HARMONICS TRACKING AND PITCH EXTRACTION BASED ON INSTANTANEOUS FREQUENCY

Toshihiko Abe, Takao Kobayashi, Satoshi Imai

Precision and Intelligence Laboratory, Tokyo Institute of Technology
Nagatsuta, Midori-ku, Yokohama, 226 Japan

ABSTRACT

This paper proposes a technique for estimating the harmonic frequencies based on instantaneous frequency (IF) of a speech signal. The main problem is how to decompose the speech signal into the harmonic components. For this purpose, we use a set of bandpass-filters, each of whose center frequencies changes with time in such a way that it tracks the instantaneous frequency of its output. As a result, the outputs of the band-pass filters become the harmonic components, and the instantaneous frequencies of the harmonics are accurately estimated. To evaluate the effectiveness of the approach, we apply it to pitch extraction. The pitch extraction is simply accomplished by selecting the correct fundamental frequency out of the harmonic frequencies. The most significant feature of the pitch extractor is that the extracted pitch contour is smooth and it requires no post-processing such as nonlinear filtering or any smoothing processes.

1. INTRODUCTION

Harmonic components of a speech signal are fundamental elements of the voiced part. Each of the harmonic components can be represented by an amplitude and frequency modulated sinusoid. For example, a speech analysis/synthesis system based on sinusoidal representation has been developed by McAulay and Quatieri[1]. Their approach attempted to extract the instantaneous amplitudes and frequencies of the underlying sinusoids.

The methods for estimating the harmonics of the speech which have been proposed so far are not satisfying with respect to both accuracy and robustness. In this paper, we propose an approach for estimating the frequencies of the harmonic components based on instantaneous frequency (IF)[2][3]. The IF is defined as the phase derivative with respect to time. To estimate the harmonics of the signal, we use a set of band-pass filters. Each harmonic component is tracked by a bandpass filter. It changes its center frequency towards the IF of the output. As a result, the output of the bandpass filter will be the harmonic component. Thus the speech signal is decomposed into the harmonic components and the harmonic frequencies are obtained as their IF's.

For performance evaluation, we apply our approach to pitch extraction. It is accomplished by simply selecting the correct fundamental frequency out of the harmonic frequencies. The frequency of each harmonic components which are overlapping with each other in the time domain is compared with the others, then the component whose location in the frequency domain is the lowest is detected as the fundamental frequency component.

2. HARMONICS EXTRACTION

2.1. Sinusoidal Representation of a Speech Signal

We assume that a speech signal $x(t)$ is represented as the summation of the harmonic components:

$$x(t) = \sum_l x_l(t). \tag{1}$$

Each harmonic component $x_l(t)$ of the speech signal can be regarded as a sinusoid modulated in both amplitude and frequency as follows:

$$x_l(t) = \begin{cases} r_l(t) \cdot \cos\left(\int_{t_l^s} t\phi_l(\tau)d\tau + \nu_l \right), & t_l^s \le t \le t_l^e \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $r_l(t)$ and $\phi_l(t)$ are the instantaneous amplitude and frequency [2][3], respectively, and ν_l denotes the initial phase. In addition, $[t_l s, t_l e]$ is the interval in which the sinusoidal component exists. In (2), the whole argument of the cosine function is the phase $\theta_l(t)$. Thus, the sinusoidal component is written by

$$x_l(t) = r_l(t) \cdot \cos(\theta_l(t)). \tag{3}$$

The instantaneous frequency (IF) is defined as the phase derivative with respect to time:

$$\phi_l(t) = \frac{d\theta_l(t)}{dt}. \tag{4}$$

0-7803-2431-5/95 $4.00

When $x_l(t)$ is given, $r_l(t)$ and $\phi_l(t)$ which satisfy (3) are not unique. However, one of the solutions which satisfy (3) is obtained using the analytic signal of $x_l(t)$[3]. If we let $z_l(t)$ denote the analytic signal derived from the harmonic component $x_l(t)$, the instantaneous amplitude and frequency are obtained as follows [2][3]:

$$r_l(t) = |z_l(t)| \tag{5}$$

$$\phi_l(t) = \frac{d}{dt}\arg[z_l(t)]. \tag{6}$$

2.2. Band-pass Filters for Analysis

To get the representation of (1) from the speech signal $x(t)$, it is necessary to decompose $x(t)$ into the harmonic components $x_l(t)$. To do this, we use a set of band-pass filters. Moreover, each of the band-pass filters is controlled to be tracking a single harmonic component.

