Robust Pitch Estimation with Harmonics Enhancement in Noisy Environments Based on Instantaneous Frequency

Toshihiko Abe, Takao Kobayashi and Satoshi Imai

Precision and Intelligence Laboratory Tokyo Institute of Technology, Yokohama, Japan
email:tabe@pi.titech.ac.jp

ABSTRACT

In this paper, we propose an approach for estimating pitch of speech in noisy environments based on instantaneous frequency(IF). First, we define the IF amplitude spectrum, which is obtained by projecting the STFT amplitude spectrum onto the IF axis. Based on the IF amplitude spectrum, we can perform harmonics enhancement by suppressing the aperiodic components. Next, we define an evaluation function to find pitch. This is done by expanding the IF amplitude spectrum to the time region. Then we propose a method for obtaining a continuous pitch contour using the dynamic programming. Experiments show accuracy and robustness of our method especially when noise exists.

1. INTRODUCTION

The difficulty of accurate and robust pitch estimation of speech is caused by several facts such as rapidly changing instantaneous pitch and formants. Moreover, the existence of noise makes it more difficult because the periodicity of speech will be ambiguous. Thus robust pitch estimation is still considered one of the most difficult tasks in speech processing.

In this paper, we present a pitch estimator based on instantaneous frequency(IF). The instantaneous frequency(IF) is a good descriptor for a signal changing its frequency in time. There have been some studies of frequency analysis of speech based on IF [1]. We have also proposed a study of harmonics estimation and pitch extraction based on IF[2].

We define the IF as a function of time and frequency. From this expression we propose a new spectral representation, the IF amplitude spectrum, which much more clearly represents the harmonic structure of speech than the short time Fourier transform(STFT) amplitude spectrum. Based on the IF amplitude spectrum, we can enhance harmonics of speech by suppressing the aperiodic components. As a result, enhanced speech makes pitch estimation more accurate and robust especially for noisy speech. For pitch estimation, we propose a local pitch evaluation function at each point of time based on IF. Then we make a global evaluation function which is an integral of the local evaluation function along each possible pitch contour on which a condition of continuity is imposed. Next, we will find a continuous pitch contour which maximizes the global evaluation function. This enables us to avoid most of gross pitch errors such as double-pitch or half-pitch. We can find such a pitch contour using the dynamic programming(DP). There have been some pitch estimators using the DP[3][4], to minimize a kind of cost function. On the other hand, our method does not need any cost functions because it simply maximizes the integral of the local evaluation function.

2. OBTAINING THE IF FROM STFT

The short time Fourier transform (STFT) of a signal $x(t)$ is defined by

$$X(\omega,t)=\int_{-\infty}^{\infty}x(\tau)w(\tau-t)\,\mathrm{e}^{-j\omega\tau}d\tau, \tag{1}$$

where $w(t)$ is a window function. Then we can obtain a filter-bank expression $F(\omega,t)$ from $X(\omega,t)$ by

$$F(\omega,t)=\mathrm{e}^{j\omega t}X(\omega,t), \tag{2}$$

We consider $F(\omega,t)$ as the coefficients when $x(t)$ is expressed by a linear combination of basis functions[5]

$$f(\omega,t)=w(t)\,\mathrm{e}^{j\omega t}. \tag{3}$$

Then $x(t)$ is reconstructed by $F(\omega,t)$ as

$$x(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}F(\omega,\tau)f(\omega,\tau-t)d\omega d\tau, \tag{4}$$

provided that $\int_{-\infty}^{\infty}\{w(t)\}^2dt=1$.

