Spectral Subtraction based on Speech State and Masking Effect

*Wooil Kim**, *Sunmee Kang***, *Hanseok Ko**
*Dep. of Electronics Engineering, Korea University, 5 ka-1, Anam-dong, Sungbuk-ku, Seoul, Korea
**Dep. of Computer Science, Seokyeong University, 16-1 Jeongreung-4dong, Sungbuk-ku, Seoul, Korea
Email : wikim@ispl.korea.ac.kr

Abstract

In this paper, a speech enhancement method based on phonemic properties and masking effect is proposed. It is a modified type of spectral subtraction wherein the spectral sharpening process is exploited in unvoiced state considering the phonemic properties. The masking threshold is used to remove the residual noise. The proposed spectral subtraction shows similar performance as that of the classical spectral subtraction method in view of the SNR. But by the proposed scheme, the unvoiced sound region is shown to exhibit relatively less signal distortion in the enhanced speech.

1. Introduction

In the common speech enhancement techniques, the noise subtraction or suppression rule is applied to whole speech region equally, so that information of phonemes in utterance is not utilized effectively[1]. The unvoiced sounds have spectral distribution similar to that of the white noise, so that their spectrum is degraded severely by the background noise and subtracted results have significant spectral distortions compared to those of the voiced sounds. In the spectral subtraction proposed in this paper, state-dependent subtraction rule is exploited. An analysis frame is developed to make decisions on the data as either voiced state or unvoiced and then, in the case of unvoiced, noise subtraction is applied to process the spectral sharpening the spectrum. The residual noise which still exists in noise-subtracted spectrum is minimized based on the masking effect. The masking effect is a primary property of human auditory system such that a soft tone can be masked by a strong tone located closed on frequency domain, essentially forcing the masked tones inaudible.

This paper is organized as follows. In section 2, an overview of spectral subtraction is presented and its problem is formulated. The proposed enhancement scheme and its detail procedure 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 the future work.

2. Overview of Spectral Subtraction and Problem Formulation

Spectral subtraction is a method for restoring the power or the magnitude spectrum of a signal observed in additive noise, by subtracting an estimate of the average noise spectrum from the noisy signal spectrum[1][2]. Assume that a windowed noise signal $n(k)$ has been added to a windowed speech signal $s(k)$, with their sum denoted by $y(k)$

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

Taking its 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.

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

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

The estimation by spectral subtraction described above is applied to all speech activity under equal criterion. This is a weakness since it is well known that the approach entails a non-uniform impact across the phoneme sequence of a speech utterance. That is, classical spectral subtraction fails to fully utilize the phonemic information carried within the signal for enhancement processing. The phonemic information includes acoustic properties which are originated in different production process of each phoneme. Therefore, when speech is contaminated by noise, each phoneme in the speech has various distortion in time and spectral domain. The vowels, typical examples of voiced sound, have the resonant frequency components, called "formants", whose energies are relatively high[3]. 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 the white noise. Consequently, the unvoiced sounds are contaminated severly and the restoration tasking is much difficult compared to the voiced sounds.

3. Proposed Methodology

The proposed enhancement method explores on the following features for improved performance.

1) Noise subtraction algorithm dependent on speech state

2) Residual noise reduction using masking

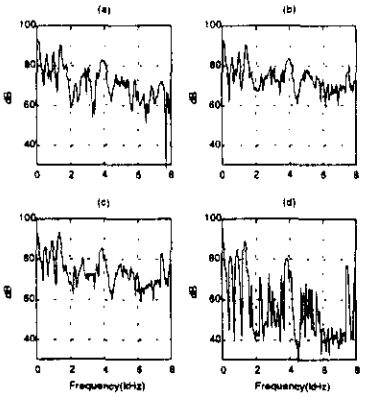

Figure 3.1 (a)Clean spectrum of a frame in speech, (b)Noisy(10dB), (c)Spectral sharpened , (d)Subtraction result

effect.

3.1 State-Dependent Subtraction

As a new approach for reducing spectral obscurity of noisy unvoiced sounds, the spectrum sharpening before noise subtraction is suggested. That is, by making high-level energy spectral component mores high and low-level ones lower, the waveform shape of the noisy spectral can be made more clear. For the spectral sharpening, a simple function is introduced and it is expressed as follows:

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

where L_{SS} is sharpening function size.

Spectral sharpening is accomplished by convoluting the noisy spectrum with suggested sharpening function.

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

$$Y_S(i) = Y_S'(i - L_{SS}/2) \quad \text{if } Y_S'(i - L_{SS}/2) \ge 0$$
$$= Y(i) \quad \textit{otherwise} \qquad (3.3)$$

Convolution with sharpening function, $SS(i)$

has the effect that noisy spectral shape becomes clear some extent through summing the close-adjacent frequency components and suppressing the far-components. Figure 3.1 (c) shows the sharpening function convoluted onto the spectrum of noisy unvoiced sound in a frame.

