Journal of Signal Processing, Vol. 12, No. 4, pp. 303-306, July 2008

SELECTED PAPER

Estimation of Local Peaks Based on Particle Filter in Adverse Environments

Seiji Tomoike and Masato Akagi

Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: {tomoike, akagi}@jaist.ac.jp

Abstract

In this paper, we propose an estimation method for local peaks of the speech spectrum using a particle filter in noisy environments. The conventional local peak estimation methods do not use the estimated peak knowledge in previous frames. These methods estimate the local peaks only in the current frame. Therefore, noises markedly affect the accuracy of local peak estimation in the current frame. The sudden incidence of peak candidates that are not appropriate for local peaks might be estimated in noisy environments. Thus, these methods have no robustness against nonstationary noise. To solve the problems of conventional peak-picking methods, we proposed a two-step estimation method for local peaks using a particle filter. The first step is to estimate the peak presence probability on the basis of the spectral envelope of the cepstrum. The local peaks can be simultaneously estimated using likelihood with the same peak presence probability in the high-probability regions. The second step is to extract peaks from the candidates of the peaks on the basis of the peak presence probability. Experimental results show that the proposed method is superior to conventional methods in terms of the frequency distance and the number of correct peaks in the nonstationary noisy environment.

1. Introduction

As one main characteristic of speech, harmonics play an important role for speech recognition, fundamental frequency (F0) estimation, speech enhancement, and so on. Harmonics are closely related to the local peaks of speech spectra in the frequency domain. Each of the harmonic components can be represented by amplitude and frequency. For the relationship between harmonics and speech, McAulay and Quatieri [1] shows that speech can be synthesized from harmonics using a speech analysis/synthesis system based on the sinusoidal model using the amplitude and frequency of harmonics. However, harmonics fluctuates at high frequencies; thus, the separate estimation of local peaks is required. Therefore, we focus on the local peaks estimation of the speech spectrum in the frequency domain in this paper.

For local peak estimation, many peak-picking algorithms have been reported thus far. The simplest method, named the second-order differential technique (SOD), first determines the local-maximum points using a second-order differential and the local peaks are decided by the slope. The improved method, named the hill-climbing technique (HC), disregards the very small amplitudes before searching the local-maximum points. The local peak estimation algorithm specialized for harmonics tracks harmonics of speech with instantaneous frequency (IF) [2]. The instantaneous frequency method uses harmonics.

The above methods often overestimate or underestimate the number of peaks. The conventional peak-picking methods have a drawback in that they do not use the already estimated peaks in previous frames. The noise markedly affects the accuracy of local peaks estimations in the current frame. Since harmonics varies gradually, the frequency positions of peaks in previous frames are important for estimating peaks in the current frame. Learning position from the previous frames facilitate the estimation of local peaks in the current frame in noisy environments.

Particle filter [4] is also used to estimate the state of the dynamic system from noisy observation. The function of a particle filter is to approximate the accurate posterior probability distribution using many particles, i.e., discrete values. The posterior probability distribution is represented according to the density of the population of the particles. As the number of particles increases toward infinity, the approximation approaches the true posterior probability. A solution for dealing with the drawbacks of the conventional methods is the accurate approximation of the posterior probability distribution by a particle filter. The posterior probability distribution gives the state transition probability. Thus, representing posterior probability distribution enables the blind prediction of local peaks.

To solve the problems of conventional peak-picking methods, we proposed a two-step estimation method for local peaks using a particle filter. The first step is to estimate peak presence probability. The likelihood of peaks is dynamically constructed using a spectral envelope of the cepstrum. The likelihood that describes the spectral envelope helps in determining whether peaks are present. To realize the simul-

taneous estimation of the state, we introduce a multidimensional likelihood. The second step is to extract peaks from the candidate peaks using the peak presence probability. The frequency bands that have maximal posterior peak presence probability become candidates of the peaks.

