4th WSEAS International Conference on ELECTRONICS, CONTROL and SIGNAL PROCESSING, Miami, Florida, USA, 17-19 November, 2005 (pp.47-52)

# Spectral Subtraction with Non-Stationary Noise Estimation Utilizing Harmonic Structure

TETSUYA SHIMAMURA and JUNPEI YAMAUCHI
Department of Information and Computer Sciences
Saitama University
255 Shimo-Okubo, Sakura-Ku, Saitama, 338-8570
JAPAN
http://www.sie.ics.saitama-u.ac.jp

*Abstract:* - The spectral subtraction (SS) method is well known as a speech enhancement technique and has been widely used. In this paper we study the noise spectrum estimation required for the SS method. It is set out to estimate directly the noise spectrum from the noise-corrupted speech frame. To accomplish this, a fundamental frequency of speech is detected and harmonic structure is constructed and utilized. We assume that non-stationary noise consists of its stationary part and non-stationary part. The non-stationary part is estimated from the harmonics obtained, while the stationary part is estimated in the conventional way. Both parts are combined, resulting in an accurate estimate of the noise spectrum. It is shown by experiments that the proposed method provides a performance improvement relative to the conventional SS methods in non-stationary noise environments.

*Key–Words:* - Speech enhancement, spectral subtraction, noise spectrum, non-stationary noise estimation, harmonic structure

## 1 Introduction

For the purpose of reducing noise in a noisy speech signal obtained by a single microphone, there exist many approaches. Wiener filter [1], spectral subtraction (SS) [4] [5], and MMSE [6]are classified into the frequency domain approach. Time domain class includes comb filter [3]. Kalman filter [2], Hidden Markov Model [7] and EVRC [8] may be of parametric model based approach.

In this paper, we consider the SS method because it is easy to implement and has been widely utilized in real speech processing systems. In the SS method, however, the noise spectrum must be estimated and subtracted from the noisy speech spectrum. To estimate the noise spectrum, non-speech segments are often used. This is based on the assumption that the noise is stationary. Thus, in the case where the noise is non-stationary, the technique is not satisfied. Martin's approach [5] conquers this problem, and provides a good result of speech enhancement. Even for Martin's method, however, the result is dependent on the noise characteristics. In this paper, we propose a new technique to estimate the noise spectrum directly from the noise-corrupted speech frame, and apply it to the SS method. By experiments, it is shown that the proposed method has a superior capability to track a highly non-stationary noise.

## 2 Spectral Subtraction

The SS method is described briefly in this section. Assume that a noisy speech signal is expressed as

$$y(n) = x(n) + d(n) \tag{1}$$

where $x(n)$ and $d(n)$ are clean speech and noise, respectively. In the frequency domain, the above equation is expressed as

$$Y(k) = X(k) + D(k) \tag{2}$$

where $Y(k)$, $X(k)$ and $D(k)$ are discrete-time Fourier transforms (DFT) of $y(n)$, $x(n)$ and $d(n)$, respectively. Based on this frequency domain expression, the SS method is implemented in the frame as follows:

$$\begin{cases} \tilde{X}(k) = \left( \dfrac{|Y(k)|^\gamma - a|\tilde{D}(k)|^\gamma}{|Y(k)|^\gamma} \right)^{1/\gamma} Y(k), \\ \qquad |Y(k)|^\gamma - a|\tilde{D}(k)|^\gamma > \beta|\tilde{D}(k)|^\gamma \\ \tilde{X}(k) = \left( \dfrac{a|\tilde{D}(k)|^\gamma}{|Y(k)|^\gamma} \right)^{1/\gamma} Y(k), \\ \qquad \text{otherwise.} \end{cases} \tag{3}$$

where $a$, $\gamma$ and $\beta$ in (3) are parameters to be set for implementation. They are called as weighting factor, power coefficient, and spectral flower, respectively.

The estimate of the speech spectrum, $\tilde{X}(k)$, is transformed into the time domain by inverse DFT, and results in an enhanced speech signal $\tilde{x}(n)$.

In a real world, $D(k)$ cannot be obtained directly. Thus, an estimate of $D(k)$ is used. In this paper, we propose a new technique to estimate $D(k)$.