The band-pass filters are assumed to have relatively narrow pass-band and linear phase characteristics. The speech signal $x(t)$ is passed through a bank of contiguous band-pass filters. Let the impulse response of the nth band-pass filter be

$$g_n(t) = h(t) \cdot \exp(j\omega_n t), \quad \omega_n > 0 \tag{7}$$

where $h(t)$ is the impulse response of a prototype low-pass filter and ω_n is the center frequency. The output of the nth band-pass filter $y_n(t)$ is given by the convolution of $x(t)$ with $g_n(t)$. Since the band-pass filter with the impulse response of (7) has a pass-band only in the positive frequency region, $y_n(t)$ contains only positive frequency components. As a result, the output $y_n(t)$ becomes an analytic signal.

Therefore, from eqs.(5) and (6), the amplitude and the phase derivative of $y_n(t)$

$$r_n(t) = |y_n(t)| \tag{8}$$

$$\phi_n(t) = \frac{d}{dt}\arg[y_n(t)] \tag{9}$$

give estimates of the instantaneous amplitude and the IF of a harmonic component of $x(t)$.

2.3. Control of the Center Frequencies of the Band-pass Filters

To make each of the bandpass-filters track a single harmonic component, the center frequency of the filter need to be controlled. We form a follow-up control system whose target is the IF $\phi_n(t)$ by letting the center frequency of the filter ω_n be a function of time, $\omega_n(t)$. The system is described as

$$\frac{d}{dt}\omega_n(t) = a(\phi_n(t) - \omega_n(t)) \tag{10}$$

where a is a time constant. Then, the filter will be tracking one of the harmonic components.

Each harmonic component tends to attract the filters whose center frequencies are close to the IF of the harmonic component. Therefore some filters may track the identical harmonic component at the same time. In such a case, the system controls the filters to make the only one filter track a single harmonic component at the moment. If the difference of any two center frequencies of the filters is below a prescribed threshold ω_{min}, the filter whose output is smaller than the other in amplitude will be forced to move away toward the sparsest place in frequency among the filters. For this purpose, the difference between the center frequencies of any two adjacent filters is measured and the largest difference is searched. If the difference between ω_m and ω_n is found to be the largest, the center frequency of the filter which have to moved off is given as the average point of the frequencies ω_m and ω_n.

We should prescribe the frequency range to be analyzed in advance. The pass-bands of the filters should not go out of the analysis frequency range. If the pass-band of any filter is about to go out of the range at any moment, the system makes the filter keep the same center frequency as just before.

As a result, the intervals of the center frequencies of the filters do not become too small nor too large, and the placement of the filters in the analysis frequency range will be relatively uniform. Therefore, the whole band of the speech signal can be analyzed with a enough number of the filters.

2.4. Decomposition into Harmonic Components

The approach we described in the previous section makes the filters successfully track the harmonic components, but its information is only a set of the outputs of the band-pass filters. Now, we need to determine the duration of the harmonic component, $[t_l s, t_l e]$.

At first, we define the part in which the amplitude of the output continuously keeps itself above the prescribed threshold r_{min} as a single component part. Next, the component whose duration is shorter than a prescribed threshold d_{min}, is regarded as a noisy fragment which should be ignored. Such a component is discarded and the remainder is determined as the harmonic components of the speech signal.

2.5. Practical Implementation

Here we show an example of a harmonic estimation experiment. To realize the approach described above on the digital computer with sampled data, the phase derivative of (9) is approximated by difference[2]. The differential equation of (10) is also calculated by difference.

