PAPER

# Fundamental Frequency Estimation for Noisy Speech Using Entropy-Weighted Periodic and Harmonic Features

**Yuichi ISHIMOTO**[†], ***Student Member***, **Kentaro ISHIZUKA**[††], **Kiyoaki AIKAWA**[†††], ***and*** **Masato AKAGI**[†], ***Members***

**SUMMARY** This paper proposes a robust method for estimating the fundamental frequency (F0) in real environments. It is assumed that the spectral structure of real environmental noise varies momentarily and its energy does not distribute evenly in the time-frequency domain. Therefore, segmenting a spectrogram of speech mixed with environmental noise into narrow time-frequency regions will produce low-noise regions in which the signal-to-noise ratio is high. The proposed method estimates F0 from the periodic and harmonic features that are clearly observed in the low-noise regions. It first uses two kinds of spectrogram, one with high frequency resolution and another with high temporal resolution, to represent the periodic and harmonic features corresponding to F0. Next, the method segments these two kinds of feature plane into narrow time-frequency regions, and calculates the probability function of F0 for each region. It then utilizes the entropy of the probability function as weight to emphasize the probability function in the low-noise region and to enhance noise robustness. Finally, the probability functions are grouped in each time, and F0 is obtained as the frequency with the highest probability of the function. The experimental results showed that, in comparison with other approaches such as the cepstrum method and the autocorrelation method, the developed method can more robustly estimate F0s from speech in the presence of band-limited noise and car noise.

***key words:*** *fundamental frequency estimation, entropy, instantaneous amplitude, periodic feature, harmonic feature*

## 1. Introduction

Extraction of the fundamental frequency (F0) of speech is an important problem as regards various areas of speech signal processing, such as speech recognition, speech analysis/synthesis, and speech segregation. For example, in speech recognition, prosodic features of speech can be used for prosodic phrase segmentation in order to improve the recognition accuracy [1]. In speech analysis/synthesis, F0 is the factor controlling the pitch of speech, so F0 extraction in the analysis part is necessary for synthesizing natural speech sounds [2]. In auditory scene analysis, such as speech segregation, it is considered that the human auditory system uses pitch, which is related to F0, as a cue to segregate concurrent speech signals [3]. There have been many studies in computational auditory scene analysis that use F0s of target speech in noisy environments [4], [5]. F0 extraction from speech in noisy environments is thus important for applications of speech signal processing in real environments. However, it is difficult because noises distort the harmonic components of speech.

Various methods of F0 estimation have been developed [6]–[8], and most of them make use of the periodic features of speech in the time domain or harmonic features in the frequency domain. To extract F0s from the periodic features of speech in the time domain, the autocorrelation function of the speech waveform has been used. The autocorrelation method is robust against white noise, but it is not robust against colored noise. To extract F0s from the harmonic features in the frequency domain, the cepstrum method and comb filtering of the amplitude spectrum of speech have been used. The cepstrum method can extract accurate F0s from clean speech, but the accuracy is easily deteriorated by noise. Comb filtering of the amplitude spectrum is more robust than the cepstrum method, but it cannot estimate F0s of noiseless speech with similar accuracy to that of the method that uses instantaneous frequency (explained below).

The instantaneous frequency of speech has recently been used for accurately estimating F0s [9]–[12]. The instantaneous frequency is one of the features that can accurately represent the periodic feature of speech signals; however, the accuracy of such representation is easily deteriorated by noise. For example, Kawahara et al. proposed a method of F0 extraction (STRAIGHT-TEMPO, based on the stability of instantaneous frequencies) to construct a speech analysis/synthesis system [11]. This method can accurately extract F0s of clean speech, but it is not very effective in noisy environments, especially when the signal-to-noise ratio (SNR) is below 10 dB. The instantaneous amplitude of speech obtained by time-frequency analysis can also represent harmonic components of speech [13], and it is more robust against noise than the instantaneous frequency. Unoki and Akagi constructed a sound-segregation model using F0 estimation based on the comb filtering of the instantaneous amplitude [5]. This method can estimate F0s of vowels in noisy environments; however, its accuracy deteriorates for continuous speech mixed with noise.

Manuscript received February 10, 2003.
Manuscript revised June 9, 2003.

[†]The authors are with Japan Advanced Institute of Science and Technology, Ishikawa-ken, 923–1292 Japan.
[††]The author is with NTT Communication Science Laboratories, NTT Corporation, Atsugi-shi, 243–0198 Japan.
[†††]The author is with Tokyo University of Technology, Hachioji-shi, 192–0982 Japan.

We previously proposed a robust and accurate method based on instantaneous amplitude and frequency for estimating the F0 of noisy speech [14]. This method combines two F0 estimation ways: one based on the periodic and harmonic features of instantaneous amplitude (PHIA) for robust estimation in noisy environments, and the other based on the stability of the instantaneous frequency in relation to accurate estimation. PHIA can more robustly estimate F0s from continuous speech under white-noise environment than the methods using only periodic or harmonic features. However, the performance of this approach is not very good under real environmental noise. The reason for this is as follows. The energy produced by environmental noise does not distribute evenly in the time-frequency domain; for example, most energy produced by car noise exists in low-frequency region and varies momentarily. PHIA uses mainly harmonic components of speech in the low-frequency region. Environmental noise, which, like car noise, has high energy in low-frequency region, thus affects the accuracy of PHIA. We consider, therefore, that F0 estimation by using instantaneous amplitude in real environments should not use a particular frequency region but must use actively low-noise regions taken from the whole frequency. It is thus important to understand how the F0 estimation method utilizes the low-noise time-frequency regions.

In light of the above-mentioned background, we have developed a robust F0 estimation method that uses periodic and harmonic features weighted by entropy in real environments. This method is considered to be used as the first estimation part of [14]. Although the first part of [14] requires the robustness and the accuracy against real environmental noise, the robustness is more important than the accuracy because the accuracy is given by another part of [14]. The key feature of the proposed method is to make a point of estimating F0s from the low-noise regions in the time-frequency domain. The method segments the time-frequency plane into regions that consist of a short time section and a frequency band with several harmonics. It then obtains the probability functions of the F0 produced from the periodic and harmonic features in the regions. Although the correct F0 can be extracted from each probability function in the segmented plane in the case of clean speech, the probability function will be distorted by noise in the case of speech mixed with environmental noise. Accordingly, to enhance the noise robustness, the method utilizes the low-noise regions. If noise exists in a region, entropy of the probability function becomes high in that region. The method thus emphasizes low-noise regions by integrating regional features by using entropy-based weight, and more robustly estimate F0 from speech in the presence of real environmental noise.

## 2. Algorithm

### 2.1 Overview

Figure 1 shows an overview of the proposed method. In the first step, this method analyzes the observed signal by using two filterbanks and represents the instantaneous amplitude of the signal in the time-frequency domain. If the signal has a harmonic complex tone like speech, this time-frequency representation of the instantaneous amplitude of the signal will have periodic and harmonic features corresponding to the F0s of the signal. F0 can be extracted from any local region of the two time-frequency planes. Therefore, even if some local regions are contaminated by noise, F0 can be extracted from the local regions that are not contaminated by noise. The main idea of this paper is to improve reliability of F0 estimation by using entropy to weight the two features in the cleaner local regions as mentioned below.

In the next step, multiple F0 candidates are extracted from the periodic and harmonic features of the instantaneous amplitude by autocorrelation in the time and frequency domains. The autocorrelation is processed in each time-frequency region and, accordingly, the F0 candidates are mapped in the time-frequency plane. There are some low-noise regions of noisy speech in real environments because the energy produced by real environmental noise does not distribute evenly in the time-frequency domain. The method segments the time-frequency plane, which is allocated the F0 candi-

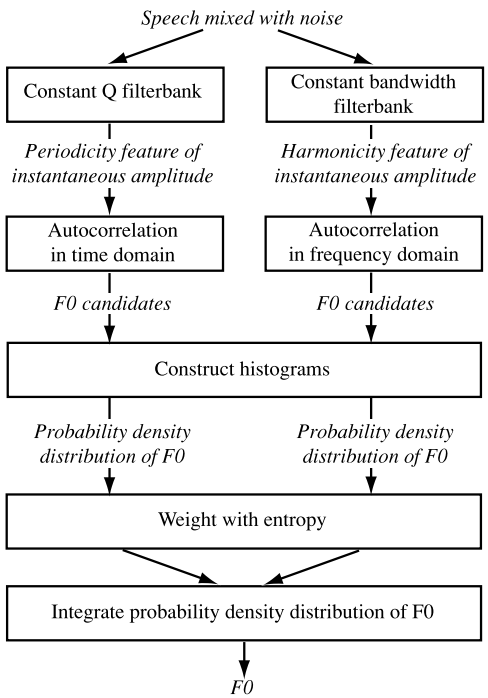



**Fig. 1** Block diagram of the proposed method.

dates, into narrow regions to select the low-noise regions.

A histogram of the F0 candidates is then constructed in each region. The histogram is considered to be a probability function of the F0 in the region. The distribution of the probability function in the low-noise region has a definite peak, whereas the distribution in the high-noise region is broad and has no definite peak. Entropy of the narrow distribution is low and that of the broad distribution is high. Hence, the method weights against the probability function using the reciprocal of entropy. It can emphasize the probability function with a correct peak of the F0 in the low-noise regions, and reduce influence of the probability function with incorrect peaks in the noisy regions. In other words, by utilizing entropy, this method can reduce distortion caused by noise.

In the final step, the weighted probability function of the F0 from the periodic feature and that from the harmonic feature are integrated. The frequency with the highest probability is extracted as the F0 of the signal.

### 2.2 Time-Frequency Representation of Instantaneous Amplitude

The time-frequency representation of the instantaneous amplitude of speech is obtained as follows. An input signal $s(t)$ is analyzed by using a filterbank. A band-limited signal $s_k(t)$ is then obtained for each filter, where $k$ is the channel number in the filterbank. Under the assumption that $\hat{s}_k(t)$ is the Hilbert transform of $s_k(t)$ with all frequency components delayed 90 degrees, the analytic signal $\tilde{s}_k(t)$ is given by

$$\tilde{s}_k(t) = s_k(t) + j\hat{s}_k(t). \quad (1)$$

The absolute value of the analytic signal $\tilde{s}_k(t)$ is the instantaneous amplitude of signal $s(t)$ in the center frequency of channel $k$.

This time-frequency analysis is performed by two filterbanks that represent the periodic and harmonic features of the instantaneous amplitude of speech. The instantaneous amplitude provided by a constant Q filterbank with high temporal resolution represents the periodic feature, which is a fluctuation corresponding to the fundamental periods of speech in the time domain. Similarly, the instantaneous amplitude provided by a filterbank with a constant narrow bandwidth represents the harmonic feature, which is a fluctuation corresponding to the F0s of speech and its multiples in the frequency domain. This is because a filterbank with the narrow bandwidth has a high frequency resolution. Figures 2 and 3 show the periodic and harmonic features of the instantaneous amplitude for the male vowel /a/. Candidates for F0 can be obtained from the periodic and harmonic features, because the peak intervals of the periodic feature equal the reciprocal of the F0s and

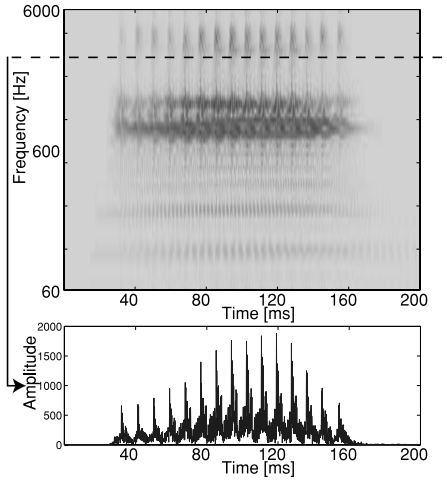



**Fig. 2** The periodic feature of instantaneous amplitude represented by the constant Q filterbank for the male vowel /a/ in the time-frequency domain.