# 3 Conventional Noise Estimation Methods

Two noise spectrum estimation methods are described in this section.

## 3.1 Noise Estimation from Non-Speech Segments

This method is commonly used in many references. In the recent paper by Virag [4], the following version of this method is used:

$$\begin{cases} \tilde{D}(k) = |Y(\lambda,k)|^2, \\ \qquad \lambda = 1 \\ \tilde{D}(\lambda,k) = \alpha \cdot \tilde{D}(\lambda-1,k) + (1-\alpha)|Y(\lambda,k)|^2, \\ \tilde{D}(k) = \tilde{D}(\lambda,k) \\ \qquad \lambda \le T_I \\ \tilde{D}(k) = \tilde{D}(T_I,k), \\ \qquad \lambda > T_I \end{cases} \tag{4}$$

where $\lambda$ is the number of frame. The $\alpha$ is the forgetting factor, whose values are ranged for 0.5~0.9. The $T_I$ corresponds to the last non-speech frame number before the speech frame.

This method is essentially used for stationary noise case. For non-stationary noise case, it is impossible to track the time variation the noise has.

This method is referred to as non-speech pause (NSP) method in this paper.

## 3.2 Minimum Statistics Method

Martin's method [5] is so-called the minimum statistics (MS) method. For the MS method [5], a subband noise spectrum is calculated from the input signal spectra $Y(\lambda,k)$ as

$$\tilde{D}_Y(\lambda,k) = \alpha \cdot \tilde{D}_Y(\lambda-1,k) + (1-\alpha)|Y(\lambda,k)|^2 \tag{5}$$

where $\alpha$ is the forgetting factor, which is set as 0.9~0.95. Next, the minimum spectrum is found from the calculated subband noise spectra $\tilde{D}_Y(\lambda,k)$ as

$$\begin{aligned} \tilde{D}_{min}(\lambda,k) = min[&\tilde{D}_Y(\lambda-M,k), \\ &\cdots \tilde{D}_Y(\lambda-1,k), \tilde{D}_Y(\lambda,k)] \end{aligned} \tag{6}$$

where $M$ is adjusted so that the searched length becomes about 0.8~1.4 [s]. The last step is to obtain the noise spectrum estimate from the minimum spectrum $\tilde{D}_{min}(\lambda,k)$ as

$$\tilde{D}(k) = omin \cdot \tilde{D}_{min}(\lambda,k) \tag{7}$$

where *omin* is a compensation factor.

The MS method has the potential to track the non-stationary noise, but cannot avoid a time delay invoked. This means that the MS method is useful for only slowly time-varying noise.

# 4 Statistical Properties of Noise

Environmental noises are classified into stationary noise and non-stationary noise.

## 4.1 Stationary Noise

Figure1(a)–(d) show the case of a car noise. This noise is recognized as the stationary noise. From Figure 1, it is observed that the variance of each frequency is very small.

## 4.2 Non-Stationary Noise

Figure 2(a)–(d)show the case of a babble noise. This noise is recognized as the non-stationary noise. From Figure 2, it is observed that the spectrogram has short-term spectral peaks at several parts, while the waveform is similar with that of the stationary noise. The long-term spectrum has two main peaks at 100 Hz and 600 Hz, but both variances are very large. From these observations, we assume that the non-stationary noise consists of the addition of stationary components and non-stationary components. This assumption is a base to derive the proposed noise spectrum estimation method.

# 5 Proposed Method

For the proposed noise spectrum estimation method, the stationary noise spectrum components are estimated by the conventional method, and the non-stationary noise spectrum components are estimated by utilizing the harmonic structure the speech signal has. Both of the estimates are combined, resulting in the estimate of the noise spectrum. A block diagram of the proposed noise spectrum estimation method is shown in Figure 3.