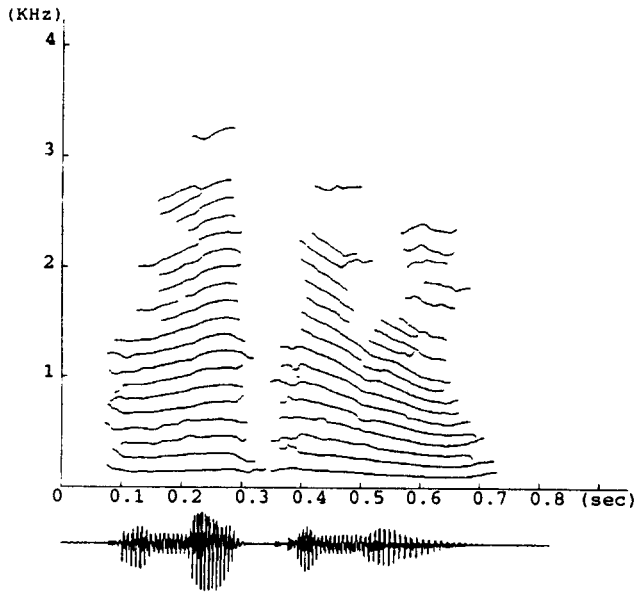

Figure 1: Harmonic frequencies of a Japanese word /tomakomai/ uttered by a male speaker.

Fig.1 shows an example of the harmonics extracted from a real speech. The sampling frequency of the speech data used for the experiment was 10kHz. The speech was quantized at 12-bit. The frequency range to be analyzed is from 0kHz to 5kHz. From the results of the preliminary experiment, we set the thresholds ω_{min}, and d_{min} to be 31.8Hz and 60msec, respectively. A 40msec-long Blackman window is used as the low-pass filter. Its 3.0-dB bandwidth is 42Hz. The number of the band-pass filters is 40. The center frequencies of the band-pass filters are uniformly placed at the beginning. In Fig.1, the waveform of the speech signal is shown in the lower part. A harmonic structure is clearly seen in the figure.

3. PITCH EXTRACTION

3.1. Algorithm

We have developed a pitch extractor based on the harmonics extraction system described in the previous section. If the harmonic components are accurately estimated, the fundamental frequency components should be included among them. The pitch extractor selects the components which have fundamental frequencies out of the other components. To achieve this, each frequency of the components whose locations are overlapping in the time domain, namely, existing at the same time, is compared with the others, and the components whose frequencies are not the lowest is excluded from the candidates for the fundamental frequency component. Consequently, the rest of the components and their frequencies are regarded as fundamental components and fundamental frequencies, respectively. The part where the fundamental components exist is classified as voiced.

3.2. Experiments

The proposed pitch extractor was tested on real and synthetic speech data. The synthetic speech data are generated using the mel-ceptstral analysis[6] and the MLSA filter[7] synthesis system with isolated words in the Tohoku-university Matsushita data base uttered by three male speakers and three female speakers. The synthetic speech is about 42 seconds long in total.

Fig.2 shows an example of the pitch contours extracted from the speech uttered a female speaker adopted from the ATR Japanese phrase database. It is seen from the figure that there are no leaps to the value like double-pitch or half-pitch. It has been found that such leaps never occur over all the speech data used for the experiment.

The performance of the proposed pitch extractor and a pitch extractor based on cepstrum analysis[5] are evaluated. The cepstrum pitch extractor is performed using a simple peak-picking algorithm within the fixed range of the pitch period from 2 to 15 msec. The pitch contours extracted from the pitch extractor are compared with the standard pitch contour which has been given to the synthesis system. The error is defined as the difference between their pitch periods. If the error is greater than 30 percent of the greater value of the two pitch periods, the error is classified as a gross pitch error. Otherwise, the error is classified as a fine pitch error. The error is evaluated at every 5msec. A standard deviation of the fine pitch error and a ratio of the number of gross pitch errors to the number of the evaluated points are measured. The standard deviation and gross pitch error rate measured on the synthetic speech data are shown in Table I and II, respectively. It is evidently shown that the proposed pitch extractor has better performance than the cepstrum pitch extractor. It should be noted that no gross pitch errors are caused by the proposed pitch extractor.

4. CONCLUDING REMARKS

We have proposed a technique for the harmonics estimation and developed a pitch extractor based on the technique. We have graphically observed the harmonic frequencies like Fig.1 on all of the speech data used for the experiment and found that the technique provides good performance with respect to stability and smoothness.

