# Extended Fundamental Frequency Extraction Using Exponentiated Amplitude Spectrum with Band-Limitation

Saori Motegi and Tetsuya Shimamura

*Abstract*—**In this paper, we implement the fundamental frequency extraction by using an exponentiated band-limited amplitude spectrum of speech. The exponent and bandwidth are adjusted according to male and female differences, characteristics of noise and noise amount. Basis function of the fundamental frequency extraction is calculated by inverse Fourier transform of an exponentiated amplitude spectrum. It is shown that this function invokes a performance robust against the white and car noise. Experimental results show that this method provides better performance than the conventional method at a wide range of signal-to noise ratios.**

*Index Terms*—**Component, fundamental frequency extraction, autocorrelation, band-limitation, exponentiated spectrum.**

## I. Introduction

All The fundamental frequency extraction is an important task in many speech processing systems [1][2]. Therefore, extracting the fundamental frequency of a speech signal is essential for research in speech processing and many methods have been proposed [1]-[6]. However, a performance improvement in noisy environments is still desired.

Correlation based processing is known to be comparatively robust against random noise. The autocorrelation function (ACF)[4] method is classified into this category, and behaves robustly in noisy environments[1]. However, the ACF has the property of being sensitive to the formant characteristic of vocal tract. On the other hand, there is the cepstrum (CEP)[5] method, which is known as a fundamental frequency extraction method to provide a comparatively insensitive performance to the formant characteristics. The CEP method is, however, sensitive to noise. For this reason, the performance of the CEP method degrades in noisy environments.

The main difference between the ACF and CEP is how to use the spectrum. The ACF is calculated by inverse Fourier transform of a squared amplitude spectrum (power spectrum) of speech, while the CEP uses a logarithmic amplitude spectrum (log spectrum) of speech. If it is noted the fact that the power spectrum is an expanded spectrum and the log spectrum is a compressed spectrum, then it is expected to improve the performance by using both the expansion and compression of spectrum properly. The method derived in [6] is based on this idea, which provides a robust performance in white noise covering a wide range of signal-to noise ratios(SNR). In this paper, we extend the function used in [6] and investigate the performance against a periodical noise like car noise as well as white noise.

In this paper, the fundamental frequency extraction method commonly uses an exponentiated band-limited amplitude spectrum of speech. However we adjust the exponent and bandwidth of the amplitude spectrum, according to male and female differences, characteristics of noise and noise amount unlike in [6].

The remainder of this paper is organized as follows. Section II describes the principle of the present method. In Section III, we show some experimental results. Finally, we conclude this paper in Section IV.

Manuscript received June 12, 2012; revised July 22, 2012.

The authors are with School of Science and Engineering, Saitama University, Shimo-Okubo 255, Sakura-ku, Saitama, 338-8570, Japan (e-mail: abe@sie.ics.saitama-u.ac.jp, shima@sie.ics.saitama-u.ac.jp)

## II. Principle

### A. Band-Limitation

The speech signal is known to have attenuation characteristics from low to high frequency regions, which is typically -6 dB/octave. On the other hand, the energy of white noise is constant in all frequency regions. Therefore, in white noise environments, the speech power is less than the noise power in high frequency region when the SNR of input speech is comparatively low. To avoid this problem, we set out to use a band-limitation operation on the amplitude spectrum of speech in low SNR conditions. In this case, the band used in this work is mainly the low frequency one. However, in high SNR conditions, the speech power is almost beyond the noise power regardless to the frequency regions. Thus, if we use the band similar to the case in the low SNR conditions, the performance of fundamental frequency extraction would degrade because of lack of speech information. Therefore, we have to adjust a suitable band according to the additive noise level included in the noisy speech signal.

On the other hand, car noise conditions are different. A narrow band at a low frequency region is emphasized due to the existence of a car noise. Therefore, the fundamental frequency extraction is unaffected by such a noise in high frequency region. Thus, in car noise environments, we need to implement the fundamental frequency extraction without band-limitation, unlike in white noise environments.

### B. Exponentiated Amplitude Spectrum

The ACF method is useful in noisy environments. The ACF calculated by inverse Fourier transform of a power spectrum is a class of the $p$-th power amplitude spectrum. When the exponent is set to $p = 2$, it reduces to the ACF.

On the other hand, the CEP method is useful in noiseless environments. The CEP is calculated by inverse Fourier transform of a log spectrum, which also belongs to a class of the generalized function calculated by inverse Fourier transform of the $p$-th power amplitude spectrum.

