
	UV-V(%)	1.88	1.84	3.39	0.94	0
	GEC(%)	5.67	26.79	13.67	31.68	26.44
	FPEAV(ms)	-1.01e-4	2.29-3	1.58e-2	3.95e-2	-2.11e-2
	FPESD(ms)	1.34e-1	1.07e-1	1.20e-1	1.40e-1	7.11e-2
0dB	GPE(%)	4.05	15.86	11.48	19.69	18.74
	V-UV(%)	8.96	25.87	25.59	42.23	42.19
	UV-V(%)	1.63	1.35	2.19	0.49	0
	GEC(%)	9.30	41.95	25.77	42.26	42.62
	FPEAV(ms)	1.09e-3	4.33e-3	1.57e-2	4.47e-2	-2.24e-2
	FPESD(ms)	1.49e-1	1.10e-1	1.18e-1	1.41e-1	6.36e-2

TABLE III
COMPUTATION TIME COST OF THE COMPARED ALGORITHMS FOR EACH SPEECH FRAME

	FAR-Spectrum	FFT-Spectrum	SIFT	Cepstrum method	ESPS
CPU	Intel PIII 500	Intel PIII 500	Intel PIII 500	Intel PIII 500	HP715-100
Time Cost	54.4ms	12.4ms	5.43ms	12ms	7ms

Discussion on the results listed in Tables I and II is made here. The results on the GPE measurement showed that the FAR-spectrum-based scheme performed better than other compared algorithms on clean speech. That is, the fundamental frequencies of most important voiced frames with high energy were successfully estimated by this scheme. The cepstrum method failed at some tail portion of voiced sound for both male and female speakers. In noisy conditions, the results on the GPE measurement showed that the FAR-spectrum-based scheme was obviously superior to all the other algorithms. The results also indicated that the performance of the proposed FFT-spectrum-based scheme degraded in noisy conditions, especially for female speakers, although it had good performance on clean speech.

The results on the GEC measurement showed that the FAR-spectrum-based scheme performed better than the other algorithms. The FFT-spectrum-based scheme performed well for male speakers, however, it failed for female speakers. The measurements of V–UV and UV–V errors provided several interesting results. These categories cannot be examined separately because they are often intimately related. For example, a V–UV detector which is biased toward the category voiced will generally have a low V–UV error rate, but in compensation will have a high UV–V error rate [7]. The results on V–UV error and UV–V error showed that our two proposed schemes and SIFT are examples of this case. On the other hand, ESPS and cepstrum method are examples opposite to this case. A simple threshold on one or more measurements to classify a frame as voiced or unvoiced is used for the compared algorithms. In our study, a continuous pitch-tracking algorithm was adopted to make voicing decision. The property of continuity on pitch contours and the impulse measure were utilized for V/UV decision. In our thinking, it is important to detect the fundamental frequency of one frame if it is voiced, so we adjusted the threshold θ_V in the pitch-tracking algorithm to achieve a low V–UV error rate. It should be noted that the better performance of the proposed FAR-spectrum-based scheme in noisy conditions is gained at the expense of higher computation complexity. The computation time cost of all the compared algorithms for each speech frame is listed in Table III.

V. CONCLUSIONS

In this paper, we first proposed a pitch measure to detect the harmonic spectral structure of speech signals, based on which a new fundamental-frequency estimation scheme was developed. This scheme can analyze the spectrum of a speech signal and produce the corresponding pitch contours. Although the spectrum of a speech signal can be obtained by the traditional FFT, it is easily contaminated by additive noise. To enhance the robustness of the proposed estimation scheme, we developed a FAR algorithm to obtain the spectrum of a speech signal. This fast algorithm was based on the techniques of adaptive representation, divide-and-conquer, and partial FFT. The obtained robust spectrum, called FAR-spectrum, was then analyzed by the pitch-measure-based estimation scheme to obtain the pitch contours. Experimental results have demonstrated the robustness and accuracy of the proposed fundamental-frequency estimation scheme based on the FAR-spectrum, especially in comparison with SIFT, cepstrum method, and a commercial software, ESPS. The superiority of the proposed scheme is gained at the expense of higher computation cost. Although the proposed fast algorithm has reduced the computation complexity of the original algorithm by 50%, it still takes longer than the compared counterparts. The goal of our future work is to further reduce this gap of computation cost.

