Robust and accurate fundamental frequency estimation based on dominant harmonic components

Tomohiro Nakatani\textsuperscript{a)}
NTT Communication Science Laboratories, NTT Corporation, Soraku-gun, Kyoto 619-0237, Japan
Toshio Irino\textsuperscript{b)}
Faculty of Systems Engineering, Wakayama University/NTT Communication Science Laboratories, NTT Corporation

(Received 26 August 2002; accepted for publication 6 July 2004)

This paper presents a new method for robust and accurate fundamental frequency ($F_0$) estimation in the presence of background noise and spectral distortion. Degree of dominance and dominance spectrum are defined based on instantaneous frequencies. The degree of dominance allows one to evaluate the magnitude of individual harmonic components of the speech signals relative to background noise while reducing the influence of spectral distortion. The fundamental frequency is more accurately estimated from reliable harmonic components which are easy to select given the dominance spectra. Experiments are performed using white and babble background noise with and without spectral distortion as produced by a SRAEN filter. The results show that the present method is better than previously reported methods in terms of both gross and fine $F_0$ errors. © 2004 Acoustical Society of America. [DOI: 10.1121/1.1787522]

PACS numbers: 43.72.Ar, 43.66.Hg [DOS]

Pages: 3690–3700

I. INTRODUCTION

Achieving robust and accurate fundamental frequency ($F_0$) estimation is one of the most important problems in speech signal processing. $F_0$ is also cited as a major clue in relation to a person’s ability to extract a desired sound from other sounds in the real world.\textsuperscript{1} A number of useful speech applications have already been presented based on $F_0$ estimation. For example, we improved a speech recognition system by employing an $F_0$-based sound segregation system as a preprocessor.\textsuperscript{2} Recently, a very high quality vocoder, STRAIGHT, was developed based on the idea of $F_0$ adaptive processing.\textsuperscript{3} In any application, $F_0$ accuracy is of the greatest importance since any error in $F_0$ determination has a detrimental effect on system performance. Many applications also require robust $F_0$ estimation under adverse noise conditions.

A number of $F_0$ estimation methods have been proposed.\textsuperscript{4,5} Typical methods are categorized into two types, namely temporal and spectral methods. The former is mainly based on the autocorrelation of input signals and includes the maximum likelihood (ML) method. The latter is mainly based on spectral peak extraction and includes the cepstrum method. Recently, several new $F_0$ estimation methods have been proposed with the goal of providing better performance than the traditional methods.\textsuperscript{3,6–10} Some methods are more accurate, while others are more robust in the face of background noise. For example, YIN is a temporal method that is one of the most accurate $F_0$ estimators when there is no background noise.\textsuperscript{9} Shimamura’s method is a temporal one that is described as being robust against white Gaussian noise.\textsuperscript{7} TEMPO is a newly developed spectral method that enables reliable $F_0$ estimation in support of a new high quality vocoder, STRAIGHT.\textsuperscript{3} Liu and Lin presented a spectral method that is robust against white Gaussian noise.\textsuperscript{8} $F_0$ estimation methods for telephone quality speech have also been proposed recently.\textsuperscript{6,10}

Although these methods are superior to conventional methods in certain respects, the limited evaluation experiments undertaken thus far make it impossible to decide which method is the best. For example, robustness and accuracy have not been comprehensively evaluated for real world applications in the presence of both background noise and spectral distortion. The conditions under which speech sounds are recorded are not constant and are usually adversely affected by the acoustic properties of microphones and the environment. It is also necessary for the $F_0$ estimation methods to support telephone line use.

Furthermore, most methods have not been evaluated with large databases that contain reliable reference $F_0$ values. There are problems with regard to the definition of the “correct” $F_0$ value. It must be defined without using speech signals for a fair evaluation since speech sounds are time-varying filtered or “distorted” versions of glottal pulse sequences. Therefore, we believe this makes it desirable to use databases consisting of simultaneous recordings of speech and electro-glottal graph (EGG) signals because EGG signals are extracted directly from glottal vibrations.

In this paper, we present a new $F_0$ estimation method, referred to as DominAnce Spectrum based Harmonics extraction (DASH), and show its effectiveness with regard to both background noise and spectral distortion using a comprehensive evaluation method. We define a dominance spectrum to reduce the factors degrading $F_0$ estimation accuracy. We evaluate the robustness and accuracy of DASH using several large databases containing simultaneous recordings of speech and electro-glottal graph (EGG) signals.\textsuperscript{11–13} We use white noise and babble noise as background noise. We

\footnotesize\textsuperscript{a)}Electronic mail: nak@cslab.kecl.ntt.co.jp
\textsuperscript{b)}Electronic mail: irino@sys.wakayama-u.ac.jp
also apply spectral distortion to the speech signals using a SRAEN filter. This is a high pass filter, above 300 Hz, recommended by the ITU-T for simulating a telephone handset. As a result, the effectiveness of our $F_0$ estimation method is reliably examined under various kinds of adverse conditions.

In the rest of this paper, we describe the proposed $F_0$ estimation method based on dominance in Sec. II and experimental results in Sec. III.

II. METHOD

The $F_0$ estimation method proposed in this paper is an extension of a spectral method based on instantaneous frequency (IF). It is referred to as DominAnce Spectrum based Harmonics extraction (DASH). The processing flow of DASH is summarized in Fig. 1. By way of preparation, the input signal is first down-sampled (to 4 kHz), and converted into a signal in the frequency domain by short-time Fourier transformation (using 512 points with a 50-m Hanning window and a 1-m frame shift). The IF and the dominance spectrum are then extracted as described in Secs. II A and II B, respectively. The effectiveness of the dominance spectrum for both background noise and spectral distortion is explained in Sec. II C. Section II D describes a method for gathering harmonics information, namely harmonic dominance. Several techniques for increasing the robustness are described in Sec. II E. For example, Sec. II E I introduces an adaptive mechanism to decide the integration length for calculating the dominance spectrum. Finally, $F_0$ estimation is refined based on the degree of dominance of fixed points to provide very accurate results as described in Sec. II F.

