MULTI-STAGE SPECTRAL SUBTRACTION FOR ENHANCEMENT OF AUDIO SIGNALS

Masatsugu Okazaki †‡, Toshifumi Kunimoto†, Takao Kobayashi‡

†ProAudio & Digital Musical Instruments Division, YAMAHA Corporation, Hamamatsu, 430-8650 Japan
‡Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, 226-8502 Japan
E-mail: †{okazaki,kunimoto}@emi.yamaha.co.jp, ‡takao.kobayashi@ip.titech.ac.jp

ABSTRACT

This paper describes a technique based on spectral subtraction (SS) for enhancing audio signals which have nonstationary properties. In the technique, audio signals are modeled as a sum of sinusoidal components, which is called tones, of various frequencies and durations. We show that the analysis window length for SS should be adjusted in accordance with the tone duration to obtain higher performance on SS. Then we propose a multi-stage SS approach into which an appropriate window length selection mechanism is incorporated. We also propose a technique for suppressing an artifact called pre-echo which appears in the noise reduction process. It is shown from the result of a subjective evaluation test that the proposed technique gives superior performance to the conventional SS for enhancement of audio signals including speech, piano, drums, and orchestra.

1. INTRODUCTION

Recently, spectral subtraction (SS) technique and its derivatives have been widely used in the areas of speech enhancement and audio restoration. In fact, the SS technique works very well in most cases of music recording or speech recognition systems for the purpose of reducing additive noise from audio signals. Moreover the SS algorithm is simple and therefore suitable for real time implementation with fewer computational costs.

The basic idea of the SS technique is to estimate the spectrum of the original signal by subtracting the noise spectrum from that of noisy signal under the assumption that both the original signal and noise are stationary or considered to be stationary on a short-time basis. This procedure can be viewed as the filtering with a noise reduction function that maps noisy spectra into clean spectra [1][2]. For the purpose of improving the objective and subjective quality of the enhanced signals, there have been proposed a number of noise reduction functions, such as ML estimation [3], MMSE estimation [4], and MMSE estimation of log magnitude spectrum [5]. It has been also shown that some improvements can be obtained by applying noise reduction in the transformed domain using the discrete cosine transform (DCT) or the Karhunen-Loeve transform (KLT) instead of using the DFT [6][7].

One of the limitations of the original SS approach is that the target signal should be stationary. However, in audio signals, there exists a wide variety of signals which do not have stationary properties even on a short-time basis. For example, the signals such as piano sounds at the attack, impulsive sounds of drums, and plosives or beginning parts of utterances of speech sounds are obviously nonstationary. This would cause severe degradation on the quality of the restored signals after hiss reduction process in the audio recordings.

In this paper, to overcome this problem, we present a new technique for the additive noise reduction based on the SS approach in which the nonstationary signal is taken into account. We model the nonstationary signals as a sum of the sinusoidal components, which we will call the "tone signals" or simply "tones," of various frequencies and durations. This modeling is reasonable for most of the audio signals, especially music signals, because they are usually composed of a number of line spectral components. We examine the relationship between the noise reduction performance in the frequency domain and analysis window length for the case of the single tone signal having relatively short duration. As a result, we show that it is important to adjust the analysis window length in accordance with the tone duration for improving the noise reduction performance. Then we propose a multi-stage SS technique into which window length selection is incorporated for the mixture of tones with various durations. We also propose a technique for suppressing an artifact appearing in the noise reduction process called "pre-echo," which is similar to a well-known phenomenon in the transformed coding of the audio signals.

2. SINGLE-STAGE SPECTRAL SUBTRACTION

Consider an audio signal $s(t)$ which has been degraded by uncorrelated additive noise $d(t)$. The problem is restoring the original audio signal from the degraded signal $y(t)$. If the restoration process is done on a frame-by-frame basis in the discrete-time domain, the observed noisy data can be expressed by

$$y_i(n) = s_i(n) + d_i(n) \tag{1}$$

where the subscript i denotes the frame number of the windowed signal. Taking the discrete Fourier transform (DFT) of (1) gives

$$Y_i(k) = S_i(k) + D_i(k) \tag{2}$$

where k is the frequency bin number and $Y_i(k)$, $S_i(k)$, and $D_i(k)$ denote the DFTs of $y_i(k)$, $s_i(k)$, and $d_i(k)$, respectively.