APPENDIX A
PROOF FOR THE PROPERTIES OF ENERGY MEASURE

The energy measure at $\hat{\omega}_f/2$, i.e., $R_E(\hat{\omega}_f/2)$ is

$$\begin{aligned} &R_E(\hat{\omega}_f/2) \\ &\quad= \left(\sum_{n \text{ is even}} h_{\text{in}}\left(n\frac{\hat{\omega}_f}{2}\right) + \sum_{n \text{ is odd}} h_{\text{in}}\left(n\frac{\hat{\omega}_f}{2}\right)\right)/E \\ &\quad= R_E(\hat{\omega}_f) + \sum_{n \text{ is odd}} h_{\text{in}}\left(n\frac{\hat{\omega}_f}{2}\right)/E \end{aligned} \tag{18}$$

subject to (3). The value of the second term in the last equation is very small such that it can be neglected, because there is no distinct impulse located on $n(\hat{\omega}_f/2)$ for odd n. Therefore, $R_E(\hat{\omega}_f/2) \approx R_E(\hat{\omega}_f)$ and we might determine by mistake the fundamental frequency to be $\hat{\omega}_f/2$, instead of $\hat{\omega}_f$, according to the energy measure alone. It should be noted that there is no confusion between $\hat{\omega}_f$ and $2\hat{\omega}_f$, since $R_E(\hat{\omega}_f) = R_E(2\hat{\omega}_f) + \sum_{n \text{ is odd}} h_{\text{in}}(n\hat{\omega}_f)/E$ subject to (3). The value of the second term in the equation cannot be neglected, so we have $R_E(\hat{\omega}_f) > R_E(2\hat{\omega}_f)$. Hence, the energy measure does not have the pitch-halving problem.

APPENDIX B
PROOF FOR THE PROPERTIES OF IMPULSE MEASURE

Subtracting $R_I(2\hat{\omega}_f)$ from $R_I(\hat{\omega}_f)$, we obtain

$$\begin{aligned} &R_I(\hat{\omega}_f) - R_I(2\hat{\omega}_f) \\ &\quad= \frac{\sum_{n \text{ is even}} h_{\text{in}}(n\hat{\omega}_f) + \sum_{n \text{ is odd}} h_{\text{in}}(n\hat{\omega}_f)}{\sum_{n \text{ is even}} h_{\text{out}}(n\hat{\omega}_f) + \sum_{n \text{ is odd}} h_{\text{out}}(n\hat{\omega}_f)} \\ &\qquad- \frac{\sum_{n \text{ is even}} h_{\text{in}}(n\hat{\omega}_f)}{\sum_{n \text{ is even}} h_{\text{out}}(n\hat{\omega}_f)}. \end{aligned} \tag{19}$$

The condition for satisfying the inequality $R_I(\hat{\omega_f}) > R_I(2\hat{\omega_f})$ is

$$\frac{\sum_{n \text{ is odd}} h_{\text{in}}(n\hat{\omega}_f)}{\sum_{n \text{ is odd}} h_{\text{out}}(n\hat{\omega}_f)} > \frac{\sum_{n \text{ is even}} h_{\text{in}}(n\hat{\omega}_f)}{\sum_{n \text{ is even}} h_{\text{out}}(n\hat{\omega}_f)}.$$

Since it always appears that

$$\frac{\sum_{n \text{ is odd}} h_{\text{in}}(n\hat{\omega}_f)}{\sum_{n \text{ is odd}} h_{\text{out}}(n\hat{\omega}_f)} \approx \frac{\sum_{n \text{ is even}} h_{\text{in}}(n\hat{\omega}_f)}{\sum_{n \text{ is even}} h_{\text{out}}(n\hat{\omega}_f)}$$

we might determine by mistake the fundamental-frequency to be $2\hat{\omega}_f$, instead of $\hat{\omega}_f$, according to the impulse measure alone. It should be noted that the confusion between $\hat{\omega}_f$ and $\hat{\omega}_f/2$ (i.e., the pitch-doubling problem) doesn't exist in the impulse measure. This can be observed from subtracting $R_I(\hat{\omega}_f/2)$ from $R_I(\hat{\omega}_f)$

$$\begin{aligned} &R_I(\hat{\omega}_f) - R_I(\hat{\omega}_f/2) \\ &= \frac{\sum_{n \text{ is even}} h_{\text{in}}\left(\frac{n\hat{\omega}_f}{2}\right)}{\sum_{n \text{ is even}} h_{\text{out}}\left(\frac{n\hat{\omega}_f}{2}\right)} \\ &\quad - \frac{\sum_{n \text{ is even}} h_{\text{in}}\left(n\frac{\hat{\omega}_f}{2}\right) + \sum_{n \text{ is odd}} h_{\text{in}}\left(n\frac{\hat{\omega}_f}{2}\right)}{\sum_{n \text{ is even}} h_{\text{out}}\left(n\frac{\hat{\omega}_f}{2}\right) + \sum_{n \text{ is odd}} h_{\text{out}}\left(n\frac{\hat{\omega}_f}{2}\right)}. \end{aligned} \quad (20)$$

The condition

$$\frac{\sum_{n \text{ is even}} h_{\text{in}}\left(\frac{n\hat{\omega}_f}{2}\right)}{\sum_{n \text{ is even}} h_{\text{out}}\left(\frac{n\hat{\omega}_f}{2}\right)} > \frac{\sum_{n \text{ is odd}} h_{\text{in}}\left(\frac{n\hat{\omega}_f}{2}\right)}{\sum_{n \text{ is odd}} h_{\text{out}}\left(\frac{n\hat{\omega}_f}{2}\right)}$$

always holds, because there is no distinct impulse located on $n(\hat{\omega}_f/2)$ for odd n. Therefore, we always have $R_I(\hat{\omega}_f) > R_I(\hat{\omega}_f/2)$.

Appendix C
Proof for the Properties of Pitch Measure

We assume that the true fundamental frequency is $\hat{\omega}_f$. Then the most possible frequencies that are easily confused with the fundamental frequency are $\hat{\omega}_f/2$ and $2\hat{\omega}_f$. We now compare $R_P(\hat{\omega}_f)$ to $R_P(\hat{\omega}_f/2)$ and $R_P(2\hat{\omega}_f)$. The pitch measure at $\hat{\omega}_f$, $\hat{\omega}_f/2$, and $2\hat{\omega}_f$ are

$$\begin{aligned} R_P(\hat{\omega}_f) &= R_E(\hat{\omega}_f) R_I(\hat{\omega}_f) \\ R_P(\hat{\omega}_f/2) &= R_E(\hat{\omega}_f/2) R_I(\hat{\omega}_f/2) \\ R_P(2\hat{\omega}_f) &= R_E(2\hat{\omega}_f) R_I(2\hat{\omega}_f) \end{aligned} \quad (21)$$

respectively. Due to the facts that $R_E(\hat{\omega}_f) \approx R_E(\hat{\omega}_f/2)$, $R_I(\hat{\omega}_f) > R_I(\hat{\omega}_f/2)$, $R_E(\hat{\omega}_f) > R_E(2\hat{\omega}_f)$, and $R_I(\hat{\omega}_f) \approx R_I(2\hat{\omega}_f)$, we have $R_P(\hat{\omega}_f) > R_P(\hat{\omega}_f/2)$ and $R_P(\hat{\omega}_f) > R_P(2\hat{\omega}_f)$. Hence the pitch measure at the true fundamental frequency, $\hat{\omega}_f$, has the maximum value, and will not cause confusion at $\hat{\omega}_f/2$ and $2\hat{\omega}_f$. In other words, the pitch measure does not have the pitch-doubling and pitch-halving problems.