A. $F_0$ estimation based on instantaneous frequencies

If a signal is represented as $s(t) = a(t) e^{j\phi(t)}$, the IF of the signal is defined as $\dot{\phi} = d\phi/dt$. The concept of IF was introduced by Flanagan as one of the speech features for a phase vocoder. The use of IF in $F_0$ estimation was initiated by Charpentier. Recently, it was reported that IF use is effective in $F_0$ estimation under noisy conditions. In these methods, the IF is calculated for each frequency bin of the short-time Fourier transform (STFT) at each time frame. The STFT can be viewed as a set of bandpass filters as follows:

$$S(f, t) = \int_{-\infty}^{\infty} g(\tau-t)x(\tau)e^{-j2\pi ft}d\tau,$$

$$= e^{-j2\pi ft}\int_{-\infty}^{\infty} g(\tau-t)e^{-j2\pi f(t-\tau)}x(\tau)d\tau,$$

$$F(f, t) = e^{j2\pi ft}S(f, t),$$

where $x(\tau)$ is the time series of the input signal, $f$ is the center frequency of each STFT bin, $g(\tau)$ is a window function, and $t$ is the center time of each time frame. Equation (2) implies that $S(f, t)$ can be interpreted as the output signal of a bandpass filter, $g(\tau)e^{-j2\pi f(t-\tau)}$, multiplied by $e^{-j2\pi ft}$. Therefore, signal $F(f, t)$ or $e^{j2\pi ft}S(f, t)$ is the output signal of the bandpass filter. The IF, $\dot{\phi}(f, t)$, is defined as the time derivative of the phase of $F(f, t)$ in as

$$\dot{\phi}(f, t) = \frac{\partial}{\partial t} \arg(F(f, t)).$$

The IF can be calculated according to Flanagan’s method as follows:

$$\dot{\phi}(f, t) = f + \frac{a(f, t)(\partial a(f, t)/\partial t)b(f, t) - b(f, t)(\partial a(f, t)/\partial t)a(f, t)}{a(f, t)^2 + b(f, t)^2},$$

where

$$a(f, t) = \text{Re}(S(f, t)).$$

FIG. 1. Flow diagram of $F_0$ estimation.
where $h(t)$ is the time derivative of $g(t)$.

When we define a mapping function between the center frequency of the STFT bin, $f$, and the derived IF, $\phi(t)$, for a voiced speech signal, it has the appearance of a regularly spaced staircase as shown in Fig. 2. The harmonic components of the fundamental frequency produces the individual stairs. This is because each harmonic component is the only dominant signal included in the frequency bands corresponding to its neighboring frequency bins, and the IFs for the bins coincide with the frequency of the component. Thus the best estimate of the harmonic frequency is at each fixed point where the IF coincides with the center frequencies of the bins, that is, $\phi = f$. The fundamental frequency, $F_0$, is estimated as the difference between the frequencies of adjacent fixed points since the harmonic frequencies are integer multiples of $F_0$. The use of IF is superior to conventional methods, as cepstrum methods, because it can provide a precise frequency estimate of the dominant harmonic component included in each STFT bin.

Abe et al. proposed the “IF amplitude spectrum,” which is a modified version of the amplitude spectrum, $|S(f,t)|$. The magnitude of the amplitude spectrum is reorganized according to IF, $\phi(t)$, instead of the STFT center frequency, $f$. The IF amplitude spectrum, $G(\phi,t)$, is defined as

$$G(\phi,t) = \lim_{\Delta \phi \to 0} \frac{1}{\Delta \phi} \int_{\phi-\Delta \phi}^{\phi+\Delta \phi} |S(f,t)| df.$$  

Since the IF amplitude spectrum enhances the harmonic structure of speech sounds, it was considered applicable to $F_0$ estimation in the presence of background noise. However, we suspect that the performance would deteriorate in the presence of spectral distortion caused by, for example, telephone handsets. This is because the IF amplitude spectrum is also affected by spectral distortion as it is inherited from the amplitude spectrum. Moreover, the IF amplitude spectrum would not reduce additive noise components since it is defined using a derivative operation on the frequency axis, which usually enhances the noise.

Atake et al. recently proposed an $F_0$ estimation method by defining the stability of fixed points based on Cohen’s bandwidth equation. This method makes it possible to estimate $F_0$ very accurately even under low signal-to-noise ratio (SNR) conditions. However, it requires an initial rough estimate of $F_0$ to define an adaptive short-time window for precise $F_0$ estimation. The degree of robustness may largely depend on this initial estimation. Therefore, one of our goals is to develop a method using a fixed window that can provide a robust estimation of $F_0$.

In order to overcome the problems posed by existing $F_0$ estimation methods, we propose a new IF-based method that is robust and accurate in the presence of both background noise and spectral distortion.

### B. Definition of dominance

We introduce a new measure based on IF to evaluate the magnitude of a harmonic component relative to the other components in each STFT bin. We refer to this measure as “the degree of dominance” and define it as $D(f,t)$ in Eq. (10) for each frequency bin centered at $f$ at a time frame centered at $t$:

$$D(f,t) = 10 \log_{10}(1/B(f,t)^2), \quad \text{Eq. (10)}$$

$$B(f,t)^2 = \frac{\int_{f-\Delta f/2}^{f+\Delta f/2} (\phi(f',t) - f)^2 S(f',t)^2 df'}{\int_{f-\Delta f/2}^{f+\Delta f/2} S(f',t)^2 df'} \quad \text{Eq. (11)}$$

where $\phi(f',t)$ and $S(f',t)$ represent the IF and the power spectrum, respectively, for a center frequency $f'$ in a frequency range neighborhood of the IF amplitude spectrum, $S(f',t)^2$.