Table I: Fine pitch error on synthetic speech data(msec).

speaker	M1	M2	M3
proposed	0.170	0.204	0.164
cepstrum	0.284	0.300	0.320
speaker	F1	F2	F3
proposed	0.061	0.062	0.094
cepstrum	0.110	0.122	0.167

Table II: Percentage of the gross pitch error on synthetic speech data.

speaker	M1	M2	M3
proposed	0.000	0.000	0.000
cepstrum	0.845	0.868	1.257
speaker	F1	F2	F3
proposed	0.000	0.000	0.000
cepstrum	4.835	5.945	8.190

To evaluate the performance quantitatively, we have developed a simple pitch extractor. The performance of the pitch extractor shows the accuracy of the estimation of the fundamental frequency. It has been confirmed that the extracted pitch never happens to vary discontinuously, and the case that the pitch leaps to a value like double-pitch or half-pitch is never caused. This is because the transition of the extracted pitch value is essentially smooth. Therefore it does not require any smoothing processes. In addition, there is no need of fixed range of a pitch frequency. For example, the algorithm is also able to extract pitch of a pitch-modulated speech signal, moreover, even if the signal is not a human voice.

There is another approach to developing a more efficient pitch extractor. It has been found that the performance of the pitch extractor can be improved by determining pitch based on temporal transitions of the frequencies of the overall harmonic components. Even if no component of the fundamental frequency exists in the speech signal, it can detect pitch based on the other harmonic components. This makes the pitch extraction performance more robust. Then band limitation or noise addition with a speech signal will not make so serious difficulties in the pitch extraction scheme, utilizing all of the harmonic frequencies other than only a fundamental frequency component.

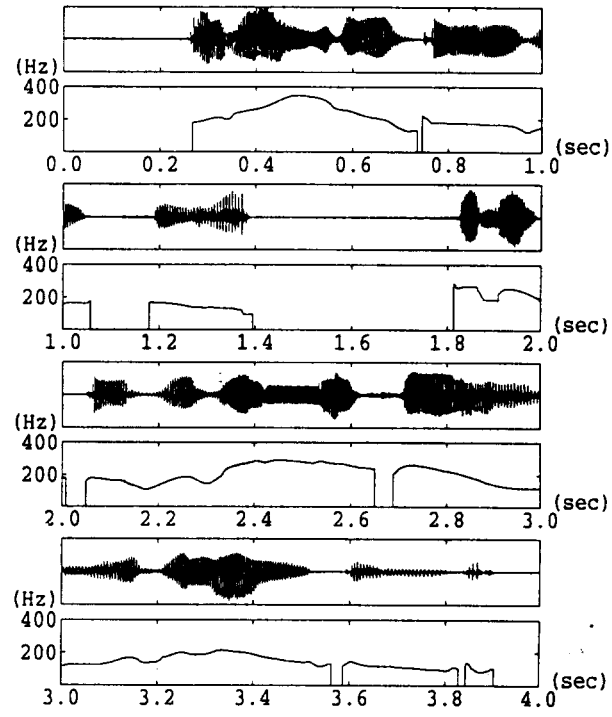

Figure 2: The fundamental frequency extracted from a Japanese sentence /arayuru genzitsuo subete zibuNno hooe nezimagetanoda/

REFERENCES

[1] McAulay, R.J. and Quatieri, T.F., "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, No.4, pp. 744-754, Aug. 1986.

[2] Flanagan, J.L. and Golden, R.M., "Phase vocoder," *Bell Syst. Tech.*, vol. 45, pp. 1493-1509, Nov. 1966.

[3] Boashash, B., "Estimating and interpreting the instantaneous frequency of a signal," *Proc. IEEE*, vol. 80, No. 4, pp. 519-568, Apr. 1992.

[4] Hess, W., *Pitch Determination of Speech Signals*, Springer-Verlag, Berlin, 1983.

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

[6] Tokuda, K., Kobayashi, T., Fukada, T., Saito, H. and Imai, S., "Spectral estimation of Speech Based on mel-cepstral representation," *Trans. IECE of Japan*, vol. J74-A, No. 8, pp. 1240-1248, Aug. 1991.

[7] Imai, S., Sumita, K. and Furuichi, C., "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Trans. IECE of Japan*, vol. J66-A, No. 2, pp. 122-129, Feb. 1983.