Appendix D
Details of FAR Algorithm

We now describe the proposed FAR algorithm in details. In the following, we assume that we are looking for the Gaussian-type function $h_p(t)$, which best describes the (residual) speech signal $s_p(t)$, at the pth search iteration for adaptive representation. Suppose we want to take N-point FFT of the N_{FL}-sample frames of the speech signal to achieve the desired resolution of the estimated fundamental frequency. At the initial step of the FAR algorithm, we take (N/d)-point FFT of the speech signal on the full frequency range, where d is a preset integer determining the number of steps for searching the best-matching Gaussian-type function $h_p(t)$. The N_{FL}-sample frames are zero-padded to (N/d) samples such that the spectrum can be obtained by (N/d)-point FFT. Then at the following steps, we take $(k(N/d))$-point FFT, for $k = 2^m$, and $1 \le m \le M = \log_2(d)$, of $s_p(t)$ on the frequency region selected by the previous $(m-1)$th step. The N_{FL}-sample frames are zero-padded to $(k(N/d))$ samples as done at the initial step. For convenience of explanation, we divide the frequency axis into different levels $\{L_{N_m}\}$ of resolution, where the N_m-level is defined as $L_{N_m} = \{\omega = 2\pi(k/N_m), 0 \le k < N_m\}$, where $0 \le m \le M = \log_2(d)$, and $N_m = (2^m)N/d$. The index m represents the mth step to search the frequency ω_p whose corresponding Gaussian-type function, $h_p(t;\ \omega_p)$, best describes $s_p(t)$, where $m = 0$ means the initial step. We can see that $L_{N_m} \subset L_{2N_m}$. At the initial step (i.e., $m = 0$), we take N_0-point FFT of the speech signal, and find the index $\hat{k}^{N_0}$, called the chosen index at N_0-level, such that $\hat{k}^{N_0} = \arg\max_k \langle s_p, h(t;\ \omega)\rangle$, where $\omega = 2\pi(k/N_0)$, $0 \le k < N_0$, and $N_0 = N/d$. At the next step, we set the search region centered at index $k = 2\hat{k}^{N_0}$ and bounded by $[k-1, k+1] = [2\hat{k}^{N_0} - 1, 2\hat{k}^{N_0} + 1]$ at N_1-level, where $N_1 = 2(N/d)$, as illustrated in Fig. 5. Then we select the index $\hat{k}^{N_1}$, called the chosen index at N_1-level, from $\{2\hat{k}^{N_0} - 1, 2\hat{k}^{N_0}, 2\hat{k}^{N_0} + 1\}$, whose corresponding FFT's value (the spectrum magnitude after FFT) is maximum. This is equivalent to taking N_1-point FFT and selecting the frequency with the maximum FFT value among three candidates, $2\pi(2\hat{k}^{N_0} - 1)/N_1$, $2\pi(2\hat{k}^{N_0})/N_1$, and $2\pi(2\hat{k}^{N_0} + 1)/N_1$. Similarly, at the third step, we can obtain the chosen index at N_2-level, $\hat{k}^{N_2}$, from the three candidates, $2\hat{k}^{N_1} - 1$, $2\hat{k}^{N_1}$, and $2\hat{k}^{N_1} + 1$ at N_2-level, where $N_2 = 4(N/d)$. We carry on this iterative step-by-step procedure until the estimated fundamental frequency lies within the desired resolution at N_M-level after M steps. The chosen index at N_M-level, $\hat{k}^{N_M} = \hat{k}^{2^M(N/d)} = \hat{k}^N$, is what we look for finally. We pick the Gaussian-type function with the frequency corresponding to $\hat{k}^N$ (i.e., $2\pi(\hat{k}^N)/N$) as the pth elementary function at the pth search iteration for adaptive representation.

REFERENCES

[1] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception*. New York: Springer-Verlag, 1972.

[2] A. V. McGree and T. P. Barnwell, III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 242–250, July 1995.

[3] S. H. Chen and Y. R. Wang, "Tone recognition of continuous Mandarin speech based on neural networks," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 146–150, Mar. 1995.

[4] T. Lee, P. C. Ching, L. W. Chan, Y. H. Cheng, and B. Mak, "Tone recognition of isolated Cantones syllables," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 204–209, May 1995.