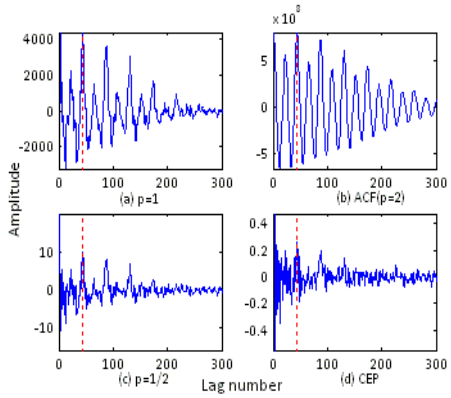



Fig. 1. True fundamental period and waveforms calculated by inverse Fourier transform of spectra in white noise

Fig. 1 shows waveforms calculated by inverse Fourier transform of *p*-th power amplitude spectra of a voiced frame of speech in a severe white noise environment where the dotted line indicates the true fundamental period. Figure 2 shows the one similar to Fig.1 in a severe car noise environment. Figure 3 shows waveforms calculated by inverse Fourier transform of *p*-th power amplitude spectra of another voiced frame of speech in a severe white noise environment where again the dotted line indicates the true fundamental period.

In Fig.1, (b) is better than (c) and (d). Fig. 1 (b) is a waveform calculated by inverse Fourier transform of 2-nd power spectrum which is recognized as an expanded spectrum. The difference between speech spectrum and noise spectrum is expanded relatively in a severe white noise environment as the exponent *p* increases. This is the reason why we can achieve the noise reduction by using larger *p*-th power amplitude spectra directly for fundamental frequency extraction, as shown in Fig.3. However, as the exponent *p* increases, the influences of the formant characteristics of vocal tract also increase. As a result, some formant peaks are emphasized, and this makes fundamental frequency extraction difficult. Therefore, it does not mean that it is good to increase the exponent *p* in a severe noise environment as well. And in a noiseless or almost noiseless condition, it is good to use smaller *p*-th power amplitude spectra. From these, it is expected that the performance of fundamental frequency extraction is improved by adjusting the suitable exponent *p* according to the additive noise level.

On the other hand, in car noise environments, in Fig.2, (c) and (d) are better than (b). In Fig.2, (c) and (d) are waveforms calculated by inverse Fourier transform of 1/2-th power and log spectra which are recognized as compressed spectra. The characteristics of car noise spectrum are similar to the formant characteristics of vocal tract. And, the CEP method is comparatively insensitive to the formant characteristic. Therefore, the fundamental frequency extraction against car noise should be implemented based on compressed spectrum. Thus, we implement the fundamental frequency extraction by using 1/*p*-th power amplitude spectrum in car noise environments.

### C. Function for Fundamental Frequency Extraction

In order to improve the performance of the ACF method, the use of auto-regressive inverse filtering has been suggested to flatten the speech spectrum. By this preprocessing, the true period peaks in the ACF are emphasized and the performance of the ACF method is improved. However, when few harmonics are present in the speech spectrum or in a noisy environment, the process of implementing the inverse filter itself may lead to erroneous results[2]. Therefore, in this paper, the ACF is calculated without this preprocessing as

$$R(\tau) = \mathrm{IDFT}(|X(f)|^2) \tag{1}$$

where $|X(f)|$ is the amplitude spectrum of the speech signal and IDFT denotes inverse Fourier transform. The region for searching the peak of fundamental period is set to be from 50 Hz to 400 Hz, which corresponds to the region of the fundamental frequencies most men and women have.

In this paper, the function for the fundamental frequency extraction by using p-th power amplitude spectrum with band-limitation is calculated by

$$R^p_{band}(\tau) = \mathrm{IDFT}(|X_{band}(f)|^p) \tag{2}$$

where $|X_{\mathrm{band}}(f)|$ means the band-limited amplitude spectrum of the speech signal.

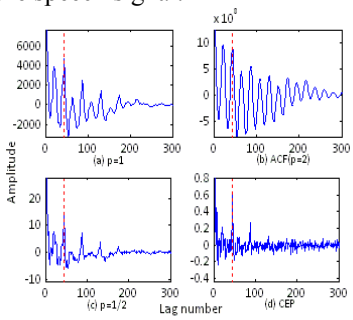



Fig. 2. True fundamental period and waveforms calculated by inverse Fourier transform of spectra in car noise

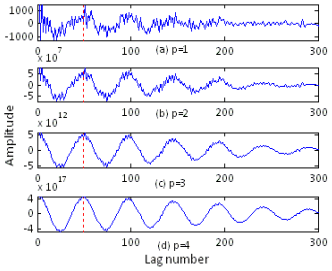



Fig. 3. True fundamental period and waveforms calculated by inverse Fourier transform of spectra in white noise

In advance, we need determine the exponent *p* and the bandwidth in order to calculate the above-mentioned function $R^p_{band}(\tau)$. The exponent *p* and the bandwidth should be adjusted according to the white noise level included in the speech signal.

