A SPECTRAL SUBTRACTION USING PHONEMIC AND AUDITORY PROPERTIES*

Sunmee Kang* · Wooil Kim** · Hanseok Ko****

ABSTRACT

This paper proposes a speech state-dependent spectral subtraction method to regulate the blind spectral subtraction for improved enhancement. In the proposed method, a modified subtraction rule is applied over the speech selectively contingent to the speech state being voiced or unvoiced, in an effort to incorporate the acoustic characteristics of phonemes. In particular, the objective of the proposed method is to remedy the subtraction induced signal distortion attained by two state-dependent procedures, spectrum sharpening and minimum spectral bound. In order to remove the residual noise, the proposed method employs a procedure utilizing the masking effect. Proposed spectral subtraction including state-dependent subtraction and residual noise reduction using the masking threshold shows effectiveness in compensation of spectral distortion in the unvoiced region and residual noise reduction.

Keywords: spectral subtraction, phonemic, auditory properties, mask effect, noise reduction

1. INTRODUCTION

In common spectral subtraction, the speech spectrum is estimated by subtracting the average of spectral components in the non-speech activity interval from the noisy speech spectrum[1]. Such a procedure is equally applied to the spectral components over the entire speech signal. But it is known that

---

* The authors gratefully acknowledge the support by the Ministry of Information and Communication under the University Basic Research Grants Program.

* Dept. of Electronic Engineering, Korea Univ.

** Dept. of Computer Science, Seokyeong Univ.

noise does not impact on all phonemes equally. Because of each phoneme's distinct acoustic property, the degree of influence by noise must be different for each. Therefore, blindly applying the same enhancement criterion over the entire speech is ineffective and ends in undesirable results. In this paper, a speech state-dependent spectral subtraction method is proposed to regulate the blind spectral subtraction for improved enhancement.

In the proposed method, a modified subtraction rule is applied over the speech selectively contingent to the speech state being voiced or unvoiced, in an effort to incorporate the acoustic characteristics of phonemes. Voiced sounds have dominant spectral components in a specific frequency band, called resonance frequency, due to its production mechanism. And these resonance frequency components are typically characterized by high energy levels. Vowels, for example, have dominant frequency components called formants in their spectral domain. On the other hand, unvoiced sounds' spectral distribution looks very similar to that of white noise. Therefore, in the spectrum of unvoiced sounds contaminated with noise, the original spectral components are obscured and subsequent noise subtraction results in severe signal distortions. To remedy the subtraction induced signal distortion, two types of state-dependent procedure (spectrum sharpening and minimum spectral bound) are employed into the spectral subtraction implementation contiguously.

The enhanced speech signal by spectral subtraction still contains the residual noise, called "musical noise", which makes a metal-like sound and annoys the human ear. In order to remove the residual noise, the proposed method employs a procedure utilizing the masking effect. The masking effect is a principal property of the human auditory system such that a band tone can be masked by an other band located close by in the frequency domain.

This paper is organized as follows. In Section 2, an overview of spectral subtraction is presented and its problem is formulated. The proposed enhancement schemes and their detailed procedures are described in Section 3. In Section 4, the representative experimental results are provided. Finally, Section 5 contains concluding remarks as well as suggestions for future work.

2. PROBLEM FORMULATION

Spectral subtraction is a method for restoration of the power or the magnitude spectrum of a signal observed in additive noise, through subtraction

of an estimate of the average noise spectrum from the noisy signal spectrum[1][2]. Spectral information required to describe the noise spectrum is obtained from signal measured during non-speech activity.

Assume that a windowed noise signal n(k) has been added to a windowed speech signal s(k), with their sum denoted by y(k). Then

$$y(k) = s(k) + n(k) \tag{2.1}$$

Taking the Fourier transform gives

$$Y(e^{j\omega}) = S(e^{j\omega}) + N(e^{j\omega}) \tag{2.2}$$

The general spectral subtraction is defined by the following equation

$$\begin{aligned} |S(e^{j\omega})|^b &= |Y(e^{j\omega})|^b - \alpha|\mu(e^{j\omega})|^b, \\ \text{if} \quad |Y(e^{j\omega})|^b &> \alpha|\mu(e^{j\omega})|^b \\ &= \beta|Y(e^{j\omega})|^b \qquad \textit{otherwise} \end{aligned} \tag{2.3}$$

where $\mu(e^{j\omega})$ is average noise value during non-speech activity. Note that $0 \leq \alpha \leq 1$ and β is 0 or a very small value.

