NOISE REDUCTION BY PAIRED-MICROPHONES USING SPECTRAL SUBTRACTION

Mitsunori Mizumachi and Masato Akagi

School of Information Science,
Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Tatsunokuchi, Nomigun, Ishikawa 923-12, Japan.
E-mail: {mizumati, akagi}@jaist.ac.jp

ABSTRACT

This paper proposes a method of noise reduction by paired microphones as a front-end processor for speech recognition systems. This method estimates noises using a subtractive microphone array and subtracts them from the noisy speech signal using the Spectral Subtraction (SS). Since this method can estimate noises analytically and frame by frame, it is easy to estimate noises not depending on these acoustic properties. Therefore, this method can also reduce non stationary noises, for example sudden noises when a door has just closed, which can not be reduced by other SS methods. The results of computer simulations and experiments in a real environment show that this method can reduce LPC log spectral envelope distortions.

1. INTRODUCTION

A large number of current speech recognition systems for clean speech display great abilities in ideal environments that have no noises. However, performances of these speech recognition systems go down extremely in real environments because noises distort speeches. Then, this paper proposes an effective method of noise reduction for them in real environments, that is based on NORPAM [1] proposed by the authors.

The noise reduction methods for speech recognition systems are broadly divided into two categories. The one approach is to change model parameters to accommodate noisy speech signals in the HMM frameworks. It is said that these methods are unsuitable for practical use, because they require large computational costs. The another approach is to equip speech recognition systems with a noise reduction system as a front-end. This proposed method adopts the later and use small 3ch. microphone array to put it practical use.

Delay-and-Sum array[3] is the most conventional and popular as the microphone array for noise reduction, but it must use huge number of microphones to achieve high accuracy. Adaptive array using only a few microphones, for example Griffiths-Jim type[4] works very well for some noises in ideal environments. However, adaptive array is weak in dealing with sudden noises or moving noises and signals received in reverberant environments. The proposed method accepts the basic concept of subtractive array, for example Griffiths-Jim type, but do not use adaptive filters. In the place of using adaptive filters, noises are analytically estimated with very small computational costs, on the basis of arrival time differences between paired microphones frame by frame. Next, the estimated noises are subtracted from noisy signal received by a microphone in the frequency domain using the Spectral Subtraction(SS)[2].

This method is evaluated by some experiments of noise reduction. Experimental results show that this method can reduce LPC log spectral envelope distortion not only in computer simulations, but also in a real environment.

2. ALGORITHM DESCRIPTION

This method uses a microphone array constructed with three linear-equally-spaced omni-directional microphones to estimate the largest noise at the position of the center microphone in each short time period. Then, noise reduction can be accomplished by subtracting the estimated noises from the signal received by the center microphone.

2.1. Estimation of Noise

Noises are estimated by using the signals received by paired microphones such that microphones locating at both ends of microphone array(main pair) or the center and one of both ends of microphone array(sub pair). Let us assume that speech signal comes from a certain direction and noises come from directions except the speech direction as shown in Fig. 1. On the assumption that speech signal $s(t)$ comes from a direction such as the difference in arrival time between main paired microphones is 2ζ, and the largest noise $n(t)$ comes from a direction such as that is 2δ, signals received at each microphone are described as follows:

$$\text{left mic.}: \quad l(t) = s(t-\zeta) + n(t-\delta), \tag{1}$$

$$\text{center mic.}: \quad c(t) = s(t) + n(t), \tag{2}$$

$$\text{right mic.}: \quad r(t) = s(t+\zeta) + n(t+\delta). \tag{3}$$

Here, let us assume that speech signals come from the front, i.e. $\zeta = 0$, to make explanation below simply. Received signals are transformed by short-term Fourier transformation(STFT) as

$$L(\omega) = S(\omega) + N(\omega)e^{-j\omega\delta}, \tag{4}$$

$$C(\omega) = S(\omega) + N(\omega), \tag{5}$$

$$R(\omega) = S(\omega) + N(\omega)e^{j\omega\delta}, \tag{6}$$

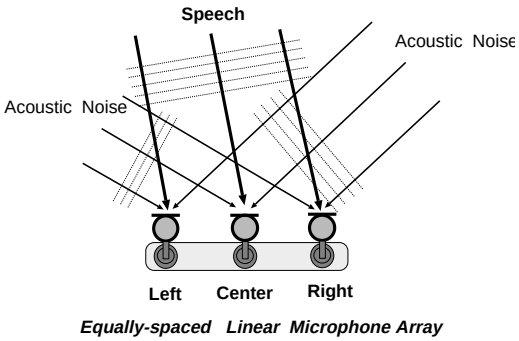

Figure 1: Relationship between a microphone array and acoustic signals.