In a car noise case, the fundamental frequency should be extracted by using 1/*p*-th power amplitude spectrum directly without band-limitation. Thus, the function is calculated by

In this paper, the function for the fundamental frequency extraction by using p-th power amplitude spectrum directly with band-limitation is calculated by

$$R^p_{band}(\tau) = IDFT(|X_{band}(f)|^{1/p}) \tag{3}$$

according to the car noise level.

In experiments in Section III, we implement the fundamental frequency extraction by calculating the function $R_{band}^{p}(\tau)$. To do this, we use a relationship between the exponent $p$ and SNR and that between the bandwidth and SNR. By a preliminary test in Section III, these are found at first. In utilizing these relationships, we need estimate the SNR of every voiced frame for the purpose of determining the exponent $p$ and bandwidth. The voiced SNR of every frame is calculated by

$$\mathrm{SNR}(\lambda) = 10\log_{10}\frac{\sum P_{speech}(\lambda, f)}{\sum P_{noise}(\lambda, f)} \tag{4}$$

where λ is the frame number, and $P_{speech}(\lambda,f)$ and $P_{noise}(\lambda,f)$ are the speech and noise power spectra. In this way, we determine the suitable exponent $p$ and bandwidth at a specified SNR.

## III. Experiments

### A. Experimental Details

Speech data in the experiments were taken from "20 Countries Language Database (NTT Advanced Technology)." We used the speech data uttered by four male and four female speakers. Each of the speech data consisted of about 10s Japanese sentences, which were sampled by a rate of 10 kHz with band-limitation of 3.4 kHz. The experiments were conducted by adding white Gaussian or car noise to the speech data. The other details in the experiments are as follows:

- window : Hamming window,
- window size : 51.2 ms,
- frame shift : 10 ms.

Each SNR of the noisy speech used in experiments was -5 dB, 0 dB, 5 dB, 10 dB. Only in the white noise case, experimental results were averaged for 10 independent trials on each speech data.

### B. Evaluation Method

The following was used for the evaluation of fundamental frequency extraction accuracy;

$$e(n) = F_{true}(n) - \hat{F}(n) \tag{5}$$

where $F_{\mathrm{true}}(n)$ is the true fundamental frequency, $\hat{F}(n)$ is the extracted fundamental frequency, and thus $e(n)$ is the extraction error for the $n$-th frame.

If $|e(n)| \geqq 10$ Hz, we recognized the extraction error rate as gross pitch error(GPE).

By inspection of clean speech waveforms in the time domain, the true fundamental period was obtained, and then $F_{\mathrm{true}}(n)$ was calculated as its inverse.

### C. Preliminary Test

At first, in order to determine the relation between the exponent $p$ and SNR, we implemented the fundamental frequency extraction by shifting the exponent $p$ at a specified SNR and calculated GPE.

| exponent p | −5 | 0 | 5 | 10 | 15 |
|---|---|---|---|---|---|
| 5 | 33.015 | 27.155 | 25.192 | 24.825 | 24.648 |
| | 30.151 | 23.95 | 21.998 | 21.619 | 21.561 |
| 4 | 26.959 | 20.312 | 18.334 | 17.847 | 17.714 |
| | 23.807 | 16.494 | 13.932 | 13.423 | 13.362 |
| 3 | 20.382 | 12.016 | 9.1744 | 8.594 | 8.485 |
| | 18.031 | 8.3591 | 4.9743 | 4.154 | 3.9531 |
| 2 | 17.186 | 6.7095 | 3.0053 | 2.1983 | 1.9409 |
| | 18.344 | 6.6095 | 2.3269 | 1.284 | 1.0006 |
| 1 | 21.602 | 7.6793 | 2.2547 | 0.9459 | 0.521 |

SNR[dB]