The spectral subtraction (SS) estimate [1] of the audio signal is given by

$$\hat{S}_i(k)=\begin{cases} [|Y_i(k)| - \mu(k)] \cdot e^{j\angle Y_i(k)}, & |Y_i(k)| > \mu(k) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $\angle Y_i(k)$ is the phase of the noisy signal spectrum and $\mu(k)$ is an estimate of $|D_i(k)|$ taken from sections where the audio signal does not exist. In [1], it is suggested to replace $|Y_i(k)|$ with its average during several frames for the purpose of reducing the estimation error.

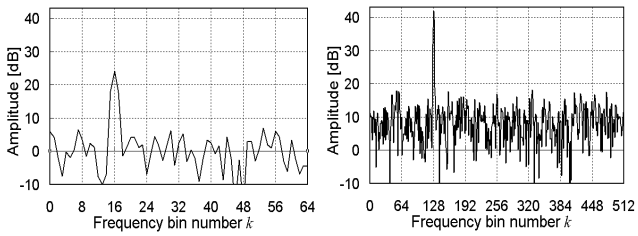

Fig. 1. Example of DFT spectrum for a tone corrupted by additive noise. (a) 128-point DFT. (b) 1024-point DFT.

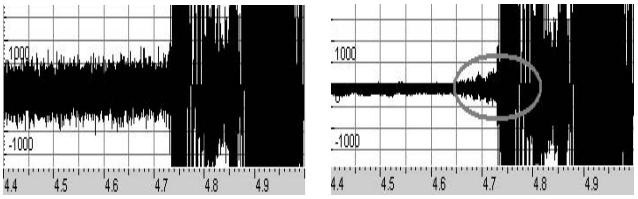

Fig. 2. Example of pre-echo caused by SS with using a long window. (a) Input signal. (b) After SS with 8192-point window.

The subtraction estimator of (3) can be expressed in a generalized form as

$$\hat{S}_i(k)=\begin{cases} [|Y_i(k)|^\alpha - \beta\mu^\alpha(k)]^{\frac{1}{\alpha}} \cdot e^{j\angle Y_i(k)}, & |Y_i(k)| > \mu(k) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where α and β are the control parameters of the SS technique. For the special case of $\alpha = 2$, it is known to be the power subtraction method.

3. TONE EXTRACTION FROM DEGRADED AUDIO SIGNALS

3.1. Extraction of sinusoidal components from noise

Since audio signals, especially music signals, are usually composed of a number of sinusoidal components which correspond to fundamental frequency component and partial tones, enhancement of audio signals can be viewed as the extraction of the sinusoidal components from background noise. Here we will refer to each sinusoidal component as the tone signal or simply tone.

In the frequency domain, the spectrum of a tone signal is given by a line-like shape. This means that its energy concentrates into a relatively few frequency bins. In contrast the spectrum of a white noise spreads over all frequency bins. Therefore, the spectral peak which corresponds to the tone signal tends to be much higher than the average value of the noise spectrum. This is illustrated by Fig. 1, where DFT spectrum for the tone signal of frequency $\omega = 2\pi/8$ corrupted by a zero-mean white noise. It is seen from the figure that the ratio of the peak amplitude of the tone spectrum to the noise spectrum increases as increasing the analysis window length. It is shown that this ratio is directly proportional to the square root of the analysis window length [8]. In other words, with using a twice-longer analysis window, we can obtain the same noise reduction performance as the case when the SNR of the input signal is 3dB higher. Another advantage of using a longer analysis window is that the frequency resolution increases with the window length. It is easy to find two tones whose pitches are close, respectively, not as beaten single tone.

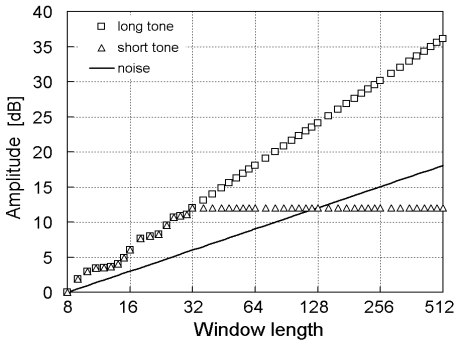

Fig. 3. DFT amplitude for stationary tone, short tone, and white noise.