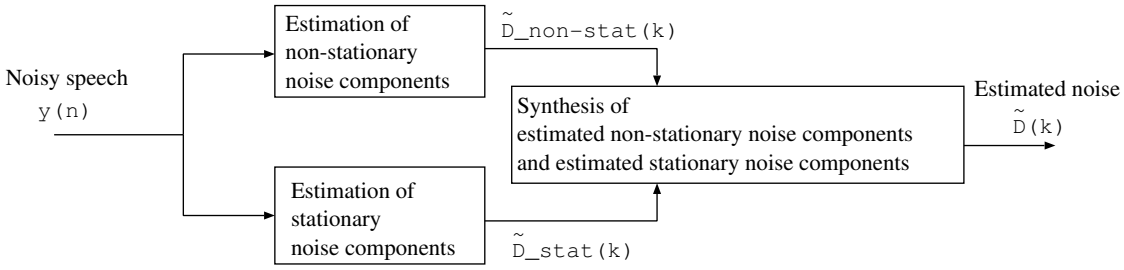



Figure 3: Block diagram of the proposed method

### 5.1 Estimation of Non-Stationary Noise Components

Once the fundamental frequency is estimated in the analysis frame, the harmonics are constructed. To achieve this, several pitch detection algorithms exist [9]. Any noise components are added to the speech components having the harmonic structure. In this case, the spectral components except for the harmonics may be of noise. Utilizing this, we can detect the noise spectrum components in the frame where the speech signal corrupted by noise exists. Figure 4 shows this as an illustration.

For the proposed noise spectrum estimation method, all spectral peaks on the noisy speech spectrum are found first. And, it is judged whether each spectral peak corresponds to the speech spectral peak or the noise spectral peak. This is conducted based on the harmonics of $F$ obtained by a fundamental frequency estimation method as

$$\begin{cases} \tilde{D}\_Peak(k) = 0, & Peak(k) \bmod F == 0 \\ \tilde{D}\_Peak(k) = 1, & \text{otherwise.} \end{cases} \tag{8}$$

However, natural speech signals do not always construct such a perfect structure of harmonics. Some variation of each harmonic component should be considered. Actually, in the proposed method a distance of $m$ is considered as

$$\begin{cases} \tilde{D}\_Peak(k) = 0, \\ \quad m < (Peak(k) \bmod F) < F - m \\ \tilde{D}\_Peak(k) = 1, & \text{otherwise.} \end{cases} \tag{9}$$

The non-stationary noise components estimate results in $\tilde{D}_{non-stat}(k) = \tilde{D}\_Peak(k)$.

### 5.2 Estimation of Stationary Noise Components

From non-speech segments, we can estimate the stationary noise components as

$$\begin{cases} \tilde{D}_{stat}(k) = |Y(\lambda,k)|^2, \\ \quad \lambda = 1 \\ \tilde{D}_{stat}(\lambda,k) = \\ \alpha \cdot \tilde{D}_{stat}(\lambda-1,k) + (1-\alpha)|Y(\lambda,k)|^2, \\ \tilde{D}_{stat}(k) = \tilde{D}_{stat}(\lambda,k) \quad \lambda \le T_I \\ \tilde{D}_{stat}(k) = \tilde{D}_{stat}(T_I,k), \\ \quad \lambda > T_I. \end{cases} \tag{10}$$

### 5.3 Synthesis of Non-Stationary Noise Components and Stationary Noise Components

The estimate of the noise spectrum is obtained as

$$\begin{cases} \tilde{D}(k) = \tilde{D}_{non-stat}(k), \\ \qquad \tilde{D}_{non-stat}(k) > 0 \\ \tilde{D}(k) = \tilde{D}_{stat}(k), \\ \qquad \tilde{D}_{non-stat}(k) = 0. \end{cases} \tag{11}$$

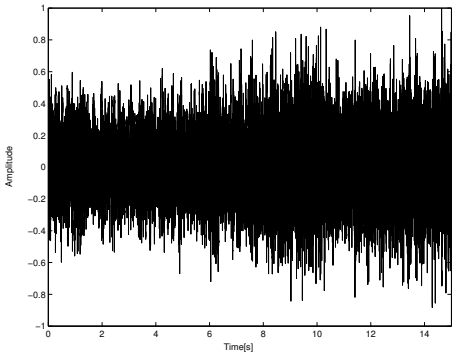
## 6 Experiments

To verify the performance of the proposed noise spectrum estimation method, the SS method was implemented by using three noise spectrum estimation types; the NSP method, the MS method and proposed method.