The value of $B(f,t)^2$ reaches a minimum when the dominant frequency component of a signal coincides with the center frequency, $f$, because the value of $\phi(f',t)$ for the adjacent frequency bins approaches $f$. This is seen as the flat stairs around the fixed points in Fig. 2. Here, the degree of dominance, $D(f,t)$, reaches its maximum since it is defined as the logarithm of the inverse of $B(f,t)^2$ in Eq. (10). By contrast, the dominance value becomes smaller and does not have a sharp peak when the frequency component is greatly affected by noise because $\phi(f',t)$ increases as $f'$ increases and thus the difference between $\phi(f',t)$ and $f$ becomes large. Consequently, the degree of dominance has sharp peaks only at fixed points corresponding to dominant frequency components.

### C. Effectiveness of dominance spectrum

When the degree of dominance, $D(f,t)$, is calculated for all STFT bins, it takes the form of a spectrum [Fig. 3(a)]. This is referred to as a “dominance spectrum.” The following sections show that basing $F_0$ estimation on the dominance spectrum makes it robust with regard to both additive background noise and spectral distortion.
1. Robustness against additive noise

Figures 3(a)–(c) show, respectively, examples of a dominance spectrum, a logarithmic power spectrum, and the power spectrum of a clean speech signal when there is no background noise. The dominance spectrum [Fig. 3(a)] includes sharp peaks that correspond to the harmonic components. The peaks are much sharper than those of a usual logarithmic power spectrum [Fig. 3(b)]. Figures 3(d)–(f) show the same kinds of spectra when the speech signal is smeared with background white noise with 0 dB SNR. Although neither the dominance spectrum [Fig. 3(d)] nor the logarithmic power spectrum [Fig. 3(e)] have sharp peaks corresponding to the harmonic components above 500 Hz, the peak-to-trough ratio is smaller in the dominance spectrum. By contrast, the peak-to-trough ratio below 500 Hz is greater in the dominance spectrum. That is, the dominance spectrum enhances the peaks of the harmonic components and suppresses the variation produced by noise. This property is particularly useful for robust $F_0$ estimation. (Further detailed analysis is provided in Appendix A to explain how the dominance spectrum changes with the level of the background white noise.)

The difference between the power spectra of clean speech [Fig. 3(c)] and noisy speech [Fig. 3(f)] is very small. In other words, the proportion of peaks derived from harmonic components and noise components is very large in the power spectrum. Therefore, the power spectrum is also considered to be a good robust measure against additive background noise.

2. Robustness against spectral distortion

The dominance spectrum exhibits the property of whitening the spectral envelope and this reduces the effect of spectral distortion. This can be clearly demonstrated by passing a speech signal through a SRAEN filter to simulate the telephone handset case. Figure 3(g) shows the dominance spectrum for the filtered sound. The first and second peaks of the dominance spectrum in Fig. 3(g) are slightly smaller than those in the original spectrum in Fig. 3(a). But this small decrement does not affect $F_0$ estimation. Figure 3(i) shows the power spectrum of the filtered sound. The peaks below 300 Hz are greatly suppressed by the filtering, and this shows that the power spectrum is much more sensitive to spectral distortion than the dominance spectrum. By contrast, in the logarithmic power spectrum [Fig. 3(b)], the peaks at low frequencies clearly remain, albeit with a small decrement. This demonstrates that the logarithmic power spectrum is robust with regard to spectral distortion.

Consequently, only the dominance spectrum is robust against both background noise and spectral distortion, and thus must be useful for $F_0$ estimation under various adverse conditions.

D. Harmonic dominance

We now define a decision measure that summarizes the degree of dominance for all harmonic components at a time frame. This measure is referred to as “harmonic dominance” and is defined as $H(f_n,t)$ in Eq. (12):

$$H(f_n,t) = \sum_{l=1}^{l_{f_n}} (D(r(lf_n),t) - D(f,t)),$$

$$D(f,t) = \sum_{f} D(f,t)/N,$$

$$F_0 = \arg \max_{f_n} \{H(f_n,t)\},$$

where $f_n$ is one of the quantized $F_0$ candidates located within a specified $F_0$ search range, $n(=0,1,2,...)$ is its index, $lf_n$ corresponds to the frequency of its $n$th harmonic component, and $r(lf_n)$ is a function that transfers $lf_n$ to the center frequency of the nearest STFT bin. Harmonic components up to $F_{\text{max}} (=1500 \text{ Hz})$ are summed as in Eq. (12). $D(f,t)$ is the average of $D(f,t)$ over the number of STFT bins, $N$, where $f$ refers to the center frequency of an STFT bin. $D(f,t)$ is a term that ensures that the measure is unbiased. Our preliminary experiment showed that specific errors known as “double pitch” and “half pitch” errors were effectively re-
duced by this unbiasing (see Appendix B). The frequency that maximizes the harmonic dominance, $f_n$, becomes the estimate of the fundamental frequency $F_0$ as in Eq. (14).

E. Robust $F_0$ estimation based on dominance

We introduce two additional techniques to improve the robustness of our $F_0$ estimation method based on dominance. One is a method for determining the optimum integration length for calculating the dominance spectrum, and the other is the use of dynamic programming (DP) to reduce discontinuous transition errors in the $F_0$ trajectory.