2. Problem Formulation

A noisy time signal $y(k)$ sampled at regular time intervals $k \cdot T$ and is composed of a clean target speech $s(k)$, which varies gradually, and additive noise $w(k)$, is given by

$$y(k) = s(k) + w(k) \tag{1}$$

After segmentation and windowing with a function $h(k)$, e.g., Hamming window, the DFT coefficient of frame t and frequency bin f is calculated using

$$Y_t(f) = \Sigma_{k=0}^{N-1} y(tL + k)h(l)e^{-j2\pi kf/N} \tag{2}$$

N denotes the DFT frame size. For the computation of the next DFT, the window is shifted by L samples. To decrease the disturbing effects of cyclic convolution, we apply overlapping. Then, we obtain the noisy DFT coefficient Y consisting of the speech part S and the noise part W. Thus, the observation model is

$$Y_t(f) = S_t(f) + W_t(f) \tag{3}$$

The relation between the state $X_t(f)$ and the speech $S_t(f)$ is

$$S_t(f) = A_t\big(X_t(f)\big) \tag{4}$$

where Eq. (4) represents the relationship between $S_t(f)$ and $X_t(f)$. We define $X_t(f)$ as the peak presence probability that is used to derive local peaks of harmonics. However, $S_t(f)$ is not completely constructed from $X_t(f)$, and A_t is the so called approximation function.

3. Particle Filter

Particle filtering can be used to estimate hidden variables representing the peak presence probability. The observed variable is contaminated by additive noise. The posterior probability distribution specifies the likelihood of each possible state given the observation. The hidden variable is estimated to maximize the distribution. The particle filtering algorithm approximates the distribution with Bayes's theorem using the likelihood $P\big(Y_t(f)|X_t(f)\big)$ and the state transition probability $P\big(X_t(f)|X_{t-1}(f)\big)$:

$$P\big(X_t(f)|Y_{1:t}(f)\big) \propto \prod_{m=1}^{M} P\big(Y_t(f)|X_t(f)\big)P\big(X_t(f)|X_{t-1}(f)\big) \tag{5}$$

where M is the number of targets, which are unknown. The particle filter is used to estimate the posterior probability distribution by using a lot of particles, i.e. discrete values. According to the central limit theorem, as the number of particles increases toward infinity, the approximation approaches the true posterior probability.

4. Estimation of Local Peaks

The state transition probability of local peaks represents the peak presence probability. The state $X_t(f)$ transits from the previous state to the current state by the nonlinear state transition function $Q_t(\cdot)$. Then we obtain the system model,

$$X_t(f) = Q_t\big(X_{t-1}(f)\big) + V_t(f) \tag{6}$$

where $V_t(f)$ is the system noise. The proposed method estimates $Q_t(\cdot)$ using the multidimensional likelihood that is dynamically updated using a particle filter. Therefore, no models of the state transition function are needed.

The first step is to estimate the peak presence probability by a particle filter. In the proposed method, the particle filter estimates the peak presence probability. To estimate local peaks of the speech spectrum, the likelihood is required. The likelihood gives the peak presence probability. The likelihood $P\big(Y_t(f)|X_t(f)\big)$ is dynamically constructed using the spectral envelope of the cepstrum given by

$$P\big(Y_t(f)|X_t(f)\big) = \begin{cases} 1 & , Y_t(f)/Cx \geq 1 \\ Y_t(f)/Cx & , Y_t(f)/Cx < 1 \end{cases} \tag{7}$$

where Cx is the smoothed spectral envelope using the cepstra.

As the frequency band of the observed spectrum, which is larger than the likelihood, gives high peak presence probability, the frequency band is set to the maximum probability. Thus, the peak presence probability has flat probabilities in the high-probability regions. The frequency band with the high probability gives a good candidate of local peak. The estimated peak presence probability $P\big(\hat{X}_t(f)\big)$ is derived as

$$\hat{X}_t(f) = P\big(Y_t(f)|X_t(f)\big)P\big(X_t(f)|X_{t-1}(f)\big) \tag{8}$$

The state transition probability is updated with

$$P\big(X_t(f)|X_{t-1}(f)\big) = \frac{\hat{X}_{t-1}(f)}{\Sigma_f \hat{X}_{t-1}(f)}, \; f = 1, 2, \cdots \tag{9}$$

The state transition probability is also constructed dynamically because the position of peaks in the next frame can be predicted by the peak presence probability.

In the particle filter, the degeneracy of a significant state often occurs by convergence. To attenuate the effects of degeneracy, a resampling method is used. The basis of the resampling is to eliminate particles that have small weights and

concentrate on new particles with large weights. The proposed method uses a resampling algorithm based on random sampling [5].

The algorithm of the proposed method is as follows:

1. Initialize variables
   $P(X_t(f)|X_{t-1}(f))= U(f)$, where $U(f)$ is uniform distribution.

2. For all time frames t
   - For all particles i
     (a) Calculate estimation (Eq. 8)
     (b) Updating state transition probability (Eq. 9)
   - Resampling step
     On the basis of [5], the certifying state space technique is introduced. The certifying state space technique put at least one particle in each state space for handling sudden change of the state transition.