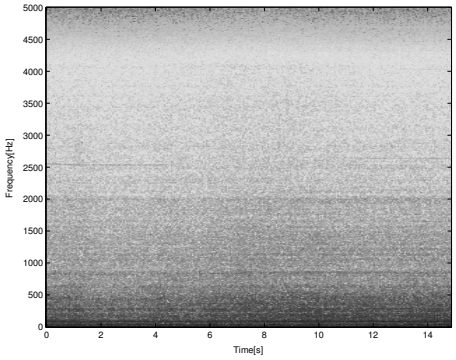
### 6.1 Speech Data and Parameters

Speech data used in the experiments are of Japanese two males and two females, which are sampled by a sampling frequency of 10 kHz with a band limitation of 3.4 kHz. Noises used are a car noise and a babble noise.

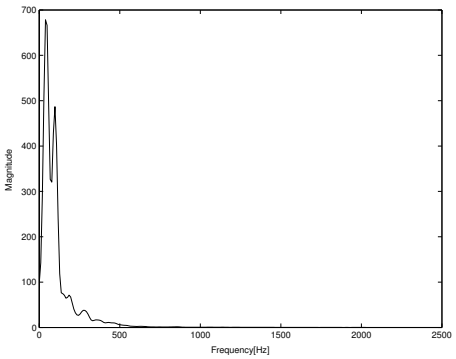
The parameters commonly used for the SS method are shown in Table 1. Those used for each noise spectrum estimation method are shown in Table 2.



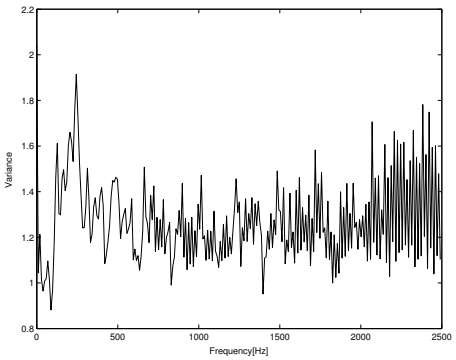

(a)




(b)

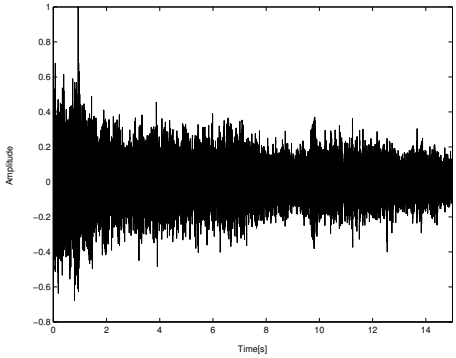



(c)

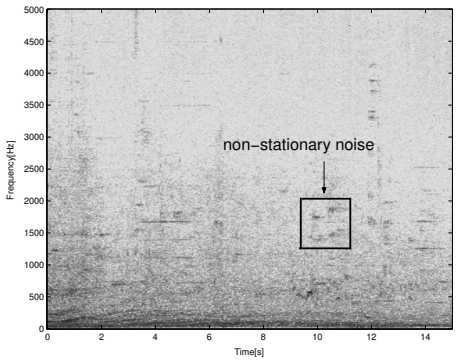



(d)

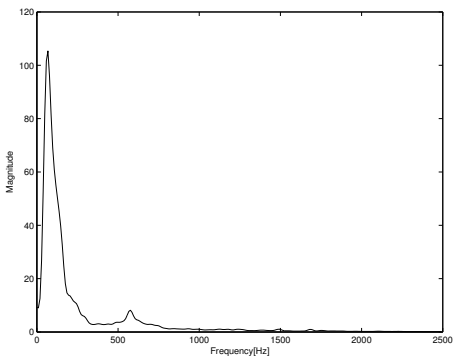
Figure 1: Stationary noise. (a) waveform, (b) spectrogram, (c) long-term spectrum, (d) variance of each frequency