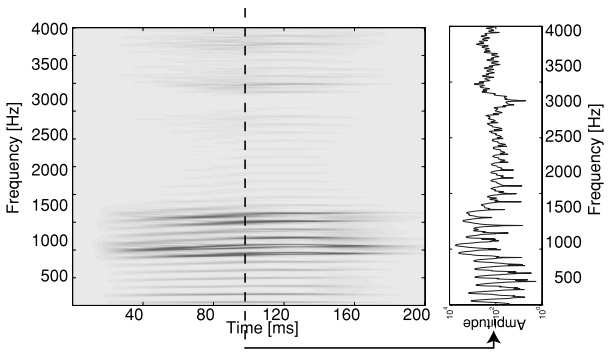



**Fig. 3** The harmonic feature of instantaneous amplitude represented by the constant bandwidth filterbank for the male vowel /a/ in the time-frequency domain.

those of the harmonic feature are the same as the F0s. In this paper, the filterbanks are constructed by using gammatone filters [15], [16]. A constant Q gammatone filterbank is constructed with 64 channels with center frequencies from 60 to 6000 Hz and bandwidth of about 90 Hz when the center frequency is 600 Hz. A constant-bandwidth gammatone filterbank is constructed with 800 channels with center frequencies from 60 to 4000 Hz and bandwidth of about 5 Hz.

### 2.3 Probability Function of F0 from Periodic and Harmonic Features

The proposed method calculates each probability function from the periodic and harmonic features in each time-frequency region.

From the periodic feature of the instantaneous amplitude of speech in the time-frequency domain, candidates for F0 are extracted using autocorrelation in each time region for each filter. Similarly, from the harmonic feature of the instantaneous amplitude, F0 candidates

are extracted using autocorrelation in each frequency region by dislocating the window of the autocorrelation. This means that we can obtain F0 candidates mapped to time-frequency planes. If the observed signal is noiseless, both F0 candidates determined from the periodic and harmonic features in the voiced section almost equal the F0s of the speech in the whole region. In the following, for the periodic feature, the frame length of the autocorrelation is about 20 ms, and the frame-shift is about 5 ms. For the harmonic feature, the frame length of the autocorrelation is about 1200 Hz, and the frame-shift is about 80 Hz.

For speech mixed with noise, the periodic and harmonic features do not appear clearly in noisy time-frequency regions. Hence, F0 candidates for the noisy signal are unsteady in the noisy region. Figure 4 shows the periodic and harmonic features of instantaneous amplitude for the male vowel /a/ mixed with band noise (1000–2000 Hz). The F0 candidates for the signal are plotted as a gray scale in the lower graphs. The F0 candidates extracted in the noisy region corresponding to 1000–2000 Hz do not show the collect F0, and they have various spurious values. However, in the low-noise region, i.e. below 1000 Hz or above 2000 Hz, the F0 candidates are the correct F0s.

To find the low-noise regions, the method divides time-frequency planes mapping the F0 candidates by narrow time-frequency windows. The F0 candidates in a window are used for constructing a histogram to distinguish whether the time-frequency region in the window is low-noise. The windows were defined as a time width of 20 ms and eight bands with equal log-frequency for the periodic feature, and eight bands with equal linear frequency for the harmonic feature. A time step of the windows is 1 ms. The histogram $h_{t,f}(k)$ is constructed from the F0 candidates in the window with time $t$ and $f$th frequency band, where $k$ is bin number of the histogram. Here, the bin width of the histograms is taken as 4 Hz. Next, the histogram $h_{t,f}(k)$ is normalized as

$$p_{t,f}(k) = \frac{h_{t,f}(k)}{\sum_k h_{t,f}(k)}, \tag{2}$$

which can be considered to be the probability function of F0 in the time $t$ and $f$th frequency region. Figure 5 shows the probability function $p_{t,f}(k)$ of the F0 of the vowel with band noise. The distribution of the probability function in the low-noise region has a sharp peak corresponding to the correct F0. The distribution in the noisy region is broad and does not have a definite peak.

In the case of both the periodic and harmonic features, the probability function $p_{t,f}(k)$ is calculated in the same way as described above.