1. Frequency range of bandwidth integral

The optimum frequency range, $\Delta f$, for the integral in Eq. (11) roughly depends on the $F_0$ of the input signal. $F_0$ can be estimated more robustly if the optimum range is given based on a rough estimate of the $F_0$. In our preliminary experiment, the optimum frequency range, $\Delta f$, was about 130 Hz for male speech and 260 Hz for female speech, which approximately coincides with the average $F_0$ of the target speech. In general, however, no a priori information on the target speech is available in advance, so the optimum range cannot be obtained in advance.

To overcome this problem, we introduce an adaptive method to decide the frequency range. With this method, the harmonic dominance is maximized twice, first for rough $F_0$ estimation, and second for more precise $F_0$ estimation. In the first estimation, we use a fixed frequency range for the integration that covers both male and female speech, and then we use the roughly estimated $F_0$ to determine the optimum frequency range for the precise estimation. In our experiment, the 260-Hz range (the same as the optimum value for female speech) was found to be suitable for the initial rough estimate, and about 67% to 110% of the initial rough $F_0$ estimate was suitable for the precise estimation.

2. The use of dynamic programming

Dynamic programming (DP) has often been used to obtain a smooth $F_0$ trajectory by tracking continuous peaks in the time series of a vector such as cepstral coefficients. In our $F_0$ estimation, the harmonic dominance calculated for each frequency bin at each time frame is treated as the vector to which DP is applied, and the frequencies corresponding to the positions of the tracked elements yield the smoothed $F_0$ trajectory.

$F_0$ tracking using DP is viewed as a processing approach that finds an optimum $F_0$ trajectory at the smallest cost overall. With our $F_0$ estimation, the time series of harmonic dominance are adopted as one cost function for DP. Let $H(f_n, t_k)$ be the harmonic dominance at the $k$th time frame for the $n$th quantized $F_0$ candidate, $f_n$, searched for in Eq. (12), and $P(t_k)$ be the signal power at the time frame. Then $d(n, t_k) = -P(t_k)H(f_n, t_k)$ is treated as the cost function in our implementation. Another cost function, $p(f_n, f_m)$, is also adopted to determine the cost of the $F_0$ transition from $f_n$ to $f_m$ between two adjacent time frames. $p(f_n, f_m) = |log(f_m/f_n)|$ is used in our implementation. This approach means that the $F_0$ tracking problem can be formalized as finding an $F_0$ sequence, $\Omega = \{f_{n_0}, f_{n_1}, \ldots, f_{n_T}\}$, that minimizes the following total cost function, $C(\Omega)$,

$$C(\Omega) = d(n_0, t_0) + \sum_{k=1}^{T} (w \cdot p(f_{n_{k-1}}, f_{n_k}) + d(n_k, t_k)),$$

(15)

where $w$ is a weight to determine the degree to which the $F_0$ transition cost is weighted in the total cost. Obviously, this cost function can efficiently be minimized using DP.

F. Accurate $F_0$ estimation based on dominance

To improve the accuracy of the $F_0$ estimation, we introduce an $F_0$ refinement method based on the IFs at fixed points. Because fixed points with large dominance values are expected to be derived from dominant harmonic components, the IFs at such fixed points are considered to be good estimates of their harmonic frequencies. Therefore, reliable $F_0$ candidates can be obtained by dividing their harmonic frequencies by their harmonic numbers. With our method, $F_0$ is determined as the weighted average of the $F_0$ candidates derived from fixed points using the degree of dominance as the weight. Because of this weight, $F_0$ is determined mainly based on the fixed points of dominant harmonic components, so the obtained $F_0$ is expected to be reliable.

The idea behind $F_0$ refinement based on fixed points was introduced by Atake et al. Cohen’s bandwidth equation was then used to evaluate the reliability of fixed points. In this paper, it is modified to use the degree of dominance.

If $F_0$ is the result of the maximization of the harmonic dominance as discussed in Sec. II E, the refined $F_0$ is defined as follows:

$$F_0 = \frac{\sum_{i=1}^{n} \sum_{\phi \in \Phi(i \cdot F_0')} \phi(i) (D(r(\phi), t) - c)}{\sum_{i=1}^{n} \sum_{\phi \in \Phi(i \cdot F_0')} (D(r(\phi), t) - c)},$$

(16)

$$c = \min_{\phi \in \Phi(i \cdot F_0'), i = 1 \sim n} \{D(r(\phi), t)\} - \epsilon \quad (\epsilon > 0).$$

(17)

Here, $\Phi(i \cdot F_0')$ is a set of IFs of fixed points that are located within $\pm 10%$ of the $i$th multiple of $F_0'$, and $n$ is the number of harmonics. Each $\phi$ is a candidate frequency of the $i$th harmonic component derived from a fixed point, so $\phi/i$ is a candidate for $F_0$, and $F_0$ is calculated as the average $\phi/i$ value weighted by the degree of dominance. “$r(\cdot)$” is used to transform a continuous frequency to its nearest STFT center frequency as used in Eq. (12). This quantization is only used to calculate the reliability of the fixed points, so it does not affect the accuracy of $F_0$ estimation. Term $c$ enables the weights for all $i$ to be greater than zero.

1. Interpolation for precise IF at a fixed point

IF $\phi$ at a fixed point is precisely determined by linear interpolation using the IFs and center frequencies of adjacent STFT bins. Let $f_1$ and $f_2$ be the center frequencies and $\phi_1$ and $\phi_2$ be their IFs. There is a fixed point between these two STFT bins if Eq. (19) holds. The interpolated value of the IF, $\phi$, at this fixed point is defined as follows:
\[ \phi = \frac{(f_2 - \phi_2)f_1 + (\phi_1 - f_1)f_2}{(\phi_1 - f_1) + (f_2 - \phi_2)}. \]  

(18)

where

\[ \phi_1 > f_1 \quad \text{and} \quad \phi_2 < f_2. \]

(19)

### G. The use of the power spectrum

In this section, we propose a simple alternative \( F_0 \) estimation method, referred to as the Ripple-Enhanced Power Spectrum (REPS), that is based on the power spectrum instead of the dominance spectrum. In the power spectrum shown in Fig. 3(f), the dominant harmonic components are represented as sharp peaks while the additive noise components are almost negligible. Thus the power spectrum also seems to be applicable to robust \( F_0 \) estimation in the presence of additive background noise although its performance would be sensitive to spectral distortions as described in Sec. II.C.2 using Fig. 3(i).

