On Effective Dual-Channel Noise Reduction for Speech Recognition in Car Environment*

Sungjoo Ahn · Sunmee Kang*** · Hanseok Ko****

ABSTRACT

This paper concerns an effective dual-channel noise reduction method to increase the performance of speech recognition in a car environment. While various single channel methods have already been developed and dual-channel methods have been studied somewhat, their effectiveness in real environments, such as in cars, has not yet been formally proven in terms of achieving acceptable performance level. Our aim is to remedy the low performance of the single and dual-channel noise reduction methods. This paper proposes an effective dual-channel noise reduction method based on a high-pass filter and front-end processing of the eigendecomposition method. We experimented with a real multi-channel car database and compared the results with respect to the microphones arrangements. From the analysis and results, we show that the enhanced eigendecomposition method combined with high-pass filter indeed significantly improve the speech recognition performance under a dual-channel environment.

Keywords: noise reduction, speech enhancement, speech recognition, noise robust

1. Introduction

In real environments, the presence of interfering noises always greatly degrades the performance of speech recognition systems. That is, although a number of commercial speech recognition systems are currently available, their performance is degraded substantially under real-world conditions. In a car for example, noise sources are particularly numerous and uncertain noises like vibrations, fan, and noises from open windows generate a spatially distributed background noise. The driver's utterances therefore needs to be enhanced before processing by a speech recognition system. Therefore, robust speech recognition emerges as one of the key technologies to produce voice control car devices.

---

* This work was supported by Ministry of Commerce, Industry and Energy.

** Department of Electronics and Computer Engineering, Korea University

*** Department of Computer Science, Seokyeong University

Many techniques have been developed to solve these problems over the past decades [1]. Among those, two different approaches have been pursued to reduce noise components. The first one uses only one microphone, whereas the second one uses multiple microphones. The most often used single channel noise reduction method is spectral subtraction [2]. However, this single channel method introduces various problems, such as musical tones. In contrast, multi-channel approaches can produce good performance but that implementation in a car is difficult and costly.

Considering low computation and easy implementation, we chose the two- channel noise reduction approach. There are various approaches such as the adaptive noise canceling (ANC) method, the delay-and-sum beamformer, the signal separation method and the method combined with the single-channel approaches [1]. The standard ANC method causes high distortion of the speech signal if any crosstalk interference exists between the two channels. In order to solve this crosstalk interference problem we can use the eigendecomposition method which is a sort of blind source separation method [3]. But if two microphones' signal components are similar, which is the case when two microphones are arranged in similar position or the power of the signals are similar, then the above eigendecomposition method cannot perform appropriately.

To cope with these problems, in this paper we reconstructed new two channel signals from two original microphones' signals. Also, generally driving car noises are dominated by lower frequency components from air flow and tire noise [4]. Thus, to reduce the low frequency components, we applied a high-pass filter to each of the microphone signals.

While a number of studies have investigated various speech enhancement and processing schemes for in-vehicle speech systems, the majority of results were recorded under controlled simulated conditions inside a room or with prerecorded car noise. Little research has been performed using actual voice data collected in a car with associated environmental noise conditions. But in this paper, we experimented with a real multi-channel car database and compared the results with respect to the microphones arrangements.

This paper is organized as follows. In Section 2, we described the two-channel noise reduction method and proposed noise reduction methods to enhance the speech recognition performance. We then conducted the representative experiments and discussed the results on the performance of the proposed methods in Section 3. Finally, in Section 4, conclusive remarks are presented.

2. Dual-Channel Noise Reduction Methods

Although humans can hear with only one ear, hearing with two ears is clearly superior. This is a key idea of the dual channel noise reduction method. In this section, we describe dual channel noise reduction methods.

2.1 Delay-and-Sum Beamformer (DS)

A delay-and-sum beamformer (DS) [1] is one of the most popular steering techniques for microphone arrays. With this method, we consider the delay of each microphone to compensate for the arrival time differences of the speech signal to each microphone. Considering the dual-channel case, the noisy speech signals from two microphones are highly correlated. If the signals of two microphones are $x(n)$ and $y(n)$, the cross-correlation is defined as

$$R_{xy}(m_1, m_2) = E[x(m_1)y(m_2)] \tag{1}$$

Among these cross-correlation values, we calculate the delay which is the point of having maximum value. After each microphone's signal is compensated with this delay, the noise reduced speech signal ($\hat{s}(n)$) is synthesized by adding these time-aligned signals together.

$$\hat{s}(n) = \frac{x'(n) + y'(n)}{2} \tag{2}$$

where $x'(n)$ and $y'(n)$ are time-aligned signals.