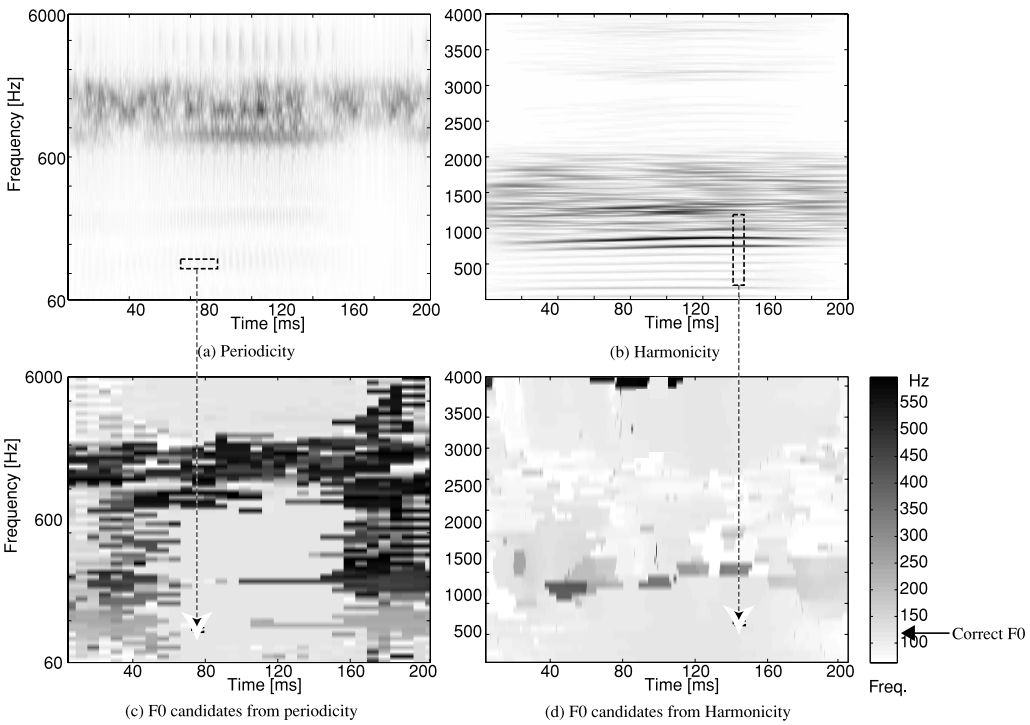



**Fig. 4** The periodic and harmonic features of instantaneous amplitude for the vowel with band noise (1000–2000 Hz), and F0 candidates. The gray scale of bottom panels indicates the F0 candidate calculated in each time-frequency region. The voiced section is 40–160 ms.

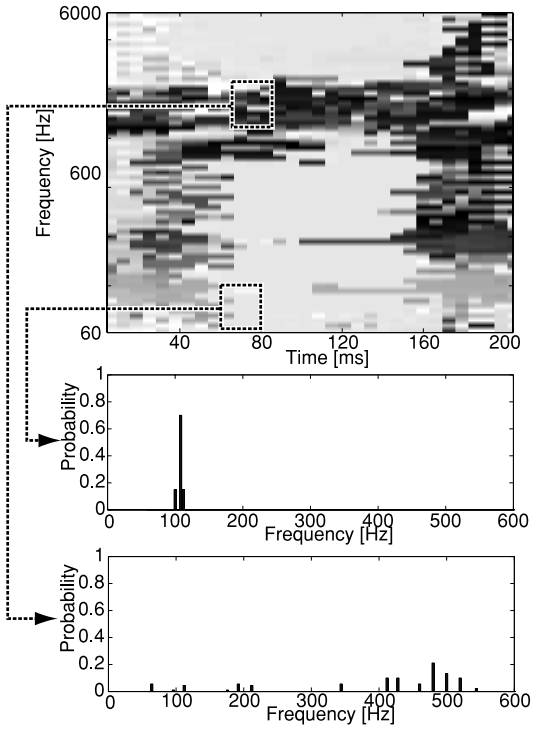



**Fig. 5** The probability function $p_{t,f}(k)$ of F0. Top panel: F0 candidates obtained from the periodic features of instantaneous amplitude of the vowel with the band noise; middle panel: the probability function of the F0 in the low-noise region; bottom panel: the probability function of the F0 in the noisy region.

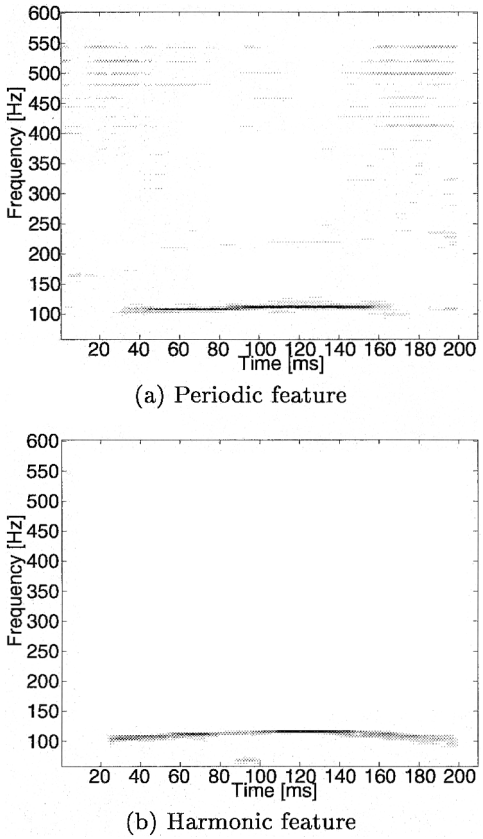



**Fig. 6** The probability function $p_t(k)$ for the clean vowel in the time-frequency domain. The gray scale indicates probabilities.

### 2.4 Weighted Probability Function of F0 by Entropy

Entropy $H_{t,f}$ is given by

$$H_{t,f} = -\sum_k p_{t,f}(k) \log p_{t,f}(k). \tag{3}$$

When the distribution of $p_{t,f}(k)$ has sharp peaks, the entropy of the distribution is low, and when the distribution is broad, the entropy is high. On this basis, the probability function $p_{t,f}(k)$ of F0 is weighted with the reciprocal of the entropy as

$$\bar{p}_{t,f}(k) = \frac{p_{t,f}(k)}{H_{t,f}}. \tag{4}$$

In other words, the entropy is used as a weight to emphasize the probability function of the F0 in the low-noise regions and reduce that in the noisy regions. The probability function $\bar{p}_{t,f}(k)$ can thus be grouped according to each time region as

$$p_t(k) = \frac{\sum_f \bar{p}_{t,f}(k)}{\sum_k \sum_f \bar{p}_{t,f}(k)}. \tag{5}$$

Figure 6 shows the probability function $p_t(k)$ of the F0 derived from the periodic and harmonic features of the clean vowel in the time-frequency domain. Both figures indicate that the regions with high probabilities correspond to the F0 contour. For noisy speech, however, the probability function without entropy-weighting has spurious peaks because of distortion by noise. Figure 7 (a) shows the probability function $p_t(k)$ without entropy-weighting in the case of the vowel with band noise. Figure 7 (b) shows the effect of entropy-weighting for the same signal. In contrast, the spurious peaks for the entropy-weighted probability function are reduced. The proposed method can thus effectively use features corresponding to F0 in the low-noise regions by utilizing entropy. Hence, it can be used to estimate reliable F0s in real environments.