Although the power spectrum has been used in some \( F_0 \) estimation methods,\(^2\) the robustness of these methods has not yet been well evaluated. We, therefore, designed another method by substituting the power spectrum for the dominance spectrum.

In a preliminary study, however, we found that the direct use of the power spectrum, \( S(f,t)^2 \), does not necessarily result in the best performance. Instead, we modified the spectrum so as to enhance the spectral ripple corresponding to the glottal pulse. The ripple-enhanced power spectrum, \( R(f,t) \), is derived by a method similar to cepstral lifting in the power spectrum (not in the log domain as with usual cepstral lifting), that is, applying inverse discrete Fourier transformation (IDFT) to \( S(f,t)^2 \), substituting zeros for the lower quefrency components, and applying DFT to the modified coefficients. Finally, we developed our alternative \( F_0 \) estimation method by substituting \( R(f,t) \) for \( D(f,t) \) in Eqs. (12), (13), (16), and (17) without changing the rest of the processing method.

### III. EXPERIMENTS

#### A. Evaluation method

When evaluating an \( F_0 \) estimation method, it is of the greatest importance to define the correct \( F_0 \) values of the target speech. \( F_0 \) values labeled in a speech database have been used in many cases. Unfortunately, such labels are usually attached by hand and thus are not necessarily accurate or large enough for reliable evaluation. For robust evaluation, the correct \( F_0 \) values are often determined from clean speech signals using the same algorithm as that being evaluated by assuming that the method is sufficiently good for clean speech. However, we need to be careful about this assumption since nonharmonic components included in certain consonants and the time-varying frequency property of the vocal tract may affect \( F_0 \) estimation even in clean speech.

In our evaluation, the correct \( F_0 \), \( F_0^{\text{cor}} \), is estimated from electro-glottal graph (EGG) signals collected at the same time as the speech signals (see Sec. III.B). Since an EGG signal is directly derived from glottal vibration and is largely unaffected by the nonharmonic components of speech, it is an ideal signal for estimating the correct \( F_0 \) value.

Since we concentrated on the evaluation of robustness, we calculated two \( F_0 \) values, \( F_0^{\text{cor}} \) and \( F_0^{\text{est}} \), from the EGG signal and the noisy speech signal, respectively, using the same \( F_0 \) estimation algorithm. We then calculated the difference between them as an error measure for model performance. We defined two measures: gross \( F_0 \) error and fine \( F_0 \) error. Gross \( F_0 \) error is the ratio of the number of frames giving “incorrect” \( F_0^{\text{est}} \) values to the total number of frames. Value \( F_0^{\text{cor}} \) is “incorrect” if it falls outside \( \pm 5\% \) of the \( F_0^{\text{cor}} \) value. Fine \( F_0 \) error is the normalized root mean square error between the \( F_0^{\text{cor}} \) value and \( F_0^{\text{est}} \) value which is not judged as “incorrect” in the gross error measurement. The formulation is the root mean square of \( (F_0^{\text{cor}} - F_0^{\text{est}})/F_0^{\text{cor}} \).

Through our evaluation, we search for \( F_0 \)'s between 50 to 500 Hz for every 1-ms frame shift, and we evaluate the estimated \( F_0 \)'s only for voiced durations based on voicing labels.
B. Databases

We used the following three databases of speech and EGG signals recorded simultaneously.

DB1 30 utterances by 14 male and 14 female Japanese speakers (total of 840 utterances, total duration of 40 min, 16-kHz sampling, and 16-bit quantization).\(^{13}\)

DB2 50 utterances by one male and one female English speaker (total of 100 utterances, total duration of 7 min, 20-kHz sampling, and 16-bit quantization).\(^{11}\) The database can be downloaded from the following URL, http://ftp.cs.keele.ac.uk/pub/pitch/Speecheval.tar.gz

DB3 Phonetically balanced text utterances by five male and five female English speakers (total duration of 9 min, 20-kHz sampling, and 16-bit quantization).\(^{12}\) The database can be downloaded from the following URL, ftp://ftp.cs.keele.ac.uk/pub/pitch/Speech

We used the voiced/unvoiced labels included in the three databases, and determined correct \(F_0\) ’s, \(F_0^{\text{cor}}\), for each \(F_0\) estimation method from the EGG signals as described in Sec. III A. We did not use the reference \(F_0\) labels present in DB2 and DB3 except as indicated in Appendix C because the reference \(F_0\) labels may have a certain bias depending on the method used for their estimation.

C. Noise and spectral distortion

We used white noise and babble noise as the additive background noise. The SNR in terms of the average power ranged from \(\infty\) to 0 dB. We calculated the average power of a signal as follows: (1) the initial average power of a signal was first obtained by calculating the average power for all frames, (2) frames with less than \(\frac{1}{3}\) of the initial average power were discarded, and then (3) the average power was calculated for the remaining frames.

The babble noise was a mixture of the normalized speech signals of ten speakers randomly selected from DB1. In a preliminary listening test involving the speech sounds in the babble noise, the target sounds were almost completely indistinguishable from the babble noise when the SNR was 0 dB and the babble noise, the target sounds were almost all distinguishable when the SNR was \(-10\) dB.