(a) male case

| exponent p | −5 dB | 0 dB | 5 dB | 10 dB | 15 dB |
|---|---|---|---|---|---|
| 5 | 14.056 | 10.218 | 9.1546 | 8.9918 | 8.8946 |
| | 13.086 | 8.6804 | 7.6576 | 7.4222 | 7.405 |
| 4 | 12.32 | 7.3605 | 6.159 | 5.8671 | 5.8174 |
| | 12.379 | 6.0444 | 4.6875 | 4.4408 | 4.4667 |
| 3 | 13.679 | 5.333 | 3.4911 | 3.1209 | 3.0461 |
| | 16.386 | 5.4835 | 2.7971 | 2.2624 | 2.1811 |
| 2 | 20.877 | 7.2671 | 2.716 | 1.723 | 1.496 |
| | 26.343 | 11.531 | 4.8233 | 2.3672 | 1.7939 |
| 1 | 32.98 | 16.926 | 8.6553 | 4.5931 | 2.5658 |

SNR[dB]

(b) female case

Fig. 4. Relation between the exponent $p$ and SNR in white noise

| exponent p | −5 | 0 | 5 | 10 | 15 |
|---|---|---|---|---|---|
| 5 | 21.039 | 10.093 | 4.918 | 2.767 | 1.8769 |
| | 20.943 | 9.9428 | 4.8575 | 2.633 | 1.8118 |
| 4 | 21.0 | 9.8343 | 4.8327 | 2.4805 | 1.6402 |
| | 21.22 | 9.9019 | 4.6819 | 2.3618 | 1.4912 |
| 3 | 21.635 | 9.545 | 4.4006 | 2.1908 | 1.365 |
| | 22.657 | 9.8553 | 4.4551 | 2.0706 | 1.2789 |
| 2 | 24.455 | 10.43 | 4.4087 | 2.0013 | 1.1704 |
| | 28.798 | 11.949 | 4.953 | 2.0884 | 1.0655 |
| 1 | 39.465 | 17.226 | 6.2269 | 2.3451 | 1.2533 |

SNR[dB]

Fig. 5. Relation between the exponent $p$ and SNR in car noise

Fig. 4 shows the result in white noise environments. In Fig.4, (a) and (b) show male and female cases. This is because male and female characteristics are different. Figure 5 shows the result of all speakers in car noise environments. This is because male and female characteristics are similar in car noise environments. In Fig.4 and 5, dark color parts mean the best performance at a specified SNR. We draw a line which goes through the dark color parts. This line indicates the relation between the exponent $p$ and SNR, which is expressed by the following equation;

- in the white noise case
  -for male case

$$p(\mathrm{SNR}) = \begin{cases} -\frac{1}{10} * \mathrm{SNR} + \frac{3}{2}, & \mathrm{SNR} < 5\,\mathrm{dB} \\ 1, & \text{otherwise} \end{cases}$$

  - for female case

$$p(\mathrm{SNR}) = \begin{cases} -\frac{1}{5} * \mathrm{SNR} + 3, & \mathrm{SNR} < 5\,\mathrm{dB} \\ 2, & \text{otherwise} \end{cases}$$

- in the car noise case

$$p(\mathrm{SNR}) = \begin{cases} -\frac{1}{8} * \mathrm{SNR} + \frac{27}{8}, & \mathrm{SNR} < 15\,\mathrm{dB} \\ \frac{3}{2}, & \text{otherwise} \end{cases}$$

Secondly, we need determine the relation between the bandwidth and SNR in white noise environments. In a way similar to the above test, we implemented the fundamental frequency extraction by shifting the bandwidth at a

specified SNR and calculated GPE. The minimum band considered in this work for the fundamental frequency is 50 Hz - 400 Hz. The maximum one is 50 Hz - 3600 Hz. The speech data are band-limited in an increment of 400 Hz.

Figure 6 shows the result. In this test, the result is averaged for male and female speakers, because male and female characteristics have similar tendencies. Using Fig.6, we determine the bandwidth at a specified SNR by searching the best performance at each SNR. It is given as;

$$\begin{cases} \text{from 50 Hz to 800 Hz,} & \text{SNR} < 2\ \text{dB} \\ \text{from 50 Hz to 1200 Hz,} & 2\ \text{dB} \le \text{SNR} < 10.5\ \text{dB} \\ \text{from 50 Hz to 3200 Hz,} & 10.5\ \text{dB} \le \text{SNR} \end{cases}$$

### A. Performance Comparison

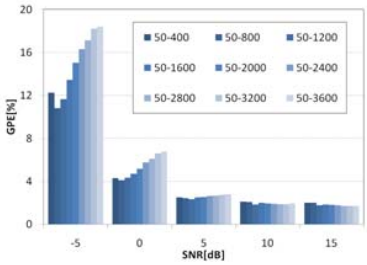



Fig. 6. Relation between the bandwidth and SNR

To investigate the accuracy of the fundamental frequency extraction, we conducted experiments. The ACF method is compared with seven methods to shift the exponent $p$ and bandwidth as follows;