Although the long analysis window gives higher noise reduction performance for stationary tone signals, some disadvantages come up when the signal has nonstationary properties. For example, if the signals such as piano sounds at the attack and the beginning parts of the speech utterances are processed based on the SS technique with a long window, an artifact might appear before the tones contained in the original signal begin. This is illustrated by Fig. 2, where a noisy speech signal sampled at 44.1kHz is enhanced by the SS technique with a window length of 8192 samples. Since this artifact is similar to the pre-echo which is well-known in the adaptive transform coding (ATC) of the audio signals, we will also refer to it as the pre-echo.

3.2. Optimum window length for short tone extraction

We model the audio signals which has nonstationary properties as a sum of tone signals of various frequencies and durations. For such signals, most of the tone signals might be shorter than the analysis window length. Hence we examine here the relationship between the amplitude of tone's DFT spectrum and the window length when the tone duration is shorter than the analysis window length.

Fig. 3 shows a curve of the DFT amplitude at frequency $\omega = 2\pi/8$ for a short tone given by $s(n) = \sin(2\pi n/8)(u(n) - u(n - 32))$, where $u(n)$ is the unit step, with changing the DFT size, i.e., rectangular window length. In the figure, the values are normalized in such a way that the value for DFT size of 8 is 0dB. The curves for a stationary tone, i.e., its duration is longer than the DFT size, of the same frequency and the expected value of the amplitude of the white noise spectrum are also shown.

It can be observed that the amplitude corresponding to the tone increases with the window length when the window length is shorter than the tone duration. However the amplitude does not change when the window length exceeds the tone duration. Since the amplitude of the noise spectrum increases with the window length, the ratio of the amplitude of the tone spectrum to the noise spectrum takes its maximum value when the window length is the same as the tone duration. Thus an optimum window length equals to the duration of the short tone. In addition, if the window length is twice longer or half shorter than the tone duration, it is equivalent to the case that the SNR of the input signal is 3dB lower.

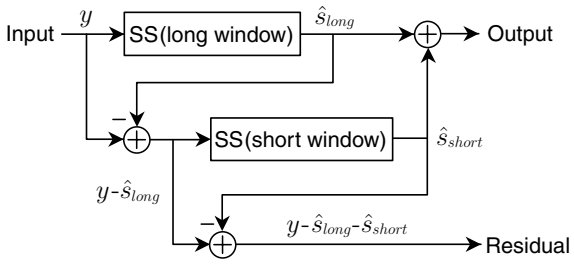

Fig. 4. Schematic block diagram of two-stage spectral subtraction.

4. MULTI-STAGE SPECTRAL SUBTRACTION

4.1. SS with using multiple windows

As mentioned in the previous section **3**, to improve the enhancement performance, it is desirable to choose the window length as long as possible for long tones and as the same as duration for short tones. However, the audio signal having nonstationary properties consists of tones of various durations, and therefore the SS with a fixed window length would not work well for such signals.

To resolve this problem, we adopt an approach of using multiple analysis windows in the SS technique and will refer to it as the multi-stage SS (MSSS). Fig. 4 shows a schematic block diagram of two-stage SS. At the first stage, the SS with using a relatively long analysis window is performed. Since only stationary tones, i.e., relatively long tones, can be extracted from noise when using the long window, the enhanced signal $\hat{s}_{long}$ mainly consists of long tones, whereas the residual signal $y - \hat{s}_{long}$ contains the noise and short tones. Then at the second stage, the SS with a shorter analysis window is applied to the residual signal of the first stage. The enhanced signal $\hat{s}_{short}$ obtained at this stage consists of shorter tones which are not extracted at the first stage. Another stage with a much shorter analysis window can be cascaded if required. Finally the enhanced signals from all stages are added and the resultant signal becomes the output of the multi-stage SS.

4.2. Extraction of long tones

Since the first stage of the MSSS uses a long analysis window, the resultant enhanced signal may suffer from the pre-echo. To suppress the pre-echo and to separate long tones from shorter tones and noise, we use the following spectral subtraction estimate at frame i:

$$|\hat{S}_i^{long}(k)| = \min\Big(|\hat{S}_{i-L+1}(k)|, \ldots, |\hat{S}_i(k)|, \ldots, |\hat{S}_{i+L-1}(k)|\Big) \tag{5}$$

where $|\hat{S}_i(k)|$ is the SS estimate given by (4) at frame i, and L is an integer which equals to the quotient of the window length divided by the frame shift. This procedure is similar to that described in [1].