[5] L. S. Lee, C. Y. Tseng, H. Y. Gu, F. H. Liu, C. H. Chang, Y. H. Lin, Y. Lee, S. L. Tu, S. H. Hsieh, and C. H. Chen, "Golden Mandarin (I)—A real-time Mandarin speech dictation machine for Chinese language with very large vocabulary," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 158–178, Apr. 1998.

[6] S. Potisuk, M. P. Harper, and J. Gandour, "Classification of Thai tone sequences in syllable-segmented speech using the analysis-by-synthesis method," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 95–102, Jan. 1999.

[7] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 399–417, Oct. 1976.

[8] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 333–338, May 1999.

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

[10] B. Boashash, *Time-Frequency Signal Analysis*. New York: Wiley, 1992.

[11] S. Qian and D. Chen, *Joint Time-Frequency Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[12] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.

[13] S. Qian and D. Chen, "Signal representation using adaptive normalized Gaussian functions," *Signal Process.*, vol. 36, pp. 1–11, 1994.

[14] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[15] J. G. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, 1993.

[16] T. A. C. M. Classen and W. F. G. Mecklenbräuker, "The Wigner distribution—A tool for time-frequency signal analysis—Parts I, II, III," *Philips J. Res.*, vol. 35, pp. 217–250, 276–300, 372–389, 1980.

[17] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[18] R. W. Schafer and L. R. Rabiner, "Digital representations of speech signals," *Proc. IEEE*, vol. 63, no. 4, pp. 662–677, 1975.

[19] ——, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, pp. 634–648, Feb. 1970.

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

[21] B. G. Secrest and G. R. Doddington, "Postprocessing techniques for voice pitch trackers," in *Proc. IEEE ICASSP'82*, 1982, pp. 172–175.

[22] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. New York: Elsevier, 1995.

Der-Jenq Liu received the B.S. degree in mechanical engineering from the National Central University, Chung-li, Taiwan, R.O.C., in 1992, and the M.S. degree in power mechanical engineering from the National Tsing-Hua University, Hsinchu, Taiwan, in 1994. He is currently pursuing the Ph.D. degree in the Department of Electrical and Control Engineering, National Chiao-Tung University, Hsinchu. His current research interests include speech recognition, neural networks, and the applications of short-range wireless communication like Bluetooth system.

Chin-Teng Lin (SM'88–M'91–SM'99) received the B.S. degree in control engineering from the National Chiao-Tung University, Hsinchu, Taiwan, R.O.C., in 1986, and the M.S.E.E. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, in 1989 and 1992, respectively.

Since August 1992, he has been with the College of Electrical Engineering and Computer Science, National Chiao-Tung University, where he is currently a Professor and Chairman of Electrical and Control Engineering Department. He served as the Deputy Dean of the Research and Development Office of the National Chiao-Tung University from 1998 to 2000. His current research interests are fuzzy systems, neural networks, intelligent control, human-machine interface, image processing, pattern recognition, video and audio (speech) processing, and intelligent transportation system (ITS). He is the coauthor of *Neural Fuzzy Systems—A Neuro-Fuzzy Synergism to Intelligent Systems* (Englewood Cliffs, NJ: Prentice-Hall, 1996), and the author of *Neural Fuzzy Control Systems with Structure and Parameter Learning* (Singapore: World Scientific, 1994). He has published about 60 journal papers in the areas of soft computing, neural networks, and fuzzy systems, including 35 paper for the IEEE.

Dr. Lin is a member of Tau Beta Pi and Eta Kappa Nu. He is also a member of the IEEE Computer Society, the IEEE Robotics and Automation Society, and the IEEE Systems, Man, and Cybernetics Society. He has been the Executive Council Member of Chinese Fuzzy System Association (CFSA) since 1995, and the Supervisor of Chinese Automation Association since 1998. He is the Chairman of the IEEE Robotics and Automation Taipei Chapter since 2000. He won the Outstanding Research Award granted by National Science Council (NSC), Taiwan, from 1997 to 2000, the Outstanding Electrical Engineering Professor Award granted by the Chinese Institute of Electrical Engineering (CIEE) in 1997, and the Outstanding Engineering Professor Award granted by the Chinese Institute of Engineering (CIE) in 2000. He was also elected to be one of the 38th Ten Outstanding Young Persons in Taiwan (2000).