Spectral distortion was introduced by using a SRAEN filter. The SRAEN filter recommended by ITU-T is designed to simulate the ideal frequency property of a telephone handset.\(^{14}\) It has defined spectral characteristics and is a high-pass filter above 300 Hz. Accordingly, this filter almost completely eliminates the fundamental component of male speech. We designed an FIR filter to fit the spectral characteristics by using the Remez algorithm.

D. \(F_0\) estimation methods for comparison

We evaluated the dominance-spectrum-based method with an adaptive integration range (DASH), and the ripple-enhanced power-spectral-based method (REPS). We compared them with a commonly-used cepstrum method (cepstrum), the \(F_0\) estimation method used in STRAIGHT, namely TEMPO,\(^{3}\) and YIN.\(^{9}\) TEMPO is an \(F_0\) estimation method based on “fundamentalness.” YIN is a modified version of the autocorrelation method. These were recently proposed as improvements to conventional methods.

As a further comparison, we provide several evaluation results in Appendix C using reference \(F_0\) labels and \(\pm 20\%\) gross error criterion, both of which are often used as evaluation conditions for recently proposed \(F_0\) estimation methods.\(^{6,9,10}\)

E. Results

1. Robustness of \(F_0\) estimation

Figure 4 shows the gross \(F_0\) errors for DB1 obtained by DASH, cepstrum, TEMPO, and YIN\(^{21}\) in the presence of additive white noise (left panel) and babble noise (right panel) for an SNR range of \(\infty\) to 0 dB without any spectral distortion by a SRAEN filter. The precise values for SNRs of \(\infty\) and 5 dB are listed in the second, third, and fourth columns of Table I. In both figure and table, DASH produces smaller errors than cepstrum, TEMPO, and YIN for all SNRs. In Tables II and III, the gross errors obtained for DB2 and DB3 are listed in the same order as in Table I.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Without SRAEN</th>
<th>With SRAEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR (dB)</td>
<td>W</td>
<td>B</td>
</tr>
<tr>
<td>DASH</td>
<td>1.52</td>
<td>2.31</td>
</tr>
<tr>
<td>REPS(_{1})</td>
<td>0.89</td>
<td>1.26</td>
</tr>
<tr>
<td>cepstrum</td>
<td>6.69</td>
<td>22.4</td>
</tr>
<tr>
<td>TEMPO</td>
<td>3.53</td>
<td>9.02</td>
</tr>
<tr>
<td>YIN</td>
<td>4.30</td>
<td>6.06</td>
</tr>
</tbody>
</table>

TABLE II. Gross \(F_0\) errors (%) for DB2 with a \(\pm 5\%\) error criterion for SNRs of \(\infty\) and 5 dB, with and without a SRAEN filter. The bold font shows the best performance. W: white noise; B: babble noise.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Without SRAEN</th>
<th>With SRAEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR (dB)</td>
<td>W</td>
<td>B</td>
</tr>
<tr>
<td>DASH</td>
<td>3.93</td>
<td>5.70</td>
</tr>
<tr>
<td>REPS(_{1})</td>
<td>2.59</td>
<td>3.76</td>
</tr>
<tr>
<td>cepstrum</td>
<td>11.0</td>
<td>29.7</td>
</tr>
<tr>
<td>TEMPO</td>
<td>8.07</td>
<td>18.6</td>
</tr>
<tr>
<td>YIN</td>
<td>5.04</td>
<td>10.7</td>
</tr>
</tbody>
</table>

TABLE III. Gross \(F_0\) errors (%) for DB3 with a \(\pm 5\%\) error criterion for SNRs of \(\infty\) and 5 dB, with and without a SRAEN filter. The bold font shows the best performance. W: white noise; B: babble noise.
proposed method, DASH, again gives smaller errors than any other method except for REPS 3 which is also our proposed method (see Sec. III E 2). Since the tendencies of the errors are consistent across the three different databases, the superiority of DASH is clearly demonstrated.

2. The use of the power spectrum

In Sec. II G, we proposed the use of the ripple-enhanced power spectrum as a substitute for the dominance spectrum. We refer to this original method as REPS 1. We evaluated this method under the same conditions as those used in Sec. III E 1. The thick solid lines (REPS 1) in Fig. 5 show the gross $F_0$ errors for DB1 with white noise (left panel) and babble noise (right panel). The errors were always more than 10% and much worse than those obtained using the dominance spectrum as shown in Fig. 4. We found that the error was very large when estimating $F_0$ from the EGG signal. This is particularly important since we assumed the estimated values to be the correct $F_0$ values in our evaluation. A detailed analysis showed that the spectral power at frequencies below $F_0$ often exceeded the power of the $F_0$ components in the EGG signal and this precluded exact $F_0$ estimation. An effective way of eliminating these characteristics in the EGG signal was to use a preemphasis filter (3 dB/oct).

We therefore applied a preemphasis filter to both the speech and EGG signals. We refer to this modified method as REPS 2. The results are shown by the thin solid lines (REPS 2) in Fig. 5. The errors were improved at high SNRs but not so much at low SNRs, especially in the presence of babble noise. We also found that the robustness of the $F_0$ estimation for the speech signals was degraded by the preemphasis. We therefore applied the preemphasis filter to the EGG signals but not to the speech signals. This is referred to as REPS 3. The thick dashed lines (REPS 3) in Fig. 5 show the results. The gross errors became much better than those obtained with REPS 1 and REPS 2. Moreover, the results were even better than those obtained with DASH shown as the solid line in Fig. 4. This is also clearly shown by the gross errors for $\infty$ and 5 dB in Tables I–III. Thus REPS 3 yielded the best performance if the signal was not smeared by any spectral distortion.

These results show that the ripple-enhanced power spectrum is robust against additive noise provided the spectral distortion is properly compensated. It is, however, difficult to specify an optimum preprocessing filter in advance when sound is recorded in the real world since the impulse response is not necessarily well-defined.