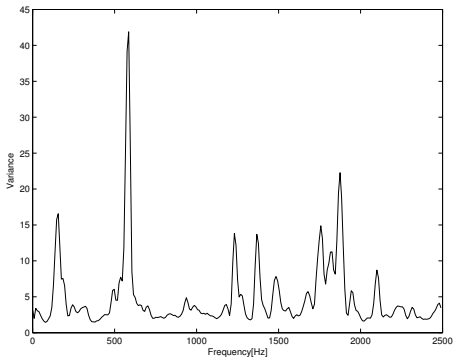

(a)




(b)




(c)




(d)

Figure 2: Non-stationary noise. (a) waveform, (b) spectrogram, (c) long-term spectrum, (d) variance of each frequency

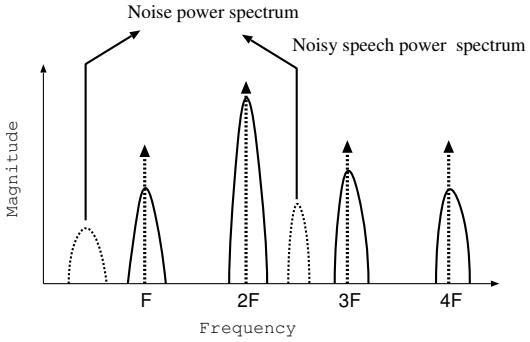



Figure 4: Estimation of non-stationary noise components

## 6.2 Evaluation

For the proposed noise spectrum estimation method, the autocorrelation method was used for pitch detection [9].

### 6.2.1 Noise Estimation Error

Noise estimation error is obtained by

$$\varepsilon(\lambda) = 10\log_{10}\frac{\sum_{k=1}^{fftn}|D(\lambda,k)| - \tilde{D}(\lambda,k)|}{\sum_{k=1}^{fftn}D(\lambda,k)} \qquad (12)$$

where $\tilde{D}(\lambda,k)$ and $D(\lambda,k)$ correspond to the estimated noise spectrum and true noise spectrum, respectively. The average of the evaluated noise estimation errors is shown in Table 3 where the proposed method (true) and the proposed method (estimate) mean to use the true fundamental frequency and its estimate, respectively.

From Table 3, we see that the proposed method provides the same level of noise estimation accuracy obtained from noise segments in the case of stationary noise. On the other hand, in the case of non-stationary noise, the proposed method is superior.

Table 1: Parameters

| Parameters in frames | |
|---|---|
| Frame length | 256 |
| Window function | Hamming window |
| Overlap | 128 |
| FFT points $fftn$ | 1024 |
| **Parameters in the SS method** | |
| Power $\gamma$ | 2 |
| Weighting $a$ | 1 |
| Spectral flower $\beta$ | 0.02 |

Table 2: Parameters for each noise estimation method

| NSP method | |
|---|---|
| Forgetting factor $\alpha$ | 0.8 |
| Last frame number $T_I$ | setting for each data |

| MS method | |
|---|---|
| Forgetting factor $\alpha$ | 0.9 |
| Number of frames $M$ | 100 |
| Compensation factor *omin* | 1.5 |

| Proposed method | |
|---|---|
| Forgetting factor $\alpha$ | 0.9 |
| Error $m$ | 4 |
| Last frame number $T_I$ | setting for each data |

Table 3: Average of $\varepsilon(\lambda)$

| Stationary noise | | |
|---|---|---|
| SNR[dB] | 10 | 5 |
| NSP method | -1.021 | -1.115 |
| MS method | 0.119 | 1.967 |
| Proposed methoditrue) | -1.012 | -1.215 |
| Proposed method (estimate) | -0.927 | -1.106 |

| Non-stationary noise | | |
|---|---|---|
| SNR [dB] | 10 | 5 |
| NSP method | -0.4534 | -0.5391 |
| MS method | 2.5203 | 0.4875 |
| Proposed method (true) | -0.9658 | -1.487 |
| Proposed method (estimate) | -0.9150 | -1.4109 |