This has the effect of reinforcing the desired speech signal while the unwanted off-axis noise signals are combined in a more unpredictable fashion.

2.2 Griffiths-Jim Beamformer (GJ)

The Griffiths-Jim beamformer [5] is a sort of adaptive beamformer. This approach is suitable in the case of an adjacent microphones arrangement. Then, we can assume that two microphones' signals are defined as

$$\begin{aligned} x_1(n) &= s(n) + n'(n) \\ x_2(n) &= s(n) + n''(n) \end{aligned} \tag{3}$$

where $s(n)$ is the speech signal which we want to estimate and $n'(n)$, $n''(n)$ are the delayed noise signal of $n(n)$.

In this situation, because the reference microphone has high speech signal components, the standard adaptive noise canceling (ANC) method cannot work well and causes speech distortion and resulting in a decrease the performance. Therefore, we need to attenuate the desired speech signal components of the reference microphone. Thus we constructed new primary and reference signals by adding and subtracting the two microphones' signals.

$$\begin{aligned} p(n) &= \frac{x_1(n) + x_2(n)}{2} \\ r(n) &= x_1(n) - x_2(n) \end{aligned} \tag{4}$$

With these two signals, we can obtain noise-reduced speech signals by adaptively reducing the noise components based on the standard ANC method.

2.3 Eigendecomposition Method (EVD)

Blind separation and deconvolution (BSD) of sources is an approach taken to estimate original source signals only the information of mixed signals observed in each input channel.

The Eigendecomposition method is a sort of multi-channel signal separation approach. The Multi-channel signal separation approach has been widely studied recently. Cao et al. [3] have proposed a dual-channel speech separation method based on eigendecomposition. The procedure of this method is as follow.

The method processes speech utterances on a frame by frame base and the correlation matrix of each microphone signal is calculated as

$$R_{Y_i} = Y_i Y_i^T \quad for \quad i = 1, 2 \tag{5}$$

where

$$Y_i = \begin{bmatrix} y_i(m) & y_i(m+1) & \cdots & y_i(m+N-1) \\ y_i(m+1) & y_i(m+2) & \cdots & y_i(m+N) \\ \vdots & \vdots & \ddots & \vdots \\ y_i(m+p-1) & y_i(m+p) & \cdots & y_i(m+p+N-2) \end{bmatrix} \tag{6}$$

and Y_i is the data matrix, N represents the length of a frame, y_i is the i-th microphone signal and the number of rows p should be selected to satisfy $P \geq 2k$, where k is the number of undesired signal components.

Then, the new ratio matrix R_{ratio} is built from the correlation matrices.

$$R_{ratio} = R_{Y_1}^{-1} R_{Y_2} \tag{7}$$

Compute the eigenvectors and eigenvalues by the decomposition of R_{ratio} and find $\lambda_{\min}$ (or $\lambda_{\max}$) as well as the corresponding eigenvector v_1 (or v_2).

Finally, design an Infinite Impulse Response (IIR) filter according to the following equation:

$$H(z) = B(z) / A(z) \tag{8}$$

where

$$B(z) = v_0 + v_1 z^{-1} + \cdots + v_{p-1} z^{(p-1)}$$
$$v = \begin{bmatrix} v_0 & v_1 & \cdots & v_{p-1} \end{bmatrix} \ : \ eigenvector \tag{9}$$

and $A(z)$ can be constructed to improve the frequency response of the eigen filter.

By using this IIR notch filter, we can obtain the noise-reduced speech signal by applying the filter to the noisy speech signal.

$$\hat{s}(n) = h(n) * y_1(n) \tag{10}$$

where *represents the convolution operator.

2.4 Enhancement of the Eigendecomposition Method (NGJ)

In this section, we propose the enhancement of the eigendecomposition method. Eigenvectors and eigenvalues of the R_{ratio} have the following two properties.

1) Signal subspace spanned by eigenvectors of the R_{ratio} coincide with a subspace spanned by the eigenvectors of R_{Y_1} and R_{Y_2}.
2) Eigenvalues of the R_{ratio} are equal to the ratio of corresponding power densities for each signal component of the two signals.

The above results can be used to separate the components $s_1(n)$ and $s_2(n)$ if the signal-to-noise ratios (SNRs)—one of s_1 and s_2 are assigned as signal and the other as noise—in the two observations are different. But if the two microphones' signal components are

similar which is the case when two microphones are arranged with similar positions or the power of the signals are similar, then the above eigendecomposition method cannot perform appropriately.

In order to enhance the noise reduction performance, we constructed new two signals similar to the Griffiths-Jim Beamformer,

$$y_1'(n) = \frac{x_1(n) + x_2(n)}{2}$$
$$y_2'(n) = x_1(n) - x_2(n) \tag{11}$$

where $x_1(n)$ and $x_2(n)$ are time-aligned microphone signals which are compensated with the time delay between two microphones.