where $S(\omega)$ is the STFT of the speech signal $s(t)$, and $N(\omega)$ is that of the largest noise $n(t)$. Then, $l(t)$ in Eq. (1) and $r(t)$ in Eq. (3) are shifted $\pm\tau$ in time, where τ is a certain constant ($\tau \neq 0$), and these make a function $g_{lr}(t)$. A function $g_{lr}(t)$ is a beamformer in time domain, and its STFT is $G_{lr}(\omega)$. They are actually defined as

$$g_{lr}(t) = \frac{\{l(t+\tau) - l(t-\tau)\} - \{r(t+\tau) - r(t-\tau)\}}{4}, \quad (7)$$

$$G_{lr}(\omega) = N(\omega)\ \sin\omega\delta\ \sin\omega\tau. \quad (8)$$

The value δ in Eq. (8) indicates the direction that the largest noise comes, so it is decided by estimating where noise comes from in each frame (this is described later). Spectrum of the noise can be calculated by setting a certain value τ at the estimation of δ and dividing Eq. (8) by $\sin^2\omega\tau$. However, it is not accurately calculated in case of $\omega\tau = n\pi$ in Eq. (8). In that frequency band, signals obtained by sub paired microphones form the another beamformer $g_{cr}(t)$ and its STFT $G_{cr}(\omega)$ as follow:

$$g_{cr}(t) = \frac{\{c(t+\tau_2) - c(t-\tau_2)\} - \{r(t+\tau_2) - r(t-\tau_2)\}}{4}, \quad (9)$$

$$G_{cr}(\omega) = N(\omega)\, e^{j\omega\frac{\delta}{2}}\ \sin\omega\frac{\delta}{2}\ \sin\omega\tau_2. \quad (10)$$

The spectrum of the largest noise $n(t)$ is estimated over all frequency range as

$$\widehat{N}(\omega) = \begin{cases} G_{lr}(\omega)/\sin^2\omega\delta, & \sin^2\omega\delta > \varepsilon_1 \\ G_{cr}(\omega)\, e^{-j\omega\frac{\delta}{2}}/\sin^2\omega\frac{\delta}{2}, & \sin^2\omega\delta \leq \varepsilon_1 \\ & \textit{and}\ \sin^2\omega\frac{\delta}{2} > \varepsilon_2 \\ G_{lr}(\omega)/\varepsilon_2^2, & \sin^2\omega\frac{\delta}{2} \leq \varepsilon_2 \end{cases} \quad (11)$$

where ε_1 and ε_2 are certain threshold values, and $\widehat{N}(\omega)$ is approximated only in the frequency bands such as $\sin^2\omega/2$ comes up to $n\pi/2$ (n: integer).

2.2. Estimation of Noise Direction

Noise directions are automatically estimated frame by frame. In this paper, two signals that speech signal is perfectly eliminated give noise directions, and they are prepared by Eq. (9), Eq. (10) and as follow:

$$g_{lc}(t) = \frac{\{l(t+\tau_2) - l(t-\tau_2)\} - \{c(t+\tau_2) - c(t-\tau_2)\}}{4}, \quad (12)$$

$$G_{lc}(\omega) = N(\omega)\, e^{-j\omega\frac{\delta}{2}}\ \sin\omega\frac{\delta}{2}\ \sin\omega\tau_2. \quad (13)$$