### 6.2.2 Comparison of Noise Reduction

Segmental SNR defined as

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{1}{L}\sum_{j=0}^{L-1}10\log_{10}\left[\frac{\sum_{n=N*j}^{N*j+N-1}x(n)^2}{\sum_{n=N*j}^{N*j+N-1}[x(n)-\tilde{x}(n)]^2}\right] \qquad (13)$$

is used where $L$ is the number of frames averaged and $N$ one frame length. The improvement in segmental SNR,

$$\mathrm{SNR}_{\mathrm{seg,imp}} = \mathrm{SNR}_{\mathrm{seg,out}} - \mathrm{SNR}_{\mathrm{seg,in}}, \qquad (14)$$

Table 4: $SNR_{seg,imp}$ (stationary noise)

| SNR[dB] | 10 | 5 |
|---|---|---|
| NSP method | 3.621 | 4.767 |
| MS method | 3.133 | 3.578 |
| Proposed method (true) | 3.541 | 5.028 |
| Proposed method (estimate) | 3.502 | 4.843 |

Table 5: $SNR_{seg,imp}$ (non-stationary noise)

| SNR[dB] | 10 | 5 |
|---|---|---|
| NSP method | 1.972 | 2.242 |
| MS method | 1.302 | 1.435 |
| Proposed method (true) | 2.205 | 3.467 |
| Proposed method (estimate) | 1.959 | 3.124 |

where $SNR_{seg,out}$ and $SNR_{seg,in}$ are the output and input segmental SNRs, is summarized in Tables 4 and 5. It is observed that the proposed method provides better performance, in particular in the case of non-stationary noise.

### 6.3 Listening Test

Six listeners listened four times random-ordered noisy speech processed by each method and scored four times. The score levels are as follows: $5 \cdots$ easy to listen, $4 \cdots$ slightly easy to listen, $3 \cdots$ usual, $2 \cdots$ slightly difficult to listen, and $1 \cdots$ difficult to listen. Tables 6 and 7 show the listening test results. Again, the superiority of the proposed method is observed.

Table 6: Average of listening test scores (stationary noise case)

| SNR[dB] | 10 | 5 |
|---|---|---|
| NSP method | 3.79 | 2.67 |
| MS method | 3.75 | 2.54 |
| Proposed method (true) | 3.92 | 2.75 |
| Proposed method (estimate) | 3.96 | 2.67 |

Table 7: Average of listening test scores (non-stationary noise case)

| SNR[dB] | 10 | 5 |
|---|---|---|
| NSP method | 3.08 | 2.38 |
| MS method | 3.17 | 2.46 |
| Proposed method (true) | 3.42 | 2.92 |
| Proposed method (estimate) | 3.38 | 2.83 |

## 7 Conclusions

In this paper, a new non-stationary noise spectrum estimation method has been presented and used with the spectral subtraction method. Some experiments have suggested that the proposed speech enhancement method works very effectively in non-stationary noise environments.

*References:*


[1] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," IEEE Trans. Signal Processing, vol. 39, no. 4, April 1991.

[2] Z. Goh, K. Tan and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," IEEE Trans. Speech Audio Processing, vol. 7, no. 5, pp. 510–524, September 1999.

[3] A. de Cheveigne, "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am., vol. 93, no. 6, pp. 3271–3290, June 1993.

[4] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," IEEE Trans. Speech and Audio Processing, vol. 7, pp. 126–137, March 1999.

[5] R. Martin, "Spectral subtraction based on minimum statistics," Proc. EUSIPCO'94, pp.1182-1185, Sept. 1994.

[6] Y. Ephraim and D. Malah, "Speech Enhancement using a minimum mean-square error log-spectral amplitude estimator", IEEE Trans. Acoustics, Speech and Signal Process., vol.ASSP-32, no.6, pp.1109-1121, Dec. 1984.

[7] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," IEEE Trans. Speech Audio Processing, vol. 6, no. 5, pp. 445–455, September 1998.

[8] TIA/EIA/IS-127, "Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital system", January 1997.

[9] L.R.Rabiner, M.J.Cheng, A.E.Rosenberg, C.A.Mcgonegal "A comparative performance study of several pitch detection algorithms," IEEE Trans. Acoustics, Speech and Signal Processing, Vol.ASSP-24, No.5, October 1976.