The entropy of the probability function $p_t(k)$ tends to be low in the voiced section and high in the unvoiced section. It may be possible that the entropy will be used for voiced/unvoiced decision, but that discussion remains for future work.

### 2.5 Integration of the Probability Functions of F0 from Periodic and Harmonic Features

If both probabilities determined from the periodic and harmonic features are high, it is considered that they

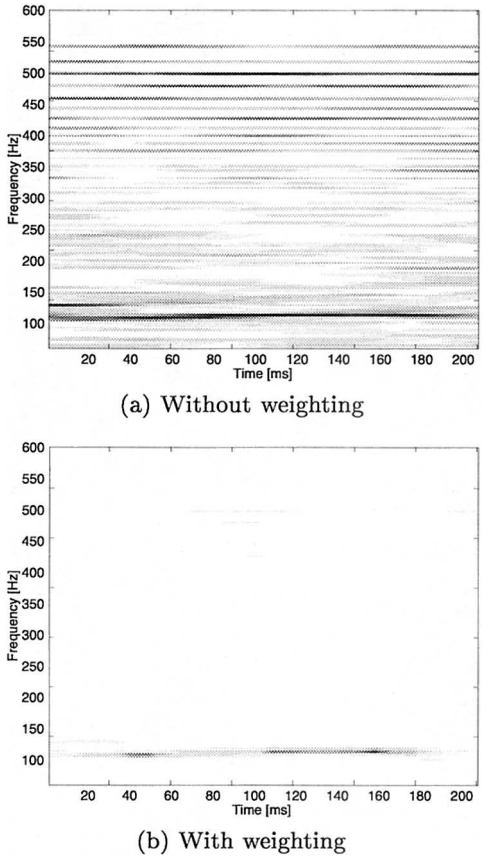



**Fig. 7** Improvement of the probability function by entropy-weighting for the periodic feature from the vowel with band noise.

indicate the correct F0. However, if one of them is high and the other is very low, they may not indicate the correct F0. To improve F0 estimation in noisy environments, the method integrates the two probability functions of F0 obtained from the periodic and harmonic features by Dempster's rule of combination

$$m(A_k) = \frac{\sum_{A_i \cap A_j = A_k} m_1(A_i) m_2(A_j)}{1 - \sum_{A_i \cap A_j = \phi} m_1(A_i) m_2(A_j)}, \tag{6}$$

where $m_1$ and $m_2$ are basic probability functions and $A_i$ and $A_j$ $(i, j = 1, 2, 3, \ldots)$ are focal elements [17]. The Dempster's rule intensifies the element in which two functions have high probability and weakens that in which one side of the two functions has low probability. This rule can deal with ignorance to represent lack of belief, although Bayesian theory deals with only belief and disbelief. However, ignorance is not dealt because of simple formulation in this paper. To use this rule for integrating the probability functions of F0, we regard the probability functions of F0 obtained from periodic and harmonic features as the basic probability function, and the frequency bin of the probability function as the focal element. That is, for the periodic features, $m(A_i)$ is corresponding to the probability function $p_{1t}(k)$ and $A_i$ is the frequency bin of the probability function. Similarly, for the harmonic feature, $m(A_j)$ is corresponding to the probability function $p_{2t}(k)$ and $A_j$ is the frequency bin of the probability function. The integrated probability function $g_t(k)$ is then calculated as

$$g_t(k) = \frac{p_{1t}(k) p_{2t}(k)}{\sum_k p_{1t}(k) p_{2t}(k)}, \tag{7}$$

Finally, the frequency with the highest probability in the probability function $g_t(k)$ is extracted as F0 in time $t$.

## 3. Experiments

To compare the robustness of the proposed method with that of others (i.e., the autocorrelation method, the cepstrum method, STRAIGHT-TEMPO [11] and PHIA [14]), we carried out two experiments using real speech mixed with band-limited noise and car noise. In the experiments, voiced/unvoiced decision was removed from STRAIGHT-TEMPO, because the other methods do not consider voiced/unvoiced decision. A database of simultaneous recordings of speech sounds and electroglottograph (EGG) [12] was used as the speech data in the experiments. The reference F0s of speech were as being equal to F0s extracted by original STRAIGHT-TEMPO with voiced/unvoiced decision from clean EGG data, because STRAIGHT-TEMPO has a high accuracy for clean speech [18]. The evaluation measures were the gross F0 error and the fine F0 error. The gross F0 error was derived by counting the samples that differ by more than 20% from reference F0s in voiced sections. The fine F0 error was defined as a standard deviation of the error within the threshold of the gross F0 error. The sampling frequency was 16 kHz.

### 3.1 Experiment 1 (Band-Limited Noise)

In this experiment, we used band-limited noise whose center frequency varied in time and whose bandwidth was fixed at 400 Hz. Figure 8 (a) shows the spectrogram of the band-limited noise. The SNR of speech mixed with the band-limited noise was 0 dB.