To obtain local peaks from the candidate peaks, a method of extracting the peaks using the peak presence probability is required. The peak presence probability is used for extracting local peaks. The peak presence probability has flat probabilities in the high-probability regions. Local peaks might exist in the frequency ranges with high peak presence probabilities at a high degree. Local peaks can be extracted by picking the medium of the peak presence probability over the threshold of peak presence probability.

Figure 1 shows an example of the local peaks estimated from a really clean speech by the proposed method. The sampling frequency is 8 kHz with a 16 bit accuracy. For each frame, the set of points forming a vertical line represents estimated positions of local peaks. A harmonic structure is clearly seen in the figure.

5. Evaluation

To show the robustness of the proposed method against nonstationary noise, the accuracy of local peak estimations was evaluated. The learning of a speech spectrum was required before estimation when we used our proposed method. However the assumption of an initial clean speech was not required.

The evaluation consisted of two experiments. In experiment 1, we synthesized a noisy speech by adding pink noise from the first frame to the last frame, and comparisons between the proposed method and conventional methods were carried out. In experiment 2, we used a synthetic speech by adding the narrowband noise whose duration was set to two frames from the fifth frame as nonstationary noise, and comparisons between the proposed method and the conventional methods were carried out.

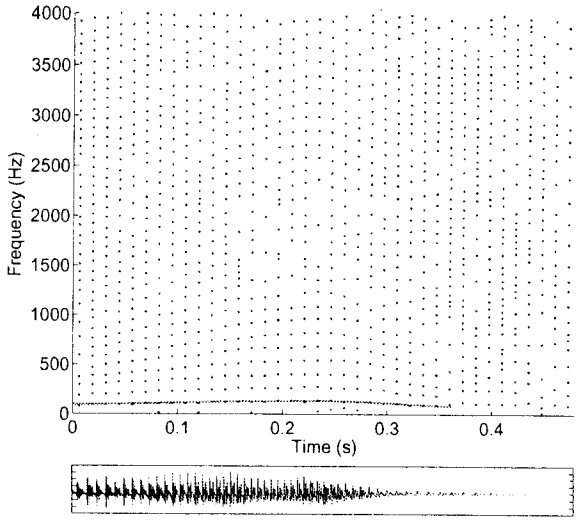

Figure 1: Harmonic frequencies of Japanese word /aoi/ uttered by male speaker. The bold line represents F0

To evaluate the performance quantitatively, we used a synthetic speech with preset positions of peaks. Synthetic speech signals were generated using a sinusoidal synthesis system [1] with the preset positions of peaks, and each amplitude of harmonic components is fixed to the same value. The evaluated methods are the proposed method (PF), second-order differential method (SOD), hill climbing method (HC) and instantaneous frequency method (IF) [2].

5.1 Evaluation measures

These methods are evaluated using two measures; the number of correct peaks and frequency distance between estimated peaks and preset peaks. We define the frequency distance between the estimated peaks and the correct peaks.

$$df = \overline{|f_{cor} - \hat{f}_{est}|} \tag{10}$$

where df is the frequency distance, f_{cor} is the number of correct peaks, and f_{est} is the number of estimated peaks. For each preset peak, the difference between the preset peak and the nearest estimated peak is calculated in the frequency domain. The distance of disaccord with the correct peaks and that of overestimated peaks are fixed to the F0. The result is averaged as the average distance.

5.2 Results and discussion

Figure 2 shows the results of experiment 1, in which the input SNRs are $-10, 0, 10, 20, \infty$. Figure 3 shows the results of experiment 2, in which the input SNRs are $-10, 0, 10, 20, \infty$. In these figures, approaching zero value indicates good estimation of local peaks.

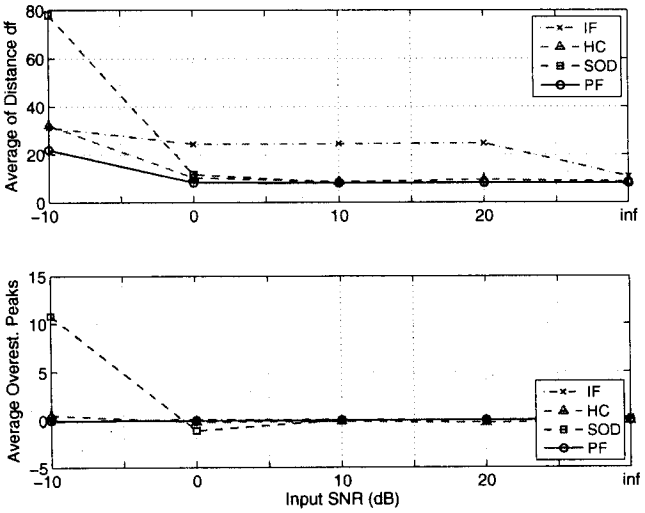

Figure 2: Average frequency distance and average number of overestimated peaks (uses pink noise, ○ is the proposed method, □ is the second-order differential method, △ is the hill-climbing method and × is the instantaneous frequency method)