With the front-end processing of two signals, we applied the eigendecomposition method to enhance the noisy speech signals.

2.5 High-Pass Filter Method (HP)

Generally driving car noise is dominated by low frequency components from air flow and tire noise. That is, car noise typically has its peak power between 100-800 Hz, depending on driving conditions and the car. Also, below 1 kHz the noise spectrum level decreases to higher frequencies by 6 dB/octave, whereas above 1 kHz the spectrum level decreases faster by about 12 dB/octave [4]. Unfortunately, the power spectrum of speech exhibits a very similar behavior. Hence, a complete separation of noise and speech may not always be achieved. But by reducing the dominant low frequency components of noise signal, we can obtain a more clean speech signal. To reduce the low frequency components of noise signal, we applied a high-pass filter to the each of microphone signals.

3. Experiments

In this section, we perform various representative experiments to make the performance comparisons of the proposed candidate methods discussed in Section 2.

3.1 Experimental Condition

To show the effectiveness of the proposed noise reduction methods, isolated speech recognition experiments were performed. The experiments were conducted using the

CAR01 corpus from the Speech Information Technology & Industry Promotion Center (SITEC) [6]. This corpus consists of car control and navigation related command words, isolated digit and four connected digits. The speech samples were simultaneous recorded using 8 channels (microphones) with the speed of 80 km/h in real car environments and sampled at 24 kHz but we re-sampled into 16 kHz.

In our experiments, we used the recordings associated with channel 3 (placed at the left-end of the sun visor) and channel 5 (placed at the right-end of the sun visor) for the near microphone's arrangement and channel 4 (placed at the center of the sun visor) and channel 7 (placed at the safety belt) for the distant microphone's arrangement. In the experiments, each speech signal was parameterized using the 12 Mel- Frequency Cepstral Coefficient plus log-energy, and their first and second derivatives. Acoustic phoneme models produced 3-state left-right continuous density Hidden Markov Models (HMMs) with 8 Gaussian mixtures per state. The corpus is divided into two different groups: train and test. The train set consists of 4,384 samples associated with channel 1 (placed at close-talk head-worn) uttered by 80 speakers. The test set consists of 1,096 samples associated with channels 3, 4, 5 and 7 uttered by 20 speakers. The experiments were conducted using the HMM Toolkit (HTK) recognizer [7]. The performance of noise reduction methods is evaluated by word accuracy rate for speech recognition.

3.2 Experimental Results

We performed various noise reduction experiments using the proposed method. As the baseline experiment, a general speech recognition experiment without any noise reduction method was conducted first to compare the performance of the proposed methods. The results obtained are shown in Table 1. As can be seen, Table 1 shows the performance in word accuracy for each microphone channel. From this result, it is shown that channel 1 (placed at close-talk head-worn) is superior to any other channels. These results show that the performance is affected by distance and direction from the speaker and the microphone.

Table 1. Recognition results without any noise reduction for each microphone channel

Channel	1	2	3	4	5	6	7
Word accuracy(%)	94.43	60.49	38.32	55.57	38.23	3.56	83.94

In the following experiment, we applied a single channel noise reduction method to compare the performance improvement of the dual-channel noise reduction method and the baseline system. Among various single channel noise reduction methods, we used the

spectral subtraction (SS) method based on minimum statistics. Table 2 shows the recognition results of SS over each microphone channel. Comparing this result with Table 1, it shows that the SS method attains a performance improvement but a more robust noise reduction method is needed to apply to real environments.

Table 2. Word accuracy for single channel noise reduction for each microphone channel

Channel	1	2	3	4	5	6	7
Word accuracy(%)	91.16	72.72	57.66	67.06	55.20	10.31	85.77

In dual channel noise reduction methods, we used two microphone displacement settings such as the recording associated with the channel pair 3 and 5 and channel pair 4 and 7.

In the following experiments, we applied the dual channel noise reduction method using the Delay-and-Sum beamformer (DS) and Griffiths-Jim beamformer (GJ). Also we applied the high-pass filter (HP) to cancel the low frequency component of the car environment. The cutoff frequency of the HP was 240 Hz. The experimental results are shown in Table 3. Comparing these results with Table 1 and Table 2, it shows that the performance is highly improved and the HP's noise reduction efficiency is very good. Also the performance improvement of channel pair 3 and 5 is higher than that using the single channel noise reduction method.