Figure 9 shows the estimated F0 contours and the gross F0 error determined by the five F0 estimation methods for a sentence uttered by a male in the presence of band-limited noise. Table 1 lists averages of the gross F0 error by each noise band for three sentences uttered by two male and two female speakers, and Table 2 lists the fine F0 error in the same conditions.

Table 1 clearly shows that the proposed method can robustly estimate F0s for all the band noises in comparison with the others. This means that the proposed

**Table 1** Gross F0 error for sentences mixed with band-limited noises.

| method | band noise [Hz] | | | | | |
|---|---|---|---|---|---|---|
| | 0–400 | 400–800 | 800–1200 | 1200–1600 | 1600–2000 | 2000–2400 |
| Autocorrelation | 55.3% | 51.5% | 42.3% | 37.4% | 36.4% | 37.3% |
| Cepstrum method | 17.3 | 9.5 | 10.8 | 10.4 | 10.9 | 3.3 |
| STRAIGHT-TEMPO | 44.2 | 61.6 | 40.0 | 1.1 | 2.0 | 0.4 |
| PHIA | 40.3 | 39.2 | 32.4 | 19.1 | 16.3 | 13.9 |
| Proposed method | 11.1 | 12.1 | 10.7 | 8.9 | 8.8 | 5.8 |

**Table 2** Fine F0 error for sentences mixed with band-limited noises.

| method | band noise [Hz] | | | | | |
|---|---|---|---|---|---|---|
| | 0–400 | 400–800 | 800–1200 | 1200–1600 | 1600–2000 | 2000–2400 |
| Autocorrelation | 13.5 Hz | 12.5 Hz | 10.4 Hz | 9.8 Hz | 8.2 Hz | 8.5 Hz |
| Cepstrum method | 6.9 | 6.6 | 6.3 | 5.8 | 5.3 | 5.3 |
| STRAIGHT-TEMPO | 4.9 | 4.9 | 2.6 | 2.4 | 2.5 | 2.6 |
| PHIA | 8.8 | 8.1 | 7.0 | 6.1 | 5.5 | 6.0 |
| Proposed method | 8.2 | 7.0 | 6.1 | 5.8 | 5.6 | 6.3 |

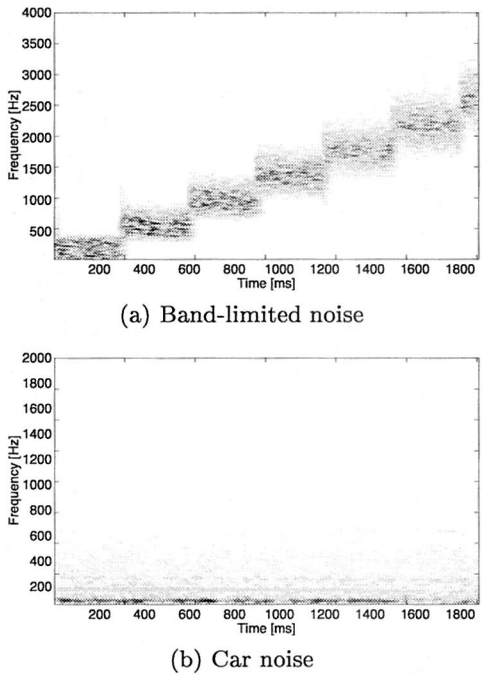



**Fig. 8** Spectrograms of noise used in the experiments.

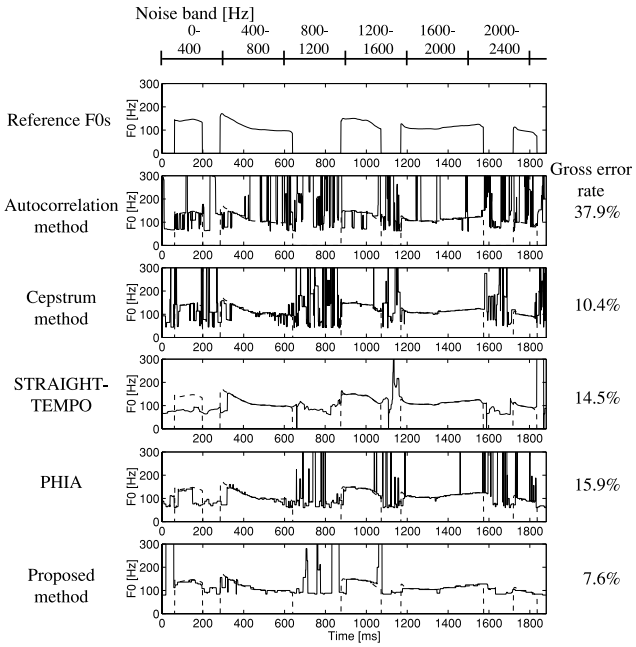



**Fig. 9** Estimated F0 contours and gross F0 error for a sentence uttered by a male with band-limited noise. The broken line indicates the same as the top panel.

method can use the local low-noise regions in the time-frequency domain and does not use a particular frequency region for extraction of F0. Table 2 shows that the proposed method has dispersion of F0 estimates against robustness to noise. It should be noted that the proposed method can estimate the F0 contours when the frequency of the band noise is 0–400 Hz, though the other methods cannot estimate them. The performance of PHIA (previously proposed by the authors) is low when the frequency of the noise band is low, because PHIA uses mainly harmonic components of speech in the low frequency regions. STRAIGHT-TEMPO uses the fundamental component for extracting F0s; therefore, the performance of STRAIGHT-TEMPO is very low when the frequency of the noise band is low. However, if there is no noise in the low frequency region, e.g., 0–1200 Hz, it has high performance. The cepstrum method has relatively high performance even for low-band noise, because it uses harmonic components of speech in the entire frequency range. However, since the cepstrum method impartially uses harmonic features in the noisy regions as well as in the low-noise regions, its performance is inferior to that of the proposed method. The autocorrelation method is sensitive to any band noises, because the high energy of noise distorts the waveform of speech even if the noise band is narrow.