Fig. 5 shows an example of the enhanced process of the two-stage SS. The input signal is a piano sound degraded by additive white noise shown in Fig. 5(a). Fig. 5(c) and (d) are the enhanced signal $\hat{s}_{long}$ and the residual $y - \hat{s}_{long}$, respectively, obtained using (4) with $\alpha = 3$ and $\beta = 1.25$. Other experimental conditions are as follows: 44.1 kHz sampling, 16-bit quantization, 8192-point Hanning window, and $L = 8$. This results correspond to those of the conventional single-stage SS with using a long window. Similarly Fig. 5(e) and (f) are the results of using (5). It can be seen

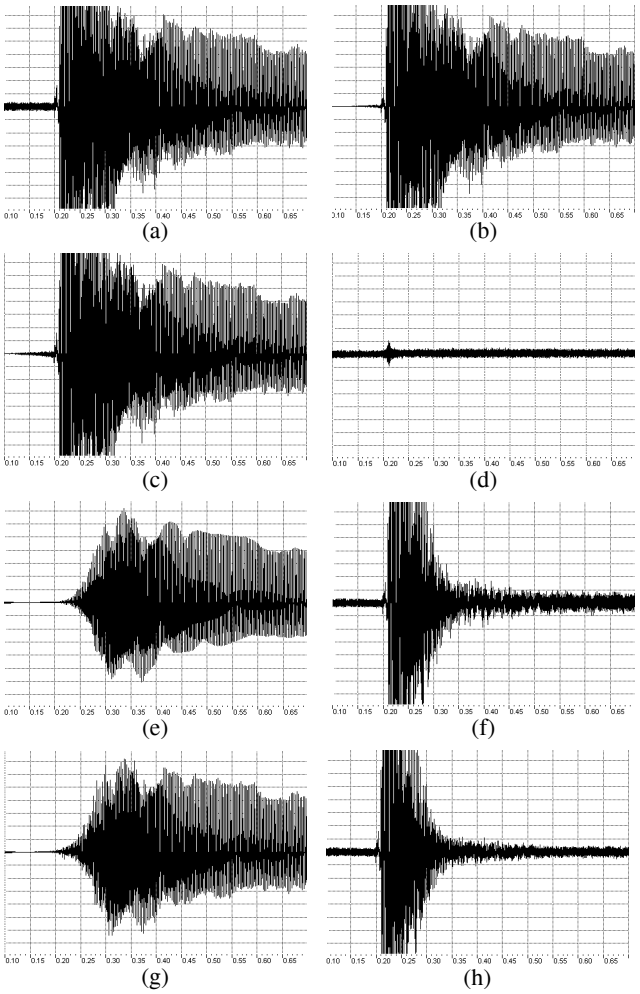

Fig. 5. Example of enhanced and residual signals from two-stage SS for a piano sound. (a) Input signal. (b) Output signal. (c) Output from single-stage SS. (d) Residual signal for (c). (e) $\hat{s}_{long}$ using (5). (f) Residual $y - \hat{s}_{long}$ for (e). (g) $\hat{s}_{long}$ using (6). (h) Residual $y - \hat{s}_{long}$ for (g).

that the pre-echo is suppressed by using (5). However it can also be seen that some long tones exist in the residual signal as shown in Fig. 5(f). This is due to the fact that the signal is underestimated when using (5).

To reduce this effect, we use an alternative estimate given by

$$|\hat{S}_i^{long}(k)| = \min\left(\frac{|\hat{S}_{i-L+1}(k)|}{A_{-L+1}}, \ldots, \frac{|\hat{S}_i(k)|}{A_0}, \ldots, \frac{|\hat{S}_{i+L-1}(k)|}{A_{L-1}}\right) \tag{6}$$

where $\{A_l\}$, $-L \le l \le L$ are the weights that control the SS estimate. Fig. 5(g) and (h) show the results of using (6). In this example, we set the weights as $\{0.02, 0.1, 0.22, 0.35, 0.49, 0.6, 0.68, 1.0, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1\}$. From the figure, it is found that the long tones in the residual are decreased. Finally we have the output shown in Fig. 5(b)

It should be noted that if the analysis window length is short enough to prevent the pre-echo, the modification of the SS estimate of (6) is not required.