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 metal-like sound and annoys 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 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[4][5].

The residual noise is minimized possibly by setting the negative subtracted spectral components to a masking threshold. The computation of the masking threshold introduced in [6] 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 absolute threshold, and 5) renormalization.

3.3 Integrated Enhancement Scheme

Block diagram in Figure 3.2 presents totally integrated speech enhancement algorithm including both the state-dependent noise subtraction and residual noise reduction using 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, it is required to study more reliable decision rule for voiced/unvoiced state.

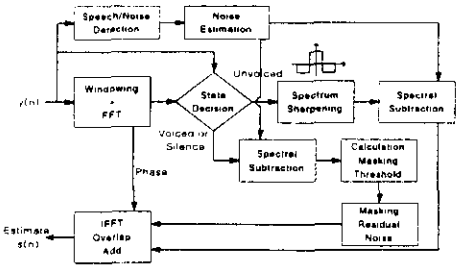

Figure 3.2 Block diagram of proposed spectral subtraction.

4. Experimental Results

For the performance of the proposed algorithms, an experiment of 3 subtractive type algorithms has 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) Proposed method : state-dependent power spectral subtraction with residual noise reduction using masking threshold (ST-SPSUB+MSKRNR)

The original clean speech data is a word "computer" which is recorded by a 20s man in a clean room and sampled at 16kHz. The noise is artificial white Gaussian. In implementation of spectral subtraction algorithm, the size of a analysis frame is 16 msec(256 points) and a overlap lag is 8 msec. Hamming windowed signal is analyzed in 256-point FFT.

Figure 4.1 (a) shows noisy speech and decided region as unvoiced state and (b) shows enhanced speech by ST-SPSUB+MSKRNR. Figure 4.2 (a), (b) present spectrogram of noisy speech and enhanced speech respectively. From figures, we can see that the residual noise as well as the background noise is considerably minimized by the proposed scheme. The listening test also shows an improvement in intelligibility of the enhanced speech. The performance comparison in terms of the input-output SNR shows that the

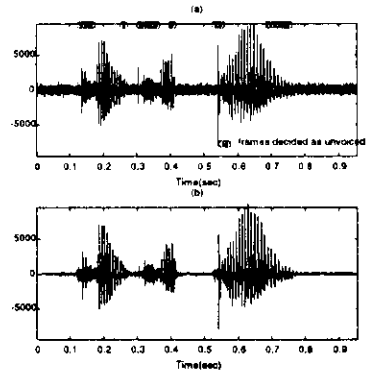

Figure 4.1 (a) Noisy speech(10dB) and decided regions as unvoiced. (b)Enhanced speech by ST-SPSUB+MSKRNR

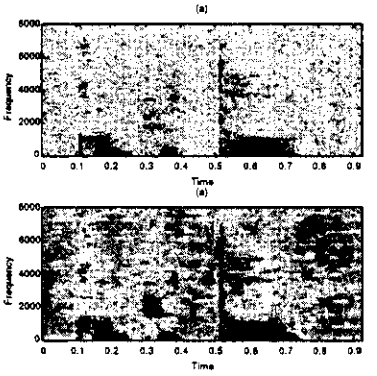

Figure 4.2 (a) Spectrogram of noisy speech. (b) Spectrogram of enhanced speech by ST_SPSUB+MSKRNR

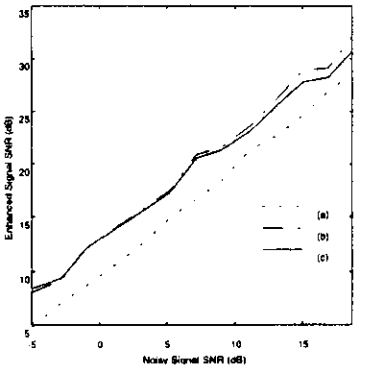

Figure 4.3 Performance comparison in terms of input-output SNR: (a) SPSUB (b) SPSUB+RNR (c) ST-SPSUB+MSKRNR

"ST-SPSUB+MSKRNR" is much better than the "SPSUB+RNR" in the low SNR cases (Figure 4.3).

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..).

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

Reference

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

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

[3] J.R.Deller Jr,J.G.Proakis, J.H.L.Hansen, Discrete-Time Processing of Speecg Signals, Prentice Hall, New Jersey, 1987.

[4] N.Virag, "Speech Enhancement Based on Masking Properties of the Auditory System," Proc. IEEE ICASSP, pp. 796-799, Detroit, Michigan U.S.A., 1995.

[5] A.A.Azirani, R.L.B.Jeannes and G.Faucon, "Optimizing Speech Enhancement by Exploiting Masking Properties of the Human Ear," Proc. IEEE ICASSP, pp. 800-803, Detroit, Michigan U.S.A., May 1995.

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