FIG. 5. Gross $F_0$ errors obtained by REPS 1, REPS 2, and REPS 3 in the presence of white noise (left panel) and babble noise (right panel). See Sec. III E 2 for definitions of REPS 1, REPS 2, and REPS 3.

FIG. 6. Gross $F_0$ errors obtained by DASH, REPS 3, cepstrum, TEMPO, and YIN for speech signals spectrally distorted by SRAEN filtering in the presence of white noise (left panel) and babble noise (right panel).
3. Robustness against spectral distortion

To evaluate robustness against spectral distortion, we applied a SRAEN filter to speech signals after adding noise. Figure 6 shows the gross \( F_0 \) errors for DB1 obtained with various \( F_0 \) estimation methods in the presence of white noise and babble noise. The gross \( F_0 \) errors for SNRs of \( \infty \) and 5 dB for DB1 with two types of background noise and a SRAEN filter are also listed in the fifth to seventh columns of Table I. Tables II and III show the gross errors for DB2 and DB3 in the same manner. In these results, the two proposed methods, DASH and REPS\(_3\), were always better than any other method. In addition, the errors obtained with DASH were not greatly affected by SRAEN filtering, and were almost always better than the errors obtained with any other method except with DB1 under low SNR conditions. This again implies that the method based on a ripple-enhanced power spectrum is not as robust against spectral distortion as the dominance-spectrum-based method, although it is most robust against background noise. This tendency becomes more obvious when we refer to the results obtained without DP as shown in Sec. III E5.

From the results presented in Secs. III E1, III E2, and III E3, we can conclude that DASH is recommended for real world applications when the spectral distortion is unknown, unpredictable, or time varying, while REPS is useful when it is possible to compensate the spectral distortion in advance.

4. Accuracy of \( F_0 \) estimation

We evaluated the fine \( F_0 \) errors of all \( F_0 \) estimation methods for DB1 using the two types of background noise and the SRAEN filter. Figure 7 and Table IV show the results. The fine errors were calculated over time frames where all the methods estimated \( F_0 \) correctly, i.e., the error was within \( \pm 5\% \), in Secs. III E1, III E2, and III E3. The errors with DASH were almost the same as those with REPS\(_3\) and were better than the errors with cepstrum, TEMPO, and YIN. This was because DASH and REPS\(_3\) both use \( F_0 \) refinement procedures based on fixed points as described in Sec. II F. This shows that the use of fixed points is effective for increasing the accuracy of \( F_0 \) estimation.

5. Effect of dynamic programming

Finally, we measured the effect of DP. We compared the gross \( F_0 \) errors obtained by the two proposed methods with and without DP using DB1. The results are listed in Table V. Although the gross errors of the methods without DP were worse than with DP, the performance of DASH and that of REPS\(_3\) are still generally superior to those of the other methods shown in Table I, except when using REPS\(_3\) in the presence of spectral distortion provided by a SRAEN filter. This implies that the \( F_0 \) estimation methods based on the dominance spectrum and the ripple-enhanced power spectrum are

### TABLE IV. Fine \( F_0 \) errors (%) for DB1 for SNRs of \( \infty \) and 5 dB, with and without a SRAEN filter. The bold font shows the best performance. W: white noise, B: babble noise.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Without SRAEN</th>
<th>With SRAEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR (dB)</td>
<td>W</td>
<td>B</td>
</tr>
<tr>
<td>DASH</td>
<td>0.68</td>
<td>0.80</td>
</tr>
<tr>
<td>REPS(_3)</td>
<td>0.66</td>
<td>0.69</td>
</tr>
<tr>
<td>cepstrum</td>
<td>0.86</td>
<td>1.43</td>
</tr>
<tr>
<td>TEMPO</td>
<td>0.85</td>
<td>1.41</td>
</tr>
<tr>
<td>YIN</td>
<td>1.05</td>
<td>1.08</td>
</tr>
</tbody>
</table>

### TABLE V. Gross \( F_0 \) errors (%) for DB1 obtained by the proposed methods with a \( \pm 5\% \) error criterion with/without DP (WDP/WODP) for SNRs of \( \infty \) and 5 dB, with and without a SRAEN filter. The bold font shows the best performance. W: white noise, B: babble noise.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Without SRAEN</th>
<th>With SRAEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR (dB)</td>
<td>W</td>
<td>B</td>
</tr>
<tr>
<td>DASH</td>
<td>1.52</td>
<td>2.31</td>
</tr>
<tr>
<td>REPS(_3)</td>
<td>2.80</td>
<td>6.34</td>
</tr>
<tr>
<td>cepstrum</td>
<td>0.89</td>
<td>1.26</td>
</tr>
<tr>
<td>TEMPO</td>
<td>1.49</td>
<td>2.32</td>
</tr>
</tbody>
</table>

FIG. 7. Fine \( F_0 \) errors obtained by DASH, REPS\(_3\), cepstrum, TEMPO, and YIN for speech signals without spectral distortion (left panel) and with spectral distortion caused by SRAEN filtering (right panel) in the presence of white noise.
both very robust against background noise, and that the former is also very robust against spectral distortion, even without any $F_0$ trajectory correction.

**IV. CONCLUSION**

We proposed a robust and accurate method for $F_0$ estimation, referred to as DASH, that is based on a new measure, namely the degree of dominance. We defined the degree of dominance to allow us to evaluate the magnitude of the harmonic components of speech signals relative to background noise. DASH was found to be more robust and accurate than conventional methods under background noise conditions of white noise and babble noise. We also demonstrated that DASH was robust against the spectral distortion created by SRAEN filtering. Moreover, we found that a simple method based on a ripple-enhanced power spectrum, referred to as REPS, was a good alternative provided the spectral shape of the input sound was carefully compensated. REPS will be useful when it is possible to compensate the spectral distortion in advance while DASH is recommended for real world applications when the spectral distortion is unknown, unpredictable, or time-varying. Further improvement is required for such applications as sound segregation in the presence of babble noise or under cocktail party conditions. It would be possible to use dual or repetitive $F_0$ estimation and a method using sound source localization information when a multi-microphone setup is available.