- the exponent is constant($p$=2) and bandwidth is "fixed"(Lim(fixed)+Exp(const)).
- the exponent is constant($p$=2) and bandwidth is "variable"(Lim(variable)+Exp(const)).
- the exponent is "fixed" and without band-limitation(Lim(non)+Exp(fixed)).
- the exponent is "variable" and without band-limitation(Lim(non)+Exp(variable)).
- the exponent is ``fixed'' and bandwidth is "fixed"(Lim(fixed)+Exp(fixed)).
- the exponent is ``variable'' and bandwidth is "variable"(Lim(variable)+Exp(variable)).
- the exponent is "fixed" and bandwidth is ``variable''(Lim(fixed)+Exp(variable)).

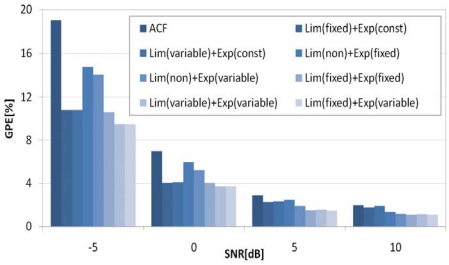
where "fixed" indicates adjusting to the value which corresponds to the SNR of noisy speech in advance, and "variable" indicates shifting to the value which corresponds to the SNR calculated by (4) at every frame.

In the car noise case, we conducted experiments where four methods are compared. The two are the ACF and CEP methods, and the others are the methods to shift the exponent $p$ as follows;

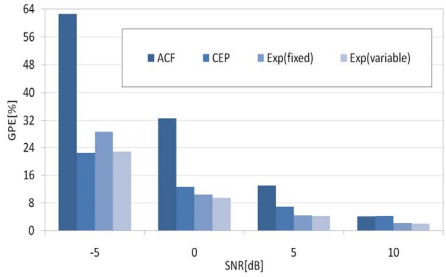
- the exponent is "fixed" (Exp(fixed)).
- the exponent is "variable" (Exp(variable)).

Figure 7 shows the results of the performance evaluation in white and car noise environments. From Fig.7(a), it is observable that the use of band-limitation provides a significant improvement at very low SNR (see Lim(fixed)+Exp(const) and Lim(variable)+Exp(const)). On the other hand, the use of exponentiated amplitude spectrum is good at high SNR (see Lim(non)+Exp(fixed) and Lim(non)+Exp(variable)). And, combining the band-limitation and exponentiated amplitude spectrum improves the performance at a wide range of SNRs (see Lim(fixed)+Exp(fixed), Lim(variable)+Exp(variable) and Lim(fixed)+Exp(variable)).

From Fig.7(b), Exp(fixed) and Exp(variable) show a good performance while there is little difference compared with the CEP method. At low SNRs, it is observable that Exp(fixed) and Exp(variable) provides significant improvement relative to the ACF method. Exp(variable) is good at a wide range of SNRs.




(a) White noise case




(b) Car noise case

Fig. 7. Results in GPE

## IV. Conclusion

In this paper, we have investigated the accuracy of the fundamental frequency extraction by using an exponentiated band-limited amplitude spectrum of speech. Experimental results have shown that the fundamental frequency extraction, which adjusts exponent and bandwidth of the amplitude spectrum according to male and female differences, characteristics of additive noise and noise amount, provides a superior performance at a wide range of SNRs in both of white and car noise environments.

## References

[1] L. R. Rabiner et al., "A comparative performance study of several pitch detection algorithms,"

[2] W. J. Hess, "Pitch and voicing determination," in *Advances in Speech Signal Processing edited by S.Furui and M.M.Sondhi*, Marcel Dekker, Inc. 1992

[3] A. Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am*, vol. 111, no .4, pp.1917-1930, 2002

[4] L. R.Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. on Acoust*, Speech and Signal Process, vol. ASSP-25, no. 1, pp. 24-33, 1977

[5] A. M. Noll, "Cepstrum pitch determination," *J.Acoust. Soc. Am*, vol. 41, no. 2, pp. 293-309, 1967

[6] T. Shimamura and H. Takagi, "Noise-robust fundamental frequency extraction method based on exponentiated band-limited amplitude spectrum," in *Proc. IEEE International Midwest Symposium on Circuits and Systems*, II, pp.141-144, 2004