The estimation by spectral subtraction described above is applied to the entire speech interval with equal criterion. This is a weakness since it is well known that noise has non-uniform impact across the phoneme sequence of a speech utterance. That is, the classical spectral subtraction fails to fully utilize the phonemic information carried within the signal for enhancement processing[3]. The phonemic information includes the acoustic properties which are originated in the different production process of each phoneme. Therefore, when the speech signal is contaminated by noise, each phoneme in the speech can show various degrees of distortion on the time and spectral domains. The vowels, typical examples of voiced sound, have the resonance frequency components, called formants, whose energies are relatively high[4]. Therefore the noise spectrum has less influence on formants than on other frequency components. On the contrary, unvoiced sounds have spectral distribution similar to that of white noise. Consequently, unvoiced sounds are contaminated severely

and estimated signals by spectral subtraction shows severe spectral distortion compared to voiced sounds.

3. PROPOSED METHODOLOGY

The two-stage speech enhancement method proposed for improved performance is ;.

1) Noise subtraction dependent on speech state
2) Residual noise reduction using masking effect

3.1 State-Dependent Subtraction

In this section, two types of state-dependent noise subtraction methods are introduced to avoid the spectral distortion in unvoiced sounds.

3.1.1 Spectrum sharpening processing

As a new approach for reducing spectral obscurity of unvoiced sounds corrupted by noise, spectrum sharpening is proposed as a kind of preprocessing. That is, by making high-level energy spectral components higher and low-level ones lower, the shape of the noisy spectrum can be more pronounced. For spectrum sharpening, a simple function is introduced and it is expressed as following.

$$SS(i)=\begin{cases} -0.5 & -L_{SS}/2 \le i \le -L_{SS}/6 \\ 1 & -L_{SS}/6 \le i \le L_{SS}/6 \\ -0.5 & L_{SS}/6 \le i \le L_{SS}/2 \\ 0 & otherwise \end{cases} \qquad (3.1)$$

where L_{SS} is sharpening function size.

Spectrum sharpening is accomplished by convoluting the noisy spectrum $Y(i)$ .with the suggested sharpening function $SS(i)$.

$$Y_S' = Y(i) * SS(i) = \sum_j Y(i) SS(i-j) \qquad (3.2)$$

$$Y_S(i)=\begin{cases} Y_S'(i-L_{SS}/2) & \text{if } Y_S'(i-L_{ss}/2)\geq 0 \\ Y(i) & otherwise \end{cases} \tag{3.3}$$

The function for spectrum $SS(i)$ has the effect that noisy spectral shape becomes clear to some extent through summing close-adjacent frequency components and suppressing far-component spectrum

3.1.2 Minimum spectral bound

As described in Section 2, in the general spectral subtraction when a noisy spectral component is smaller than the average noise spectrum, the noise subtraction result is set as a small value proportional to noisy spectrum or zero to avoid becoming minus value. In the case of the unvoiced sound, its spectral similarity with white noise makes such processing occur frequently. The spectrum set as that possibly have significant difference from original clean spectrum, so that restored signal contains serious spectral distortion. As the SNR decreases, such a situation becomes more pronounced.

In the proposed approach, spectral minimum bound is computed using power spectrum estimated based on the linear predictive model (all-pole model) from noise subtracted spectrum. Since estimated power spectrum has a similar envelope to the original spectrum, it can provide information as spectral bound mentioned above. The minimum spectral bound is calculated as follows.

1) The time domain signal x is estimated through the inverse FFT of combination noise-subtracted spectrum and the phase of noisy signal.

2) The excitation gain is calculated using the Parseval's theorem.

$$\frac{1}{N}\sum_{f=0}^{N-1}\frac{g^2}{\left|1-\sum_{k=1}^{P}a_k e^{-j\pi fk/N}\right|^2}=\sum_{m=0}^{N-1}x^2(m) \tag{3.4}$$

where a_k is linear predictive coefficients attained from samples of x in a frame and P is linear predictive order.