Table 3. Word accuracy for dual channel noise reduction using the DS and GJ methods (%)

Method / Channel	DS	HP+DS	HP+GJ
Ch 3 + Ch 5	46.90	84.95	83.03
Ch 4 + Ch 7	82.85	90.97	88.23

In the following experiment, we applied the eigendecomposition (EVD) method. The results are shown in Table 4. As shown, the performance is significantly better than the results of other previous experiments. In the case of channel pair 3 and 5 , the performance of the eigendecomposition method is lower than that of the previous SS or DS. But the performance of the EVD method combined with the HP and NGJ (two new reconstructed signals) method is highly improved. From the results, it is shown that the best performance is obtained when EVD method is combined with the HP and NGJ together. For the case of channel pair 4 and 7, the performance of the EVD is similar to that of EVD combined with the HP or NGJ. But for the case of channel 3 and 5 pair, the

performance of EVD method combined with HP and NGJ is superior to that of EVD or EVD with the HP or NGJ. This is because the case of channels 4 and 7, the microphone's poison is apart from each other and the each microphone signal is somewhat different. On the other hand, for the case of channels 3 and 5 the position of microphones is close to each other and the distance of between the speaker and microphones is similar, thus the two microphone signals are very similar.

Table 4. Word accuracy for dual channel noise reduction using the EVD method (%)

Method / Channel	EVD	HP+EVD	NGJ+EVD	HP+NGJ+EVD
Ch 3 + Ch 5	63.41	83.94	78.56	89.14
Ch 4 + Ch 7	89.96	91.61	89.69	91.88

Table 5 shows the overall experimental results. The best performance of each noise reduction method shown in the table. From the results, the proposed method performs better than other noise reduction methods. These results demonstrate that the dual channel noise reduction method using the proposed approach outperforms that of the baseline system and single channel noise reduction method.

Table 5. Word accuracy performance comparison of various noise reduction methods (%)

Method / Channel	baseline	SS	HP+GJ	HP+DS	HP+NGJ+EVD
Ch 3 + Ch 5	38.32	57.66	83.03	84.95	89.14
Ch 4 + Ch 7	83.94	85.77	88.23	90.97	91.88

4. Conclusions

In this paper, we presented an effective dual-channel noise reduction method based on eigendecomposition. In particular, we compared the SS and various dual-channel noise reduction methods for robust speech recognition. Also, we experimented with a real multi-channel car database and compared the results with respect to the microphone's arrangement. From the analysis and experimental results, we showed that the enhanced eigendecomposition method combined with a high-pass filter indeed significantly improved speech recognition performance under a dual-channel environment. By applying the proposed robust noise reduction method, the speech recognition system outperforms that of the standard EVD method by a factor of up to 70%. Future study will be continued to

further increase the performance and to find more robust noise reduction methods.

References

[1] Huang, X., Acero, A. and Hon, H., 2001. *Spoken Language Processing*, Prentice Hall PTR.
[2] Boll, S. F., 1979. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *ASSP-27*, No.2, pp. 113-121, April.
[3] Cao, Y., Sridharan, S., and Moody, M., 1997. "Multichannel Speech Separation by Eigendecomposition and Its Application to Co-talker Interference Removal," *IEEE Transactions on Speech and Audio Processing*, Vol.5, No.3, pp. 209-219, May.
[4] Campbell, D. R., and Shields, P. W., 2003. "Speech enhancement using sub-band adaptive Griffiths-Jim signal processing," *Speech Communication 39*, pp. 97-110.
[5] Aubauer, R., and Leckschat, D., 2001. "Optimized second-order gradient microphone for hands-free speech recordings in cars," *Speech Communication 34*, pp. 13-23.
[6] http://www.sitec.or.kr
[7] Young, S., Evermann, D., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2002. *The HTK Book (for HTK Version* 3.2).

Received: February 10, 2004
Accepted: March 10, 2004

▲ Sungjoo Ahn
Department of Electronics and Computer Engineering, Korea University
5Ka-1, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea
Tel: +82-2-927-6115 (O), H/P: 011-9099-0258
Fax: +82-2-3291-2450
e-mail: sjahn@ispl.korea.ac.kr

▲ Sunmee Kang
Department of Computer Science, Seokyeong University
16Ka-1, Chongnung-Dong, Sungbuk-ku, Seoul, 136-704, Korea
Tel: +82-2-940-7291 (O), H/P: 011-9760-7144
Fax: +82-2-919-0345
e-mail: smkang@skuniv.ac.kr

▲ Hanseok Ko
Dept of Electronics and Computer Engineering, Korea University
5Ka-1, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea
Tel: +82-2-3290-3239 (O), H/P: 011-9001-3239
Fax: +82-2-3291-2450
e-mail: hsko@korea.ac.kr