Here, speech signals have no effect on estimating noise directions, as $G_{cr}(\omega)$ in Eq. (10) and $G_{lc}(\omega)$ in Eq. (13) do not include speech signals at all. Setting τ_2 in Eq. (10) and Eq. (13) arbitrary, the following is calculated,

$$d(t) = \mathrm{IFFT}\left[\frac{G_{lc}(\omega)G_{cr}^*(\omega)}{|G_{lc}(\omega)||G_{cr}(\omega)|}\right], \quad (14)$$

$$d = \underset{t}{\mathrm{argmax}}\,[\,d(t)\,]\,. \quad (15)$$

Then, the value δ, that is half of the difference in arrival time between main paired microphones is given as d in Eq. (15). However, noise directions are sometimes miss-estimated by estimating them in very short time frames. Noise directions are finally decided with estimated value d in Eq. (15) in three frames by

$$\delta = \underset{d}{\mathrm{argmax}}\left[\underset{i=1,\,2,\,3}{\mathrm{histgram}}(d_i)\right]. \quad (16)$$

2.3. Reduction of Noise

After estimating the spectrum of noise, it is neccessary to subtract it from that of noisy speech signal received by the center microphone. Since this method takes aim at noise reduction as a front-end for speech recognition systems, it is enough to deal with amplitude spectra only. For the subtraction of amplitude spectra, this method employs the SS. The noise spectrum may be lager than that of input noisy speech signal because of approximating the noise spectrum in Eq. (11) or miss-estimating the noise directions. This method uses the non-linear SS such as

$$|\widehat{S}(\omega)| = \begin{cases} |C(\omega)| - \alpha\cdot|\widehat{N}(\omega)|, & |C(\omega)| \geq \alpha\cdot|\widehat{N}(\omega)| \\ \beta\cdot|C(\omega)|, & \text{otherwise}, \end{cases} \quad (17)$$

where α is a subtraction coefficient, and β is a flooring coefficient. Thus, this method can reduce the distortions of amplitude spectra cased by acoustic noises.

In regards to the SS, this method is superior to other methods. This method can cope with all types of acoustic noises by estimating the spectra of noises time by time, but other methods be poor at eliminating the non stationary noises, such as sudden noises, because they substitute signals received in the past for the noises in greater or lesser degree.

3. EXPERIMENTS AND RESULTS

Some computer simulations and experiments in a real environment have been conducted to examine the availability of this method. In this paper, non stationary noises are used for the reason of displaying the advantage of this method, since it is impossible for almost all other methods that use

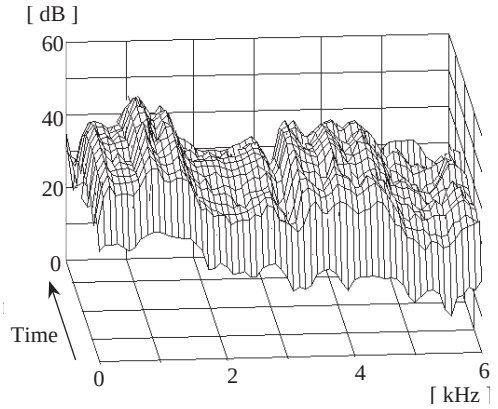

Figure 2: Clean speech /a/.

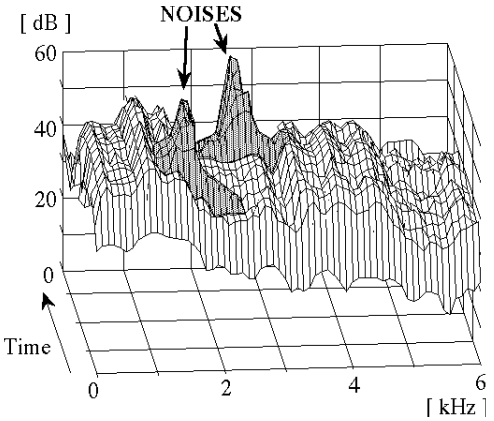

Figure 3: Noise-added speech signal(noises are blackened).

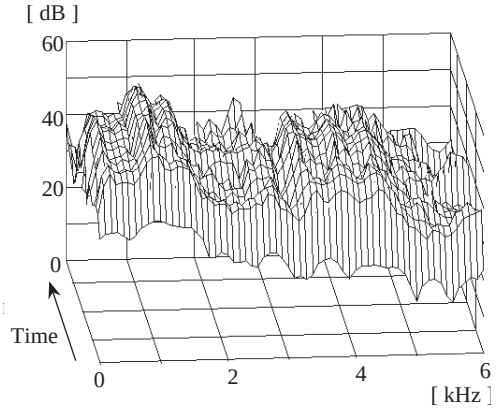

Figure 4: Noise-reduced speech signal.

the adaptive filter or the SS. All sound data prepared are sampled at 48 kHz with 16 bit accuracy.