Now we define the IF at the point (ω,t) by

$$\lambda(\omega,t)=\frac{\partial}{\partial t}\arg[F(\omega,t)]. \tag{5}$$

If we put

$$F(\omega,t)=a+jb,$$

the IF is given[6] by

$$\lambda=\frac{a\frac{\partial b}{\partial t}-b\frac{\partial a}{\partial t}}{a^2+b^2}. \tag{6}$$

Moreover,

$$\begin{aligned}\frac{\partial}{\partial t}F(\omega,t)&=\frac{\partial a}{\partial t}+j\frac{\partial b}{\partial t}\\&=\int_{-\infty}^{\infty}\left(-\frac{\partial w(\tau-t)}{\partial t}+j\omega w(\tau-t)\right)\mathrm{e}^{-j\omega(\tau-t)}x(\tau)d\tau\end{aligned} \tag{7}$$

Therefore $\partial a/\partial t$ and $\partial b/\partial t$ in (6) can be obtained by replacing the window function in (1) with $\left(-\frac{\partial}{\partial t}w(\tau-t)+j\omega w(\tau-t)\right)$.

3. THE IF AMPLITUDE SPECTRUM

3.1. Definition

Here we consider $F(\omega,t)$ and $\lambda(\omega,t)$ at a fixed time point t and use expressions $F(\omega)$ and $\lambda(\omega)$ for convinience.

Since (2) leads to $|F(\omega)| = |X(\omega)|$, $|F(\omega)|$ is equivalent to the STFT amplitude spectrum. Now, we take an integral of $|F(\omega)|$ on a subinterval on frequency ω where $\lambda_0 \leq \lambda(\omega) \leq \lambda_0 + \Delta\lambda$, and we regard the integral as the amplitude on the subinterval. Then we take a limit $\Delta\lambda \rightarrow 0$ and define the limit value of the amplitude as the IF amplitude spectrum,

$$g(\lambda_0) = \lim_{\Delta\lambda \rightarrow 0} \int_{\lambda_0 \leq \lambda(\omega) \leq \lambda_0 + \Delta\lambda} |F(\omega)| d\omega. \qquad (8)$$

Fig.1 shows some examples of the IF amplitude spectrum. Here we use a 40msec Blackman window as the window function $w(t)$. Since we cannot actually take the limit $\Delta\lambda \rightarrow 0$, we approximate it by taking a sufficiently small step $\Delta\lambda$. Fig.1(a) and (b) is quasi-periodic signals. Sharp peaks are observed where harmonics are located. On the other hand, Fig.1(c) is an aperiodic signal. Although sharp peaks are also can be seen, they are not so sharp and stable in time in comparison with those of the periodic signals. Examples of the STFT and the IF spectrograms are shown in [IMAGE A676G01.GIF] and [IMAGE A676G02.GIF] for a word utterance, [IMAGE A676G03.GIF] and [IMAGE A676G04.GIF] for white noise.

3.2. Emphasis on Harmonics

We define the normalized local second-order moment of the IF amplitude spectrum at frequency λ_0 by

$$m_\lambda(\lambda_0) = \frac{\int_{-\infty}^{\infty} \exp\left\{-\frac{(\lambda-\lambda_0)^2}{\sigma^2}\right\} |g(\lambda)|(\lambda-\lambda_0)^2 d\lambda}{\int_{-\infty}^{\infty} \exp\left\{-\frac{(\lambda-\lambda_0)^2}{\sigma^2}\right\} |g(\lambda)| d\lambda}, \qquad (9)$$

where σ corresponds to the extension of the gaussian function which works for localization in frequency. By using (8) and putting $\lambda_0 = \lambda(\omega_0)$, we obtain

$$m(\lambda(\omega_0)) = \frac{\int_{-\infty}^{\infty} |F(\omega)| \exp\left\{-\frac{(\lambda(\omega)-\lambda(\omega_0))^2}{\sigma^2}\right\} (\lambda(\omega)-\lambda(\omega_0))^2 d\omega}{\int_{-\infty}^{\infty} |F(\omega)| \exp\left\{-\frac{(\lambda(\omega)-\lambda(\omega_0))^2}{\sigma^2}\right\} \lambda(\omega) d\omega}. \qquad (10)$$

Now we again consider (10) to be a function of ω and t by putting $m(\omega,t) = m_\lambda(\lambda(\omega,t))$. The function $m(\omega,t)$ indicates the IF's dispersion with respect to ω around the point (ω,t). As is shown in Fig.1, the IF amplitude spectrum of a periodic component such as sinusoids concentrates almost at a point in frequency, therefore the moment will be very small. On the other hand, the IF amplitude spectrum of a aperiodic component has a wider dispersion and the moment will be larger.