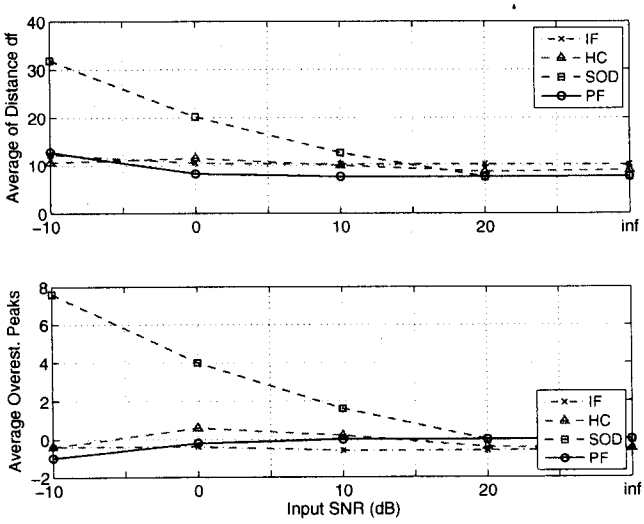

Figure 3: Average frequency distance and average number of overestimated peaks estimated (uses narrowband noise whose duration is two frames, ○ is the proposed method, □ is the second order differential method, △ is the hill-climbing method and × is the instantaneous frequency method)

In Fig. 2, PF yields good performance for both the number of overestimated local peaks and the frequency distance at input SNR of 0, 10, 20. SOD or HC estimate local peaks erratically in each input SNR. IF yields good performance for the number of estimated local peaks, while the distance is higher than that of the proposed method.

In Fig. 3, PF yields good performance for both the number of overestimated local peaks and the distance in the conditions of each input SNR. SOD overestimates most of the local peaks. HC overestimates the local peaks as input SNR decreases. The IF method yields good performance for the number of estimated local peaks, and its distance is higher than that of the proposed method.

The SOD and HC cannot distinguish local peaks of harmonics and peaks in noise; it is difficult to set optimal parameters for these methods. The IF method has to determine the number of local peaks and tracks these peaks made in the initial frame, and is affected widely by the noise in each input SNR. Since the proposed method can estimate local peaks dynamically using the learning of previous frames, the proposed method yields a good performance for both the number of overestimated local peaks and frequency distance. As a result, the proposed method can estimate local peaks correctly in nonstationary noisy environments. Therefore, the proposed method effectively uses the position of local peaks in the previous frames.

6. Conclusion

In this paper, we proposed an estimation method for local peaks of the speech spectrum using a particle filter. The proposed method effectively uses the position of local peaks in previous frames. The proposed method has great advantage for estimating local peaks of a speech spectrum in noisy environments such as short-term narrowband noise. We show the possibility of estimating local peaks of a speech spectrum in nonstationary noisy environments under the assumption that clean speech or high SNR speech exists in the initial frame.

Acknowledgement

A part of this study was supported by SCOPE (071705001) of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

[1] R. J. McAulay and T. F. Quatieri: Advances in Speech Signal Processing, Chapter 6, pp. 165-208. Marcel Dekker, 1991.

[2] T. Abe, T. Kobayashi and S. Imai: Harmonics tracking and pitch extraction based on instantaneous frequency, In Proc. IEEE ICASSP, pp. 756-759, 1995.

[3] R. E. Kalman: A new approach to linear filtering and prediction problems, Trans. ASME, J. Basic Engineering, Vol. 82-D, pp. 35-45, 1960.

[4] G. Kitagawa: Monte Carlo filter and smoother for non-Gaussian nonlinear state space models, J. Computational and Graphical Statistics, Vol. 5, pp. 1-25, 1996.

[5] M. S. Arulampalam, S. Maskell, N. Gordon and T. Clapp: A tutorial on particle filter for online nonlinear/non-Gaussian Bayesian tracking, IEEE Trans. Signal Processing, Vol. 50, No. 2, pp. 174-188, 2002.