Noise reductions have been conducted under the conditions: frame length is 5.3 msec, frame shift is 2.7 msec, window function is hamming, threshod values ε_1 and ε_2 are 0.6 and 0.2, and coefficients α and β for the SS are 1 and 0.001 respectively. Here, frame length is set as possible as small to decrease the distortions caused by the SS, and other parameters are set experimentally.

3.1. Computer Simulations

For computer simulations, the noises that suddenly arise and disappear and sweep tones in the frequency range that speech signals exist are prepared.

A clean speech signal shown in Fig. 2 is the utterance of Japanese vowel /a/ in the ATR speech database. And two sudden noises are prepared such as narrow-band noises with the center frequency of 1500 Hz or 2500 Hz and a bandwidth of 200 Hz, and both of them exist only 50 msec. Then, they are mixed in a computer on the assumption that a speech signal comes from the front and both of two noises come from about 30 degrees to the right. The noise-added speech signal is shown in Fig. 3, where noises are blackened. As a simulation result, noise-reduced speech signal is shown in Fig. 4. Compared Fig. 4 with Fig. 2 and Fig. 3, it is obvious that sudden noises can be well enough reduced by this method.

For the objective evaluation, the following LPC log Spectrum envelope Distortion (LPC-SD) is calculated. LPC-SD is good for this method as a front-end of speech recognition system because recent speech recognition systems use LPC analysis to calculate the distance between reference patterns and input speech. LPC-SD is actually caluculated as,

$$\text{LPC-SD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\{S_x(i) - S_c(i)\}^2} \qquad [\text{ dB }], \quad (18)$$

where $S_c(\omega)$ is the LPC log spectrum envelope of clean speech signal, $S_x(\omega)$ is that of the speech to evaluate, and N = 6kHz. LPC analyses have been done under the condition: sampling frequency is 12 kHz, frame length and frame shift are 21.3 msec and 5.3 msec, window function is hamming, order of LPC analysis is 16, and pre-emphasis is 0.98.

Then, the change of LPC-SD in each frame by adding and reducing sudden noises is shown in Fig. 5, where the dotted and solid lines show LPC-SD before and after noise reduction respectively. In all frames that speech signal and sudden noise exist together, LPC-SD has been reduced and the mean improvement of LPC-SD is 7.07 dB. SNR of noise-added speech signal shown in Fig. 5 is -10 dB. On the other hand, the mean improvements of LPC-SD under the other conditions that speech signals are less noisy are shown in Fig. 6 with the cases of broad-band sudden noises that bandwidths of sudden noises are changed from 200 Hz to 1000 Hz.

Next, a sweep tone as another noise is prepared, and it varies continuously from 1 kHz to 6 kHz in 1 sec. Clean speech signal made by looping the stationary part of Japanese vowel /a/ and a sweep tone are mixed in the same way above. As a result of noise reduction, LPC-SD of noise-added speech signal and that of noise-reduced speech signal are shown in Fig. 7 with the peak frequency of a sweep tone in each frame, and the mean improvement of LPC-SD is 8.61 dB. Consequently, it is not too much to say that this method can reduce noises whose characteristics vary continuously in both time and frequency domains.

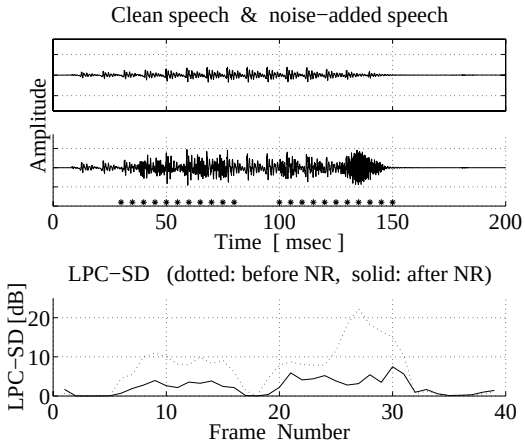

Figure 5: Clean speech signal, noise-added speech signal(the mark '∗' shows that noise exsits in that time, and LPC-SD (the mean improvement is 7.07 dB).

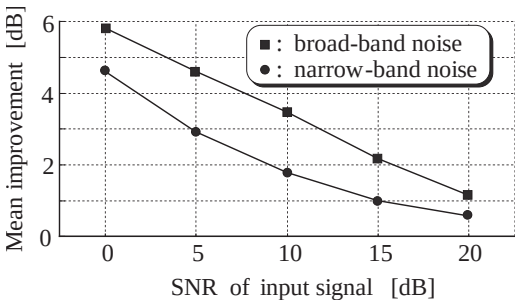

Figure 6: Mean improvements of LPC-SD for narrow-band and broad-band noises.

3.2. Experiments in the Real Environment

In addition to computer simulations, experiments in a real environment have been conducted. Sound data are prepared by presenting the clean speech signal and noise signal through loud speakers that located at the front and about 30 degrees to the right respectively, and recording by the microphone array. The room used for experiments is soundproofed, and its reverberation time(RT) at 500 Hz is about 50 msec but RTs at the lower frequencies are fair long.

Here, one of the experimental results is shown on account of space consideration. Sound data prepared is the same as the computer simulation(Fig. 5) that uses narrow-band sudden noises. The result of noise reduction is shown in Fig. 8 and the mean improvement of LPC-SD in frames both speech signal and sudden noise exist is 5.30 dB. LPC-SD has been reduced in not only computer simulations but also a real environment, but reverberations let performance of this method decrease. This method often fails to estimate noise directions in real environments, since the mean improvement rises to 6.02 dB by giving noise directions.

4. CONCLUSION

This paper proposed a method as a front-end noise reduction processor for speech recognition systems. This method

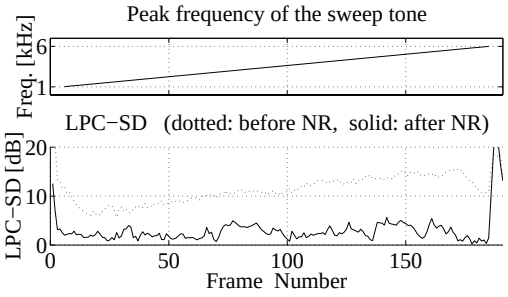

Figure 7: Peak frequency of a sweep tone as a noise, and LPC-SD (SNR of noise-added speech signal is about −6 dB and the mean improvement is 8.61 dB).

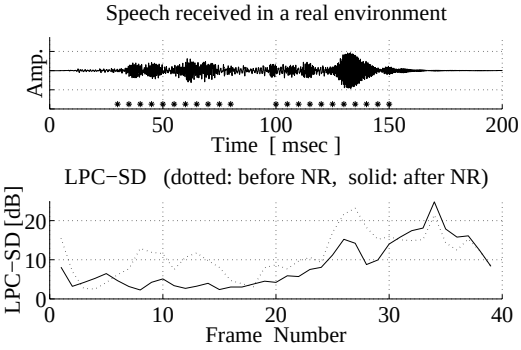

Figure 8: Speech signal received by the center microphone, and LPC-SD (SNR of recorded speech signal is about −12 dB and the mean improvement is 5.30 dB).

estimates noises using a subtractive microphone array and subtracts them from the noisy speech signal by the SS. This method can estimate noises analytically not depending upon these acoustic properties in each frame, so this can also reduce non stationary noises. It is confirmed by computer simulations and experiments in a real environment.

In the future, if noise directions can be estimated more precisely in real environments, this method will be practically achieve its purpose.

5. REFERENCES

[1] M. Akagi and M. Mizumachi: "Noise Reduction By Paired Microphones," Proc. of EUROSPEECH 97, vol. 1, pp. 335-338 (1997)

[2] S. F. Boll: "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. ASSP, vol. 27, no. 2 (1979)

[3] J. L. Flanagan, D. A. Berkley, G. W. Elko, J. E. West and M. M. Sondhi: "Autodirective microphone systems ," ACUSTICA, vol. 73, no. 2 (1991)

[4] L. J. Griffiths and C. W. Jim: "An Alternative Approach to Linearly Constrained Adaptive Beamforming," IEEE Trans. AP, vol. 30, no. 2 (1982)