### 3.2 Experiment 2 (Car Noise)

In this experiment, to investigate the robustness of the proposed method for real environmental noise, we used car noise recorded in the cabin of a moving car. The car noise is included in the JEIDA noise database [19]. Figure 8 (b) shows the spectrogram of the car noise, which clearly has high energy in the low-frequency regions and low energy in the high-frequency regions.

Figure 10 shows the estimated F0 contours and the gross F0 error determined by the F0 estimation methods for a sentence uttered by a male speaker with the car noise. The SNR of speech mixed with the car noise was 3 dB. The proposed method could robustly estimate F0s in almost all of the voiced sections. Moreover, the gross F0 error of the proposed method is more than 10% lower than that of the others.

Figure 11 shows the gross F0 error and the fine F0 error of the F0 estimation methods as a function of SNR. In this experiment, we used 14 sentences for 14 male and 14 female speakers. The speech data were mixed with car noise and the SNRs were 10, 5, 3, and 0 dB. The figure shows that the gross F0 error of the proposed method is the lowest when SNR is below 10 dB. In point of robustness to noise, it can be concluded that the performance of the proposed method is superior to that of the other methods in car noise environment. This entropy-weighted method developed with the aim of improving the robustness to real environmental noise, and the proposed method can be replaced with PHIA as the first part of [14] from this result. Because the fine F0 error of the proposed method was almost same as that of PHIA, the robustness of [14] to real environmental noise will improve by using the proposed method instead of PHIA. The proposed method could not provide the same accuracy as STRAIGHT-TEMPO for frequency resolution of the fiterbanks and the bin width of the histogram. However, we consider that the accuracy of the proposed method is permissible because the accuracy is improved by another part of [14].

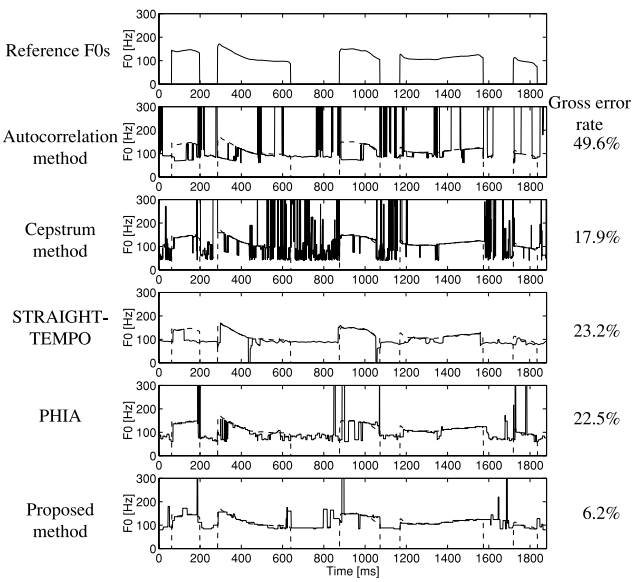
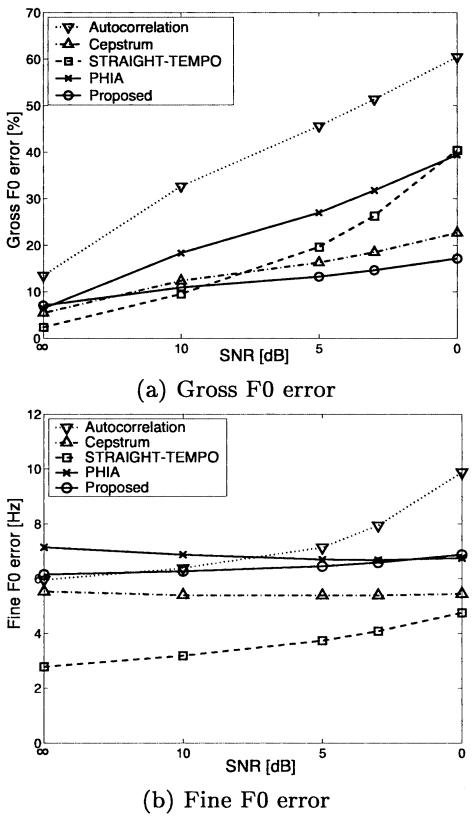



**Fig. 10** Estimated F0 contours and gross F0 error for a sentence uttered by a male with car noise. The broken line indicates the same as the top panel.

## 4. Conclusions

A robust F0 estimation method that employs periodic and harmonic features weighted by entropy in real environments was developed. The spectral structure of real environmental noise varies momentarily, and noise energy does not distribute evenly in the time-frequency domain. Therefore, segmenting a spectrogram of speech mixed with environmental noise into narrow time-frequency regions, will produce many low-noise regions. A feature of this method is to use entropy for emphasizing low-noise regions. The proposed method calculates the probability function of F0 from periodic and harmonic features of speech in each narrow time-frequency region. The probability function is then weighted using the reciprocal of entropy. This means that the method emphasizes the probability function in the low-noise regions to enhance the noise robustness.




(a) Gross F0 error

(b) Fine F0 error

**Fig. 11** Performance of the F0 estimation methods as a function of SNR between speech and car noise.

The method thus robustly estimates F0 from speech in the presence of real environmental noise. Experiments for speech mixed with band-limited noise and car noise were carried out to evaluate the robustness of the proposed method in noisy environments. The experiment results show that the proposed method can correctly estimate F0 from speech with noise in the low frequency region like 0–400 Hz, though the low-frequency noise distorts the fundamental component of speech. The performance of the method is superior to that of other F0 estimation methods when the SNR of speech to car noise is below 10 dB. In particular, its gross F0 error for speech with car noise is more than 5% lower than that of the others when the SNR is 0 dB.