**ACKNOWLEDGMENTS**

The authors express their gratitude to Professor H. Kawahara of Wakayama Univ. for providing the database of speech with EGG, Dr. A. de Cheveigné of IRCAM for updating his $F_0$ estimation program, YIN, Dr. S. Katagiri for research support, and members of NTT for helpful discussions. This work was partially supported by CREST of JST.

**APPENDIX A: DOMINANCE SPECTRUM DEPENDENCE ON THE BACKGROUND NOISE LEVEL**

Figure 8 shows the dominance spectrum and logarithmic power spectrum of a harmonic sound with background white noise. The harmonic sound is composed of two sinusoids of 200 and 400 Hz. White noise is added to the harmonic sound so that the SNR becomes 30, 10, 0, and $-20$ dB.

The two sinusoidal components have sharper peaks in the dominance spectrum than in the logarithmic power spectrum even when the SNR is low. The peaks and dips produced by the noise affect the preciseness. It is desirable for the spectral representation to yield distinctive peaks corresponding to harmonics and a flat response in the noise region. This shows that the dominance spectrum is a better representation for $F_0$ estimation than the logarithmic power spectrum.

**APPENDIX B: METHOD FOR AVOIDING DOUBLE/HALF PITCH ERRORS**

The dominance spectrum is always negative when there is no subtraction of $D(f, t)$ in Eq. (12). Then, the harmonic dominance of the larger $f_n$ in Eq. (12) tends to become large, resulting in an increase in the double pitch error. This is because harmonic numbers below $F_{\text{max}}$ become smaller as $f_n$ becomes larger, and thus the number of negative values decreases in the harmonic dominance calculation. Conversely, when the dominance spectrum is set positive by adding a large constant value, which often produces negative values increases in the harmonic dominance calculation. Therefore, it is necessary to normalize the spectrum adequately to avoid these two kinds of errors. For this purpose, we choose the average of the dominance value for all frequencies as the

![Fig. 8. Dominance spectrum (left panels) and logarithmic power spectrum (right panels) for a mixture of two sinusoidal components and white noise when the SNRs are 30, 10, 0, and $-20$ dB.](image)
These tables also contain the results obtained by YIN in our with previous reports.

The formulation of the degree of dominance is similar in appearance to 11 P. C. Bagshaw, S. M. Hiller, and M. A. Jack, “Enhanced pitch tracking and the processing of f0 contours for computer aiding intonation teach-

APPENDIX C: GROSS F0 ERRORS CALCULATED USING COMMONLY USED F0 LABELS

Although it is very difficult to define universally correct F0 values, common F0 labels used for several previous F0 evaluation reports 6,9,10 may provide better references for a rough comparison of F0 estimation methods. For this purpose, we evaluated our F0 estimation methods with the F0 labels used by de Cheveigne.6 For DB1 and DB2, we used the labels made by de Cheveigne. For DB3, we used the labels present in the DB itself. All the labels were estimated from EGG signals. In addition, we modified the definition of the gross F0 errors as the ratios of the estimated F0 values that fell ±20% beyond the correct F0 value, in accordance with previous reports.

Table VI shows the results for DB1, DB2, and DB3. These tables also contain the results obtained by YIN in our experiments, and those reported by Wang et al. 6 and by Kasi et al.,10 referred to as Wang and Kasi. Note that Wang and Kasi are evaluated under somewhat different conditions. Therefore, direct comparison remains difficult. For example, they did not evaluate their methods under adverse noise conditions. Furthermore, the gross error rates of Kasi were calculated only on frames for which both reference F0 labels and their voicing decision method indicate voiced frames. The voicing errors of their voicing decision method were reported as being more than 10%, therefore we suspect that their gross error rates would be much worse when they are calculated for all the voicing durations of the reference F0 labels.

The results in the three tables clearly demonstrate that our proposed methods, DASH and REPS3, are superior to the other methods.


19 The integration in Eq. (11) is the numerical integration of the values at discrete points. The frequency quantization is 1 Hz when calculating Eq. (12). Each harmonic frequency, f0*, in Eq. (12) is transformed to the nearest center frequency by an operator “(*)”. Also, in Eq. (16) for F0 refinement, the operator “(*)” is again used to transform each IF into a discrete center frequency.

20 The formulation of the degree of dominance is similar in appearance to Cohen’s bandwidth equation, i.e., \( B^2 = \int (f - \bar{f})^2 S(f)^2 df \int S(f)^2 df \), where \( \bar{f} \) is an average frequency. However, Cohen’s definition does not represent the bandwidth of the instantaneous frequency around a fixed point, but the width of the distribution in the power spectrum around the center of the distribution. The former is considered to be more suitable for evaluating the stability of fixed points. Also, Cohen’s definition does not provide a way of localizing integration areas around specified frequencies. Therefore, we use the degree of dominance to improve the accuracy of F0 estimation.

21 Under clean speech conditions, the gross pitch error achieved by YIN was 4.3% as shown in Table I, and the result was worse than 0.3% originally reported by de Cheveigné. This is because our evaluation conditions were more severe than theirs. In their work, the F0 gross errors were calculated according to whether the obtained F0 was within ±20% of the correct F0. Moreover, they used their own F0 labels for the same database. In our preliminary study, we were able to improve the performance of YIN to 0.44% gross errors with their F0 labels and a ±20% error criterion to simulate their conditions. Note that YIN software was downloaded from de Cheveigné’s www page.