Blind Separation for Real-World Speech Signals

KIYOTOSHI MATSUOKA, SEIJI YAMADA, MASAFUMI MATSUNO
Department of Brain Science and Engineering
Kyushu Institute of Technology
Hibikino, Wakamatsu, Kitakyushu

TAKAYOSHI YAMAMOTO
Kyushukyohan Co. Ltd.
Kogane, Kokurakita, Kitakyushu
JAPAN

Abstract: - Among potential applications of blind source separation (BSS) a most promising one might be separation of speech signals. In real-world situations, however, any BSS algorithm for sound signals suffers from a difficulty. From a practical point of view the microphone array should be made compact, but then the mixing matrix becomes almost singular, inducing certain instability in the algorithm execution. This paper describes some experