14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

A HARSH NOISE ASSESMENT MEASURE FOR SPECTRAL SUBTRACTION

Kohei Yamashita and Tetsuya Shimamura

Department of Information and Computer Sciences, Saitama University,
255 Shimo-Okubo, Saitama 338-8570, Japan
{k-yama, shima}@sie.ics.saitama-u.ac.jp

ABSTRACT

In this paper, we derive a new assessment measure, which is suitable for harsh noise like musical noise. The measure uses an image processing technique and a filter based on auditory loudness. The aim of the measure is to evaluate the harshness of noise without subjective experiments. The effectiveness of the proposed measure is shown with experiments on real speech data.

1. INTRODUCTION

Background noise is added to speech and degrades the performance of speech processing systems. This is true in systems of speech communication, speech analysis, speech recognition and so on. In such systems, clean speech is desired. This is the reason why various speech enhancement techniques have been studied up to now for the purpose of eliminating noise. Spectral subtraction (SS) is one of the methods to enhance speech in noise [1][2]. The SS has been widely used since it can suppress noise effectively with simple computation. It is, however, known that an artificial noise occurs due to estimation error of noise spectrum, resulting in degradation of the performance of the SS. The artificial noise is called "musical noise" because of its strange tones. To conquer this problem, modified SS methods have been proposed [3]–[8]. Recently, the SS methods working effectively in very low signal-to noise ratio (SNR) environments have been devised [17]–[20].

On the other hand, assessment measures exist to evaluate the usefulness of a speech enhancement technique. Signal-to-Noise Ratio (SNR) is the most widely used measure in which the signal and noise powers are compared. However, SNR is a poor indicator of speech quality. This is because speech waveform distortions are not considered and speech and noise durations are mixed in SNR. A frame-based SNR called Segmental SNR is more matched with our auditory perception. Even for the segmental SNR, however, noise durations are included in the calculation, resulting in an assessment measure depending on noise. The frequency weighted segmental SNR measure may be the most accurate SNR evaluation in which SNR is calculated for each frequency band of auditory perception based frequency division [9]. For evaluation methods utilizing frequency characteristics, several approaches exist such as Log Area Ratio [10], Weighted Spectral Slope Measure [11][12] and Itakura Measure [13]. It is known that the frequency based measures are closed to the subjective measure [14].

In the case where the SS method is used in a very low SNR environment around of 0dB, the quality of speech processed by the method is sufficiently affected by the residual noise including musical noise. In that case, even if the distortion of speech itself is little, the noise remained in the processed speech gives harshness. If we evaluate the quality of the processed speech by using the conventional speech quality measures, then the results will be obtained based on the speech characteristics, particularly frequency characteristics. Then, the speech intelligibility will not be accurately assessed. Thus, in this paper, we consider to separate the speech intelligibility and noise amount in the processed speech, and set out to measure the occupied noise amount from the view point of auditory perception in the case of low SNR conditions. By using the subjective measure, for example, Mean Opinion Score (MOS), the scores may be separated into the speech intelligibility part and noise amount part. As well known, however, it takes a long time to evaluate them, because many listeners have to be gathered and listen to the speech data. In this paper, we propose an objective measure to assess the remained noise, and evaluate the harshness which results in a score of the dissonance of noise.

2. PROPOSED MEASURE

We assume to have a speech signal $x(n)$ corrupted by an additive noise $w(n)$. Then the noisy speech signal is described by

$$y_k(n) = x_k(n) + w_k(n) \tag{1}$$

where k is a frame number. In such a situation, the SS is considered as filtering in the frame, which is described by

$$|\hat{X}_k(f)| = |Y_k(f)| - |\hat{W}_k(f)| \tag{2}$$

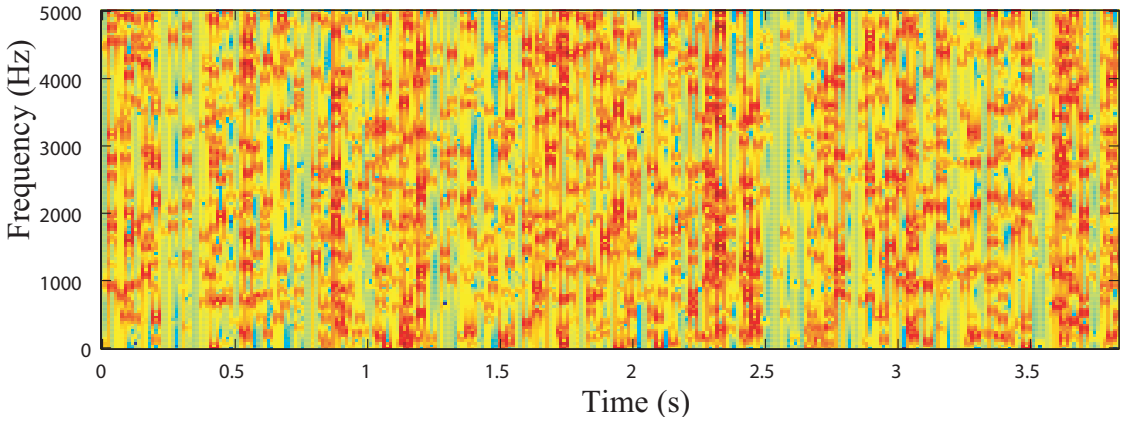

Fig. 1. Spectrogram of a musical noise.

where $Y_k(f)$, $X_k(f)$ and $W_k(f)$ are the spectra of windowed signals $y_k(n)$, $x_k(n)$ and $w_k(n)$, respectively. The $\hat{\cdot}$ means estimation. Generally, $|\hat{W}_k(f)|$ is obtained from non-speech segments. The SS is implemented based on the assumption that the characteristic of noise is stationary. The characteristic of noise, however, may not be stationary completely in each segment. Therefore, $|\hat{X}_k(f)|$ includes residual noise produced by noise estimation error.

For analyzing the residual error, let us assume here that the residual noise is obtained as

$$E_k(f) = \begin{cases} |\hat{X}_k(f)| - |X_k(f)| & |\hat{X}_k(f)| > |X_k(f)| \\ 0, & \text{else.} \end{cases} \quad (3)$$

Goh et al.[8] addressed an approach to suppress the musical noise included in $E_k(f)$ by post-processing after the implementation of SS. In the method, the spectrogram of the processed speech is graphically manipulated so as to reduce the musical noise. The principle of the manipulation is that the musical noise appears as random spectral peaks on the spectrogram and eliminating those spectral peaks leads to reduction in musical noise. Figure 1 shows the spectrogram of a musical noise produced imitatively from the principle of the above random spectral peaks by a computer. We can actually perceive musical tones on this noise.

We utilize an idea of edge extraction used in image processing so as to detect the random spectral peaks as impulsive estimation errors. The estimation error score is obtained for each frequency in the spectrogram as

$$P_k(f) = \frac{|-E_{k-w}(f) - ... + 2wE_k - ... - E_{k+w}|}{2w+2} \quad (4)$$

where w is an integer and $2\omega+1$ corresponds to the number of frames used for the calculation as one block. When the difference in spectral magnitude between priori and posteriori frames for each frequency is large, $P_k(f)$ also becomes large. Figure 2 shows an image figure to implement (4).

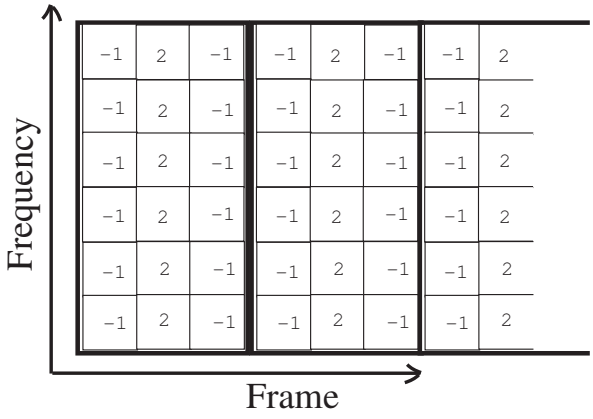

Fig. 2. A filtering image of the proposed measure in the case of $w = 1$.

It is, however, known that human auditory perception is different in each frequency. For this reason, we should calculate the auditory scores so as to compensate for the perception gap. Accordingly, we use a loudness sone $L(f)$ as a frequency filter[15]. The loudness sone is a frequency ratio of the degree of human auditory perception. This ratio is used in frequency weighted segmental SNR [16] which is an improved measure based on SNR. The score of the frequency weighted segmental SNR gives the highest one between objective and subjective correlations[14]. Figure 3 shows the frequency ratio of loudness sone. By using the loudness sone, the measure of harsh noise, HArsh Measure(HAM), is calculated in each frame by

$$\Phi_k(f) = P_k(f) * L(f). \quad (5)$$

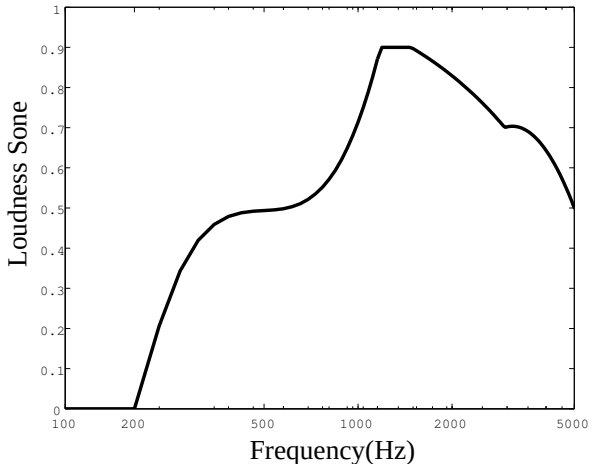

Fig. 3. A loudness sone.

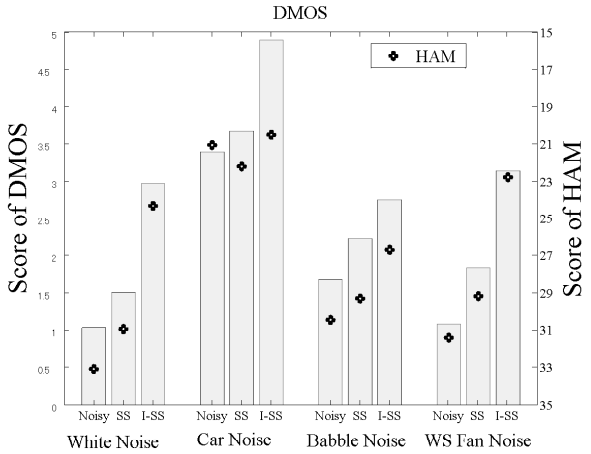

Fig. 4. Comparison between Score of Normalized DMOS and several objective measure (Input SNR = 0dB.)

The final HAM is obtained by averaging

$$\Phi = \frac{1}{K}\sum_{k}\frac{1}{F}\sum_{f}\Phi_k(f) \tag{6}$$

$$HAM = 10\log_{10}\Phi \tag{7}$$

where K is the frame number used to calculate the HAM segmentally.

3. COMPARISON

To validate the effectiveness of the proposed measure, we conducted experiments on real speech data. Speakers are a Japanese male and female. Each speech length is about 10 seconds. The sampling frequency is 10 kHz and band limitation 3.4 KHz.

The parameters for experiments are as follows. The frame length is 51.2ms(512 points), which is obtained by windowing by Hanning window. Each frame is half-overlapped. The Fast Fourier Transform (FFT) used in the SS is with 512 points. The w used in HAM is 1.

We use two conventional speech enhancement methods to compare the effectiveness of the proposed assessment measure. One is the standard SS method[3], because it is the most standard. And the other is an iterative SS with weighting factors[17][19] because the iterative SS method aims to residual noise reduction, eliminating musical noise simultaneously. These methods use a full-wave rectification processing commonly.

First, we calculate HAM on these speech data corrupted by 4 kinds of noise. The noises are white noise (white), car engine noise (car), babble noise at exhibition (babble) and work station fan noise (fan). The white noise was generated by a computer and the other noises obtained from a database. Table 1 shows the results of HAM. Table 1 shows that some differences exist by noise characteristics in the same SNR cases. This suggests that the harshness of noise is not able to be judged by only SNR. On the other hand, for a comparison of the same noise characteristics at each SNR, each score of HAM becomes larger as each noise level becomes higher.

Table 1. Scores of HAM

SNR	Process	White	Car	Babble	Fan
0dB	Noisy	32.98	20.97	30.36	31.30
	SS	30.85	22.11	29.22	29.09
	Iterative SS	24.23	20.42	26.62	22.69
5dB	Noisy	30.53	18.38	27.84	28.86
	SS	28.04	19.36	26.47	26.30
	Iterative SS	21.47	18.01	23.38	19.96
10dB	Noisy	28.07	15.78	25.30	26.39
	SS	25.29	16.36	23.72	23.59
	Iterative SS	18.72	15.14	20.55	17.65

Next, we compare the proposed evaluation measure with the conventional SS methods by listening test. We used a DMOS (Differential Mean Opinion Score) as a subjective evaluation criterion. In this experiment, we only considered the harshness of noise. Thus, the quality of speech was scored based on the following 5 levels:

- 5 : no harsh.
- 4: little harsh.
- 3: so so.
- 2: harsh.
- 1: cannot bear to hear.

Speech data about 10 listeners were chosen randomly for each evaluation on a noisy speech data where the SNR is 0dB. Each listener was listened twice to each speech data prepared in a random order. And the scores obtained from the listeners for each speech data were averaged. Figure 4 shows the comparison between the results of HAM and DMOS. In Figure 4, evaluation between the HAM and the DMOS is compared as the ratio of the value because an evident basis value does not exist in HAM. Figure 4 shows that the proposed measure almost correctly matches the result of the listening test. By comparing the SS-based methods, we see that the HAM measure expresses the harshness of the noise because the iterative SS method is more excellent with respect to musical noise reduction. In the case of car noise, the score of HAM processed with SS is worse than noisy. The differences may be due to that the energy of car noise is concentrated on the low frequency range mainly.

4. CONCLUSIONS

We have derived a new assessment measure. The measure represents the degree of harshness of noise. The measure is calculated by using an edge extraction technique in image processing with a loudness sone based on human auditory perception. Some experiments on real speech data show the effectiveness of the proposed measure.

5. REFERENCES

[1] J. S. Lim, "Speech Enhancement,"Prentice-Hall, Englewood Cliffs, 1983.

[2] J. S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," *IEEE Trans. Acoust. Speech and Signal Processing*, Vol.ASSP-26, no.5, pp.471-472, Oct. 1978.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech and Signal Processing*, Vol.ASSP-27, pp.113-120, Apr. 1979.

[4] M. Berouti, T. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. IEEE ICASSP*, pp.208-211, April,1979.

[5] R. Martin, "Spectral subtraction based on minimum statics," *Proc. EUSIPCO*, pp.1182-1185, September 1994.

[6] V. Stahl, A. Fischer and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," *Proc. IEEE ICASSP*, pp. 1875-1878, June 2000.

[7] I. Cohen, "Noise estimation by minima controlled recursive averaging for robust speech enhancement,' *IEEE Signal Processing Letters*, Vol.9, no.1, pp.12-15, January 2002.

[8] Z. Goh, K. Tan and B.T.G. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Trans. Speech and Audio Processing*, Vol.6, no.3, pp.287-292, May,1998.

[9] J. M. Tribolet, P. Noll, B. J. McDermott and R. E. Crochiere, "A study of complexity and quality of speech waveform coders," *Proc. IEEE ICASSP*, pp.586-590, April,1978.

[10] T. P. Barnwell III, M. A. Clements, S. R. Quackenbush and E. P. Farges, "Improved objective measures for speech quality testing," Final Report no.100-83-C-0027, DCA, Sep 1984.

[11] J. H. L. Hansen and L. M. Arslan, "Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit card corpus" *IEEE Trans. Speech and Audio Processing*, Vol.3, no.3, pp.169-184, May 1995.

[12] D. Klatt "Prediction of perceived phonetic distance from critical-band spectra: A first step" *Proc. IEEE ICASSP*, pp.1278-1281, May 1982.

[13] F. Itakura, "Minimum prediction residual applied to speech recognition," *IEEE Trans. Acoust. Speech and Signal Processing*, Vol.23, no.1, pp.67-72, Feb. 1975.

[14] J. R. Deller Jr., J. G. Proakis and J. H. L. Hansen, "Discrete-Time Processing of Speech Signals," Macmillan, 1993.

[15] "Auditory and Speech," Corona Pub., 1980. (in Japanese)

[16] R. E. Crochiere, L. R. Rabiner, N. S. Jayant and J. M. Tribolet, "A study of objective measures for speech waveform coders," *Proc. Zurich Seminar on Digital Communications*, March, 1978

[17] K. Yamashita, S. Ogata and T. Shimamura, "Spectral subtraction iterated with weighting factors", *Proc. IEEE Speech Coding Workshop*, pp.138–140, Oct. 2002.

[18] J. Beh and H. Ko, "A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech", *Proc. IEEE ICME* Vol. 3, no. III, pp.633-636 July, 2003.

[19] K. Yamashita, S. Ogata and T. Shimamura, "Improved spectral subtraction utilizing iterative processing", *IEICE Trans.* Vol.J88-A, no.11, 2005 (in Japanese).

[20] H. Nakashima, Y. Chisaki, T. Usagawa and M. Ebata, "Spectral subtraction based on statistical criteria of the spectral distribution", *IEICE Trans. Fundamentals*, Vol.E85-A, no.10, pp.2283-2292, Oct. 2002.