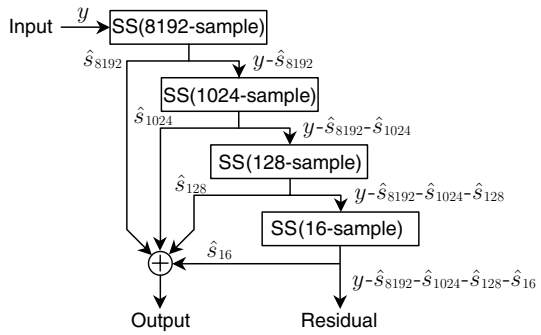

Fig. 6. A four-stage spectral subtraction system.

5. EXPERIMENTS

5.1. Experimental conditions

We implemented a four-stage SS system for enhancement of audio signals as shown in Fig. 6. We set window length as 8192, 1024, 128, and 16 for the first, second, third, and fourth stages, respectively. We used Hanning window as the analysis/synthesis window. Frame shift was one eighth of the window length. From the result of preliminary experiments, we set $\alpha = 3$ for all stages and $\beta = 1.37$ except for the fourth stage where $\beta = 1.94$ was used. The modified estimate of (6) was used except the last stage, and the control weights $\{A_l\}$ were set as the same values as described in **4.2**. The audio signal was sampled at 44.1 kHz and quantized linearly with 16-bit precision. Then a white Gaussian noise with the variance of 240^2 was added to the original signal.

5.2. Results

To evaluate the performance of the MSSS, we conducted subjective and objective evaluation tests. We used speech, piano, drums, and orchestral sounds as the original signals. For comparison, single-stage SS with the fixed window length is also applied to the degraded signals.

Table 1 shows the subjective performance measured by the SNR. It is found that longer analysis window gives higher SNR value for music signals when the single-stage SS is applied. The MSSS gives comparable or slightly lower values than the single-stage SS with a window length of 8192 samples.

Table 2 shows the result of a degradation category rating (DCR) test. In Table 2, the value of each entry represents the degradation mean opinion score (DMOS). In the DCR test, ten listeners were presented with the original signal as a reference before they listen the enhanced signal. The task for the listener is to rate the degradation perceived when comparing the enhanced signal to the reference. A 5-point scale was used, that is, 5: degradation not perceived, 4: perceived but not annoying, 3: slightly annoying, 2: degradation annoying, and 1: very annoying. It can be shown from the table that the MSSS provides the highest DMOS scores for all input signals. This is due to the fact that much difference between MSSS and 8192-sample SS are not perceived about the residual noise level, whereas degradation caused by pre-echo is obviously perceived for the 8192-sample SS.

Table 1. Objective performance measured by SNR in dB.

	MSSS 4-stage	Single SS 8192	Single SS 1024	Single SS 128	Input
Speech	22.3	23.6	**24.0**	11.3	21.0
Piano	**25.1**	24.5	24.4	11.0	19.4
Drums	18.0	**19.9**	16.8	5.6	12.5
Orchestral	21.7	**23.2**	21.4	10.6	16.4
Average	21.8	**22.8**	21.7	9.6	17.4

Table 2. Subjective performance measured by DMOS.

	MSSS 4-stage	Single SS 8192	Single SS 1024	Single SS 128
Speech	**3.6**	3.3	2.9	1.7
Piano	**3.8**	3.5	2.4	1.5
Drums	**4.0**	**4.0**	2.9	1.3
Orchestral	**3.4**	3.1	1.8	1.6
Average	**3.7**	3.5	2.5	1.5

6. CONCLUSION

In this paper, we described a new technique based on SS for enhancing audio signals which has nonstationary properties. We have shown that it is important to adjust the analysis window length in accordance with the tone duration for improving the noise reduction performance. Then we proposed a MSSS technique into which window length selection is incorporated for the mixture of tones with various durations. We also proposed a technique for suppressing the pre-echo. We have shown that the proposed technique gives superior performance to the conventional SS. Our future work is evaluating the MSSS performance using various types of audio signals.

7. REFERENCES

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[3] R.J. McAulay and M.L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[6] I.Y. Soon, S.N. Koh, and C.K. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech Communication*, vol. 24, pp. 249–257, 1998.

[7] Y. Ephraim and H.L. Van-Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[8] A.V. Oppenheim and R.W. Schafer, *Discrete-Time Signal Processing*, chapter 11, Prentice-Hall, 2nd edition, 1999.