3) Calculate an estimate of the power spectrum of speech model.

$$P_{XX}(\omega) = \frac{g^2}{|1 - \sum_{k=1}^{P} a_k e^{-j2\pi jk/N}|^2} \tag{3.5}$$

4) Scaling at each critical band for sufficient value as the minimum bound.

$$MB(\omega) = P_{XX}(\omega) \frac{\sum_{\omega=bl_i}^{bh_i} P_{YY}(\omega)}{\sum_{\omega=bl_i}^{bh_i} P_{YY}(\omega) + \sum_{\omega=bl_i}^{bh_i} |\mu(\omega)|^2} \tag{3.6}$$

where bl_i is the lower boundary of critical band i, bh_i is the upper boundary of critical band, $P_{YY}(\omega)$ is noisy power spectrum and $\mu(\omega)$ is average noise spectrum.

Computed minimum bound $MB(\omega)$ is set as noise subtracted spectrum when the noisy spectrum is smaller than noise spectrum.

3.2 Residual Noise Reduction based on Masking Effect

The enhanced speech signal by the proposed spectral subtraction still bears the residual noise, called "musical noise", which makes a metal-like sound and annoys the human ear. In order to remove the residual noise, the proposed method exploits a process based on masking effects. The masking effect is a principal property of the human auditory system. When tones are produced simultaneously, masking occurs in which louder tones can completely obscure softer tones. In other words, the physical presence of sound certainly does not ensure audibility and conversely can ensure inaudibility of other sound[5][6][7].

Residual noise is possibly minimized by setting the negative subtracted spectral components to a masking threshold. The computation of the masking threshold introduced in [8] is used. It is composed of following steps : 1) critical band analysis of the signal, 2) applying the spreading function to the critical band spectrum, 3) calculating the spread masking threshold, 4) accounting for the absolute threshold, and 5) re-normalization.

3.3 Integrated Enhancement Scheme

The block diagram in Figure 1 presents totally integrated speech enhancement algorithm including both the state-dependent noise subtraction and

residual noise reduction using the masking threshold. Detection of speech state, i.e., voiced or unvoiced, is made by means of energy and ZCR (Zero Crossing Rate). However because reliability of decision by ZCR is dependent on noise amount, more study is required to find a more reliable decision rule for the voiced/unvoiced state.

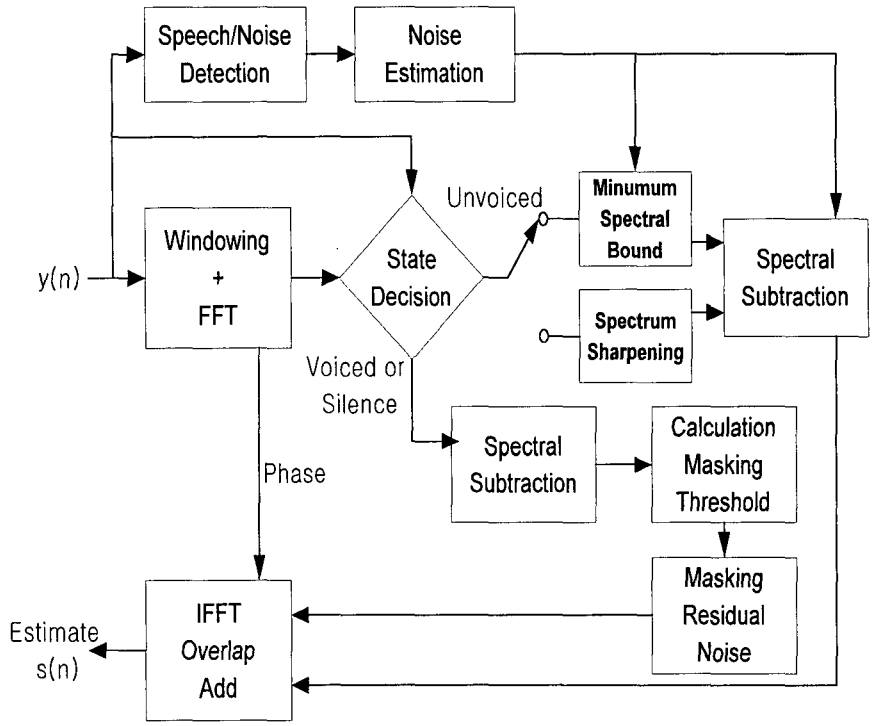

Figure 1. Block diagram of proposed spectral subtraction

4. EXPERIMENTAL RESULTS

For the performance of the proposed algorithm, experiments of 4 subtractive types of algorithm have been conducted as follows.