By utilizing this property of the moment, we can achieve harmonics enhancement by suppressing aperiodic components as follows. Here we weight $F(\omega,t)$ according to $m(\omega,t)$ as follows.

$$\tilde{F}(\omega,t) = \exp\left\{-\frac{(m(\omega,t))^2}{\rho^2}\right\} F(\omega,t) \qquad (11)$$

This makes $\tilde{F}(\omega,t)$ small where $m(\omega,t)$ is large and vice versa. And the constant ρ controls the weighting rate. We can obtain the enhanced signal $\tilde{x}(t)$ by inverse-transforming $\tilde{F}(\omega,t)$ according to (4). The enhanced signal $\tilde{x}(t)$ has its aperiodic component suppressed and as a result its harmonic components get enhanced. In order to preserve the harmonic components with as less distortion as possible, we should set σ in (9) to be small enough for the gaussian function to separate a single harmonic component from adjacent ones. Moreover, satisfying this condition, we should choose σ as large as possible to capture the wide dispersion of aperiodic components so that the aperiodic components will be suppressed. The enhanced signal $\tilde{x}(t)$ will be used for the following pitch estimation. An example of the harmonics enhancement is shown in [IMAGE A676G05.GIF].

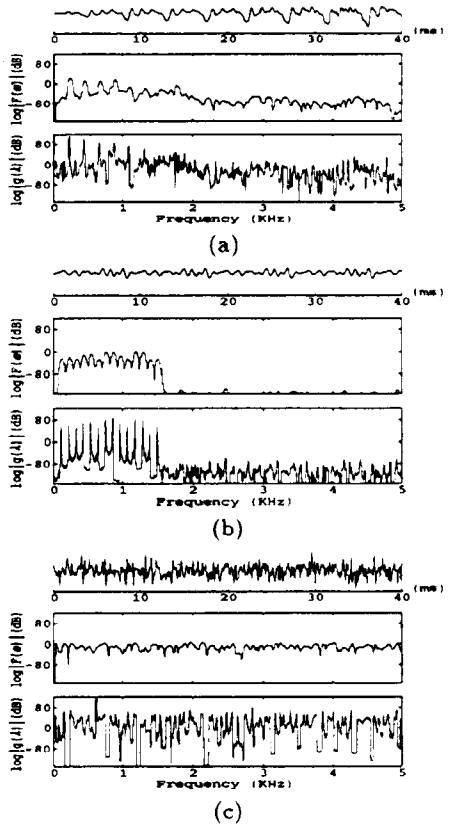

Figure 1: Waveforms, amplitude spectra and the IF amplitude spectra (a) beginning of a female utterance 'arayuru ⋯,' (b) composition of several sinusoids, (c) white noise.

4. PITCH ESTIMATION

4.1. Principle

In this section, we will show how to achieve pitch estimation based on the IF amplitude spectrum $g(\lambda)$. Here we define a transform of $g(\lambda)$

$$v(T,t) = 10^{-\alpha T} \int_{\lambda_0}^{\lambda_1} g(\lambda) s(\lambda,T)\, d\lambda \qquad (12)$$

where

$$s(\lambda, T) = \begin{cases} 0, & \lambda T < \pi \\ \frac{1}{2}(\cos \lambda T + 1), & \lambda T \geq \pi \end{cases} \quad (13)$$

The function $s(\lambda, T)$ has peaks where T is located at the fundamental period $2\pi/\lambda$ and its integral multiples. We consider $s(\lambda, T)$ to indicate the likelihood that a signal component of the IF λ has periodicity of T. And $10^{-\alpha T}$ works as weighting in such a way that it gives slight priority to shorter pitch periods. The values λ_0 and λ_1 determine the range of λ used for pitch estimation. Fig.2 shows an example of $v(T, t)$ of a speech signal at a point of t.