## Acknowledgements

This work was undertaken when the first author was an intern at NTT Communication Science Laboratories. The authors would like to thank Dr. Hiroshi Murase and Dr. Makio Kashino for supporting our research. This work was supported by CREST (Core Research for Evolutional Science and Technology) of Japan Science and Technology Corporation (JST).

### References

[1] H. Singer and S. Sagayama, "Pitch dependent phone modeling for HMM based speech recognition," Proc. ICASSP92, vol.1, pp.273–276, 1992.
[2] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, vol.27, pp.187–207, 1999.
[3] A.S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound, MIT Press, Cambridge, MA, 1990.
[4] T. Nakatani, T. Kawabata, and H.G. Okuno, "A computational model of sound stream segregation with multi-agent paradigm," Proc. ICASSP95, vol.4, pp.2671–2674, 1995.
[5] M. Unoki and M. Akagi, "Signal extraction from noisy signal based on auditory scene analysis," Proc. ICSLP98, vol.4, pp.1515–1518, 1998.
[6] W. Hess, Pitch Determination of Speech Signals, Springer-Verlag, Berlin, 1983.
[7] W.J. Hess, "Pitch and voicing determination," in Advances in Speech Signal Processing, ed. S. Furui and M.M. Sondhi, pp.3–48, Marcel Dekker, New York, 1992.
[8] D.J. Hermes, "Pitch analysis," in Visual Representations of Speech Signals, ed. M. Cooke, S. Beet, and M. Crawford, pp.3–25, John Wiley & Sons, Chichester, 1993.
[9] L. Qiu, H. Yang, and S.N. Koh, "Fundamental frequency determination based on instantaneous frequency estimation," Signal Processing, vol.44, pp.233–241, 1995.
[10] T. Abe, T. Kobayashi, and S. Imai, "Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency," Proc. ICSLP96, vol.2, pp.1277–1280, 1996.
[11] H. Kawahara, H. Katayose, A. de Cheveigné, and R.D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," Proc. Eurospeech99, pp.2781–2784, 1999.
[12] Y. Atake, T. Irino, H. Kawahara, J. Lu, S. Nakamura, and K. Shikano, "Robust fundamental frequency estimation using instantaneous frequencies of harmonic components," Proc. ICSLP2000, vol.2, pp.907–910, 2000.
[13] L. Cohen, Time-frequency Analysis, Prentice Hall, New Jersey, 1995.
[14] Y. Ishimoto, M. Unoki, and M. Akagi, "A fundamental frequency estimation method for noisy speech based on instantaneous amplitude and frequency," Proc. Eurospeech2001, vol.4, pp.2439–2442, 2001.
[15] R.D. Patterson and J. Holdsworth, "A functional model of neural activity patterns and auditory images," in Advances in Speech, Hearing and Language Processing, ed. W.A. Ainsworth, E.F. Evans, and C.M. Hackney, vol.3, pp.547–563, JAI Press, London, 1991.
[16] M. Unoki and M. Akagi, "A method for signal extraction from noise-added signals," IEICE Trans. Fundamentals (Japanese Edition), vol.J80-A, no.3, pp.444–453, March 1997.
[17] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976.
[18] A. de Cheveigné and H. Kawahara, "Comparative evaluation of F0 estimation algorithms," Proc. Eurospeech2001, pp.2451–2454, 2001.
[19] S. Itahashi, "A noise database and Japanese common speech data corpus," J. Acoust. Soc. Jpn., vol.47, no.12, pp.951–953, 1991.

**Yuichi Ishimoto** received the B.E. degree in Electrical and Electronic Engineering from Utsunomiya University, the M.E. degree in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in 1997 and 2000, respectively. He is currently working towards the Ph.D. degree in Information Science at JAIST. His research interests include speech signal processing and computational auditory scene analysis. He is a member of Acoustical Society of Japan.

**Kentaro Ishizuka** received the B.E. degree in Information Science, and Master of Informatics degree in Intelligence Science and Technology from Kyoto University, Japan, in 1997 and 2000, respectively. Since 2000, he has been a member of NTT Communication Science Labs, NTT Corporation, Atsugi, Kanagawa, Japan. His research interests include acoustic, speech, signal processing, hearing psychology and physiology, speech perception, automatic speech recognition, and auditory scene analysis. He is a member of IEEE, Acoustical Society of America, and Acoustical Society of Japan.

**Kiyoaki Aikawa** received his Ph.D. from the University of Tokyo in 1980. He was then engaged in the NTT Basic Research Laboratory in 1980. He was a visiting scientist of Carnegie Mellon University in 1990. From 1992 to 1995, he was a senior researcher in Advanced Telecommunications Research Laboratories. After staying NTT Human Interface Laboratories from 1996 to 1998, he worked at NTT Communication Science Laboratories from 1999 to 2003. Since 2003, he has been with Tokyo University of Technology, where he is currently professor. He received the Sato Award from the Acoustical Society of Japan, and the Telecom-System Technology Award from the Electrical Communication Foundation in 1996.

**Masato Akagi** received the B.E. degree in Electronic Engineering from Nagoya Institute of Technology in 1979, the M.E. and the Dr.Eng. degrees in Computer Science from Tokyo Institute of Technology in 1981 and 1984, respectively. In 1984, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT). From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been with the School of Information Science, JAIST, where he is currently professor. His research interests include Speech Perception Mechanisms of Humans, and Speech Signal Processing. Dr. Akagi received the IEICE Excellent Paper Award from the IEICE in 1987, and the Sato Prize for Outstanding Paper from the ASJ in 1998.