1) Power spectral subtraction without residual noise reduction (SPSUB)

2) Power spectral subtraction with residual noise reduction proposed by S. F. Boll (SPSUB+RNR)

3) State-dependent power spectral subtraction employing spectrum sharpening processing with residual noise reduction using masking threshold (ST(1)-SPSUB+MSKRNR)

4) State-dependent power spectral subtraction employing minimum spectral bound with residual noise reduction using the masking threshold (ST(2)-SPSUB+MSKRNR)

The original clean speech data is an utterance "computer" which is recorded by a twenty five year-old man in an anechoic chamber and sampled at 16 kHz. The noise is artificially generated and white Gaussian. In the implementation of spectral subtraction algorithm, the size of an analysis frame is 16 msec (256 points) and an overlap lag is 8 msec. At each frame, Hamming windowed signal is analyzed in 256-point FFT.

Figure 2 (a) shows noisy speech and the region selected as unvoiced state and Figure 2 (b), and (c) show enhanced speech by ST(1)-SPSUB+MSKRNR and ST(2)-SPSUB+MSKRNR. Figure 3 (a), (b) and (c) present the spectrogram of noisy speech and enhanced speech by two methods respectively. From figures, we can see the residual noise as well as the background noise are considerably reduced by the proposed schemes. Listening test also shows an improvement in intelligibility of the enhanced speech. The performance comparison in terms of input-output SNR shows that the proposed spectral subtraction methods are better than SPSUB+RNR by 0.98 dB on the average in the low SNR (<0 dB), but their results degrade as SNR increases (Table 1). It is thought that such behavior in performance degradation as SNR increases occurs since the residual noise is reduced not by subtraction but setting the values below threshold in ST(1)/ST(2)-SPSUB+MSKRNR.

Table 1. Performance comparison in terms of input-output SNR in the integrated implementation.

SNR(dB)	-15.71	-9.93	-5.67	0.17	6.25	10.13
SPSUB	2.39	3.09	6.87	11.63	18.76	23.54
SPSUB+RNR	2.51	3.25	7.15	**11.93**	**19.12**	**24.03**
ST(1)-SPSUB +MSKRNR	**4.21**	**4.05**	**7.53**	11.74	18.55	23.02
ST(2)-SPSUB +MSKRNR	2.95	3.38	7.25	11.75	18.67	23.18

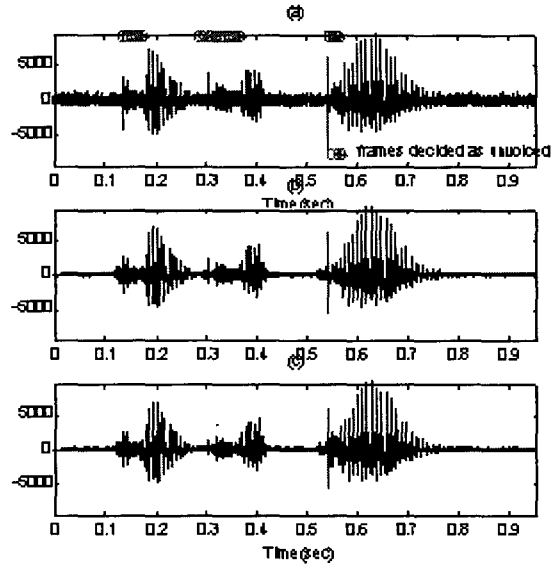

Figure 2. (a)Noisy speech (10 dB) and unvoiced decided regions. (b)Enhanced seech by ST(1)-SPSUB+MSKRNR. (c)Enhanced spech by ST(2)-SPSUB+MSKRNR.

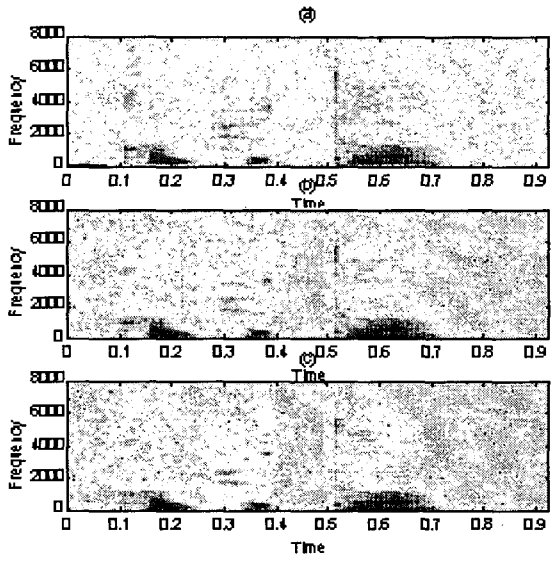

Figure 3. (a)Spectrogram of noisy speech (10dB). (b)Spectrogram of enhanced speech by ST(1)-SPSUB+MSKRNR. (c)Spectrogram of enhanced speech by ST(2)-SPSUB+MSKRNR

5. CONCLUSION

This paper proposed a spectral subtraction algorithm based on phonemic properties and masking effect. That is, an experimental trial for speech enhancement modeling speech production and perception mechanism of the human auditory system has been conducted. The proposed spectral subtraction indicates a similar performance to those of the classical spectral subtraction methods in terms of the SNR. However, in the enhanced speech by the proposed scheme, the unvoiced sound region is shown to display relatively less signal distortions. A continuing investigation for further performance improvement is being pursued in the areas of developing a more reliable state decision algorithm, utilizing various phonemic classes (stops, silences, etc..).

REFERENCES

[1] S. F. Boll. 1979. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction." IEEE Trans. on ASSP, Vol. ASSP-27, No.2, 113-120.

[2] S. V. Vaseghi. 1996. Advanced Signal Processing and Digital Noise Reduction, Wiley & Teubner, New York.

[3] B. L. Pellon and J. H. L. Hansen. 1996. Text-directed speech enhancement using phoneme classification and feature constrained vector quatization, Proc. of IEEE ICASSP96, 645-648.

[4] J. R. Deller Jr, J. G. Proakis, J. H. L. Hansen. 1987. Discrete-Time Processing of Speech Signals, Prentice Hall.

[5] N.Virag. 1995. "Speech Enhancement Based on Masking Properties of the Auditory System." Proc. of IEEE ICASSP95, 796-799.

[6] A. A. A zirani, R. L. B. Jeannes and G.Faucon. 1995 "Optimizing Speech Enhancement by Exploiting Masking Properties of the Human Ear." Proc. of IEEE ICASSP95, 800-803.

[7] J. H. L. Hansen and S.Nandkuma. 1995. Robust estimation of speech in noisy background based on aspects of the auditory process, J. Acoust. Soc. Am, Vol.97, 3833-3849.

[8] J. D. Johnston. 1988. "Transform Coding of Audio Signal Using Perceptual Noise Criteria." IEEE J. on Select. Areas Comm., Vol.6, 314-323.

접수일자 : '98. 9. 25.
게재결정 : '98. 10. 30.

▲ Sunmee Kang
Department of Computer Science, Seokyeong University
16-1 Jeongreung-4dong, Sungbuk-ku, Seoul, Korea
Tel: (02) 940-7144
e-mail : smkang@bukak.seokyeong.ac.kr

▲ Wooil Kim
Department of Electronic Engineering, Korea University,
5ka-1 Anam-dong, Sungbuk-ku, Seoul, Korea

▲ Hanseok Ko
Department of Electronic Engineering, Korea University,
5ka-1 Anam-dong, Sungbuk-ku, Seoul, Korea
Tel: (02) 3290-3230 (O), (02) 592-0425 (H)
Fax: (02) 921-0544
e-mail: hsko@kuccnx.korea.ac.kr