Since $v(T, t)$ is the integral of $s(\lambda, T)$, it indicates the sum of the likelihood of the pitch period resulted from all the harmonic components. Therefore we consider this to be a time-local evaluation function for pitch period T at time t. Eq.(12) can be achieved by the integral on ω as follows.

$$v(T, t) = 10^{-\alpha T} \int_{\omega_0}^{\omega_1} |F(\omega, t)| s(\lambda(\omega, t), T)\, d\omega \quad (14)$$

The values ω_0 and ω_1 determines the range of ω used for pitch estimation. Now we define a pitch contour on a interval $[t_0, t_1]$, which is supposed here to include the whole speech. We denote a pitch contour function which gives pitch period T_p at time t by $T_p(t)$. We assume this to be one of a set of functions $T(t)$ which satisfy a continuity condition we will mention later. Moreover, we assume that $T_p(t)$ maximizes the curvilinear integral along itself of the evaluation function $v(T, t)$. It is given by

$$T_p(t) = \underset{T(t)}{\operatorname{argmax}} \int_{t_0}^{t_1} v(T(t), t)\, dt. \quad (15)$$

We call this the global evaluation function on $[t_0, t_1]$.

4.2. Algorithm

To achieve the above pitch estimation on a discrete system, we rewrite (15) in a discrete form and use the dynamic programming (DP). Here we assume that the variables t and T take discrete values and use integral variables n and N instead. We assume the interval $[t_0, t_1]$ corresponds to the interval $[n_0, n_1]$. Now we are going to find the pitch contour $N_p(n)$ on $[n_0, n_1]$. The range of pitch period is assumed to be $N_a \leq N_p(n) \leq N_b$. From (15), the pitch contour is given by

$$N_p(n) = \underset{N(n)}{\operatorname{argmax}} \sum_{n=n_0}^{n_1} v(N(n), n), \quad (16)$$

where $N(n)$ is a set of all the possible pitch contours in the range $n_0 \leq n \leq n_1$. To obtain $N_p(n)$, we define an evaluation function for a subinterval $[n_0, n]$ as

$$V(N, n) = \max_{N(n)} \sum_{k=n_0}^{n} v(N(k), k), \quad (17)$$

And we set the initial value as

$$V(N, n_0) = v(N, n_0), \quad N_a \leq N \leq N_b \quad (18)$$

Then we use iteration

$$V(N, n) = v(N, n) + \max_{|i| \leq i_0} V(N + i, n - 1), \quad N_a \leq N \leq N_b, \quad n_0 < n \leq n_1. \quad (19)$$

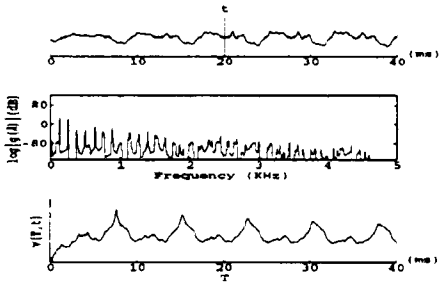

Figure 2: Waveform, the amplitude spectra and the local evaluation function $v(T, t)$ of the beginning of a male utterance 'arayuru ···,'

Finally we can find $N_p(n)$ by back-tracking the local maximum of the global evaluation function by

$$N_p(n_1) = \underset{N_a \leq N \leq N_b}{\operatorname{argmax}} V(N, n_1)$$

$$N_p(n) = N_p(n+1) + \underset{|i| \leq i_0}{\operatorname{argmax}} V(N_p(n+1) + i, n),$$

$$n = n_1 - 1, n_1 - 2, \cdots, n_0. \quad (20)$$

The constant i_0 in (19) and (20) is an integer which determines the continuity condition on the pitch contour. We set the grid sizes on n and N small enough, and set $i_0 = 1$ which means the pitch period $N(n)$ is able to increase or decrease by only 1 as n increases by 1. This leads to limitation of the pitch contour's slope depending on the ratio of the grid size on n to that on N. In general, we should make the ratio small enough to allow sharp pitch contours' slope. The grid size on n will be the frame interval and the grid size on N will be the sampling period in practice.

5. EXPERIMENTS

5.1. Experimental Conditions

We performed some experiments of pitch estimation on speech signals including noise-added signals. The sampling frequency is 10kHz. The speech data are Japanese sentences uttered by two male speakers and two female speakers. Four sentences about 15 seconds in total are used for each speaker. We made noisy signals by adding noise to the speech signals. White noise, car noise and computer room noise are used.

To obtain $F(\omega, t)$ in (1) from discrete-time signals, 512-point FFT was performed at intervals of 2msec. A 40-msec-Blackman window was used for the window function. For harmonics enhancement, we chose σ in (10) to be $2 \times (10\text{KHz})/512 = 39.1\text{Hz}$, and ρ in (11) to be $(13.8\text{Hz})^2$. For pitch estimation, we set ω_0 in (14) to be about 140Hz, ω_1 to be 5kHz and α to be $1/200\text{msec}^{-1}$. These values are based on some preliminary experiments. The SNR is calculated over the duration of each original speech signal from the database.

We compared performance of pitch estimators which are the proposed pitch estimator(proposed1), harmonics enhancement and the proposed pitch estimator(proposed2), the cepstrum pitch estimator(cepstrum) and the cepstrum pitch estimator to which the proposed DP is applied(cepstrum+DP). For the cepstrum method, the maximum value in each frame of the cepstrum in the range from 2.1msec to 14.9msec was detected as a pitch period. We confirmed that the cep-

strum+DP method made no gross errors like double-pitch or half-pitch, therefore we used those pitch contours as standard ones.

The error is defined as the deviation from the standard pitch. If an error is greater than 20 percent, it is classified as a gross pitch error(GPE). Otherwise, it is classified as a fine pitch error(FPE). The FPE is defined as the square-root of the mean square error. The evaluation points are taken at every 2 msec on the voiced part of the standard pitch.

5.2. Results

Fig.3 shows the results of pitch estimation. When SNR $= \infty$ which means noise free, it can be seen that the cepstrum+DP method causes no errors. This is because the pitch contours in this case are used as standard. We can see the cepstrum method causes a few percent GPE. Most of those errors are found to be double-pitch and half-pitch errors. In the case of low SNR, the cepstrum+DP method made less errors than the cesptrum method. This is due to the fact that the proposed DP impose a condition of continuity on pitch contours and that prevents errors like double-pitch or half-pitch. It is clearly seen that the proposed methods cause less GPE and FPE than the cepstrum methods for the most part. Moreover, we can see the GPE caused by proposed2 method are less than those caused by proposed1 method. This indicates the effectiveness of the harmonics enhancement for pitch estimation. Although we also applied the harmonics enhancement to the cepstrum methods, there was no particular improvement.

6. CONCLUSION

In this paper, we proposed a method to enhance harmonics of speech signals based on the local distribution of the IF amplitude spectrum. And we proposed a pitch estimator also based on the IF amplitude spectrum. The methods are shown to be efficient in pitch estimation. It is shown that the proposed pitch estimator has superiority with respect to GPE especially for noisy signals. This is consider to be due to the harmonics enhancement and the pitch estimator which guarantees the continuity of pitch contours. In addition, the DP method also works on the cepstrum pitch estimator for reducing GPE.

7. REFERENCES

1. M. Cooke and M. Crawford, "Tracking spectral dominances in an auditory model," in *Visual Representations of Speech Signals*,pp.197–204, John Wiley & Sons Ltd, 1993.
2. T. Abe, T. Kobayashi and S. Imai, "Harmonics estimation based on instantaneous frequency and Its Application to Pitch Determination," *IEICE Trans. Information & Systems*, vol. E78–D, No.9, pp. 1188–1194, 1995.
3. A. Kießling, R. Kompe, H. Niemann and E. Nöth, "DP-based determination of F0 contours from speech signals," *in Proc. ICASSP 92*, vol. II, pp. 17–20, 1992.
4. H. Ney, "Dynamic programming technique for nonlinear smoothing," *in Proc. ICASSP 81*, pp. 62–65, 1981.
5. P.P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, New Jersey, 1993.
6. J.L. Flanagan and R.M. Golden, "Phase vocoder," *Bell Syst. Tech.*, vol. 45, pp. 1493–1509, 1966.

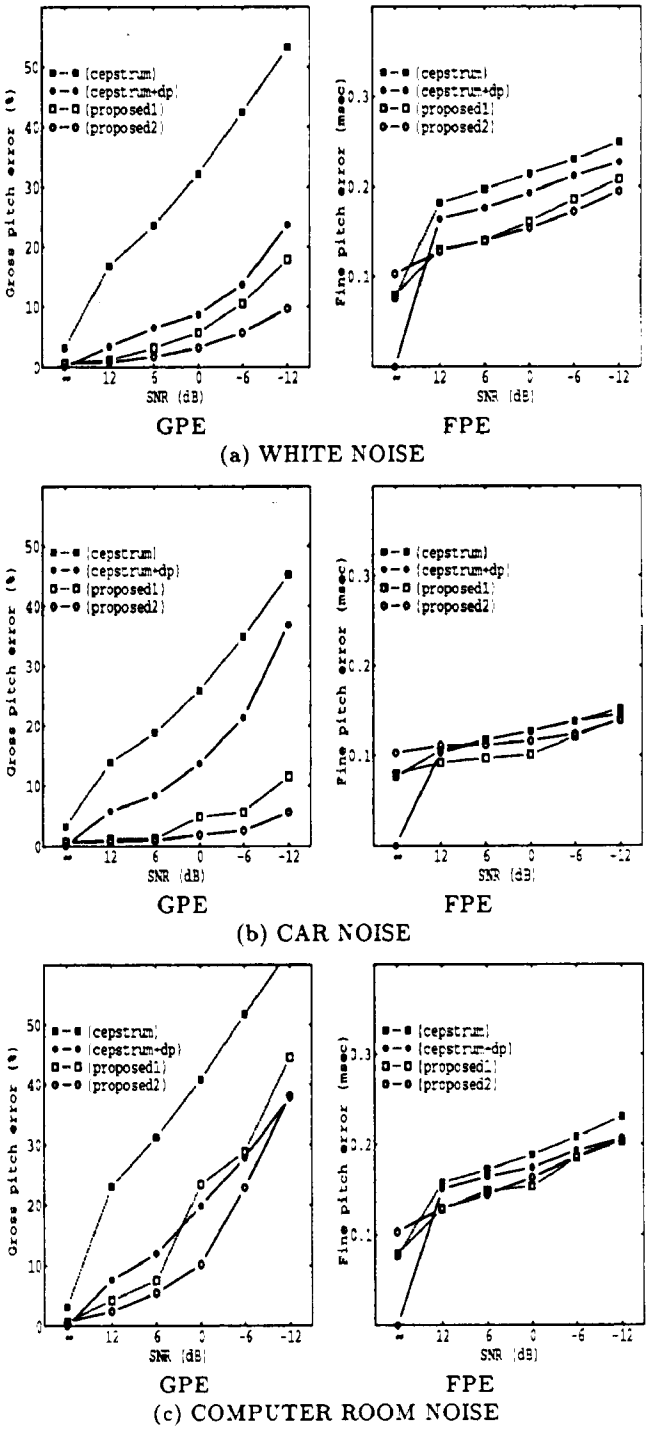

Figure 3: GPE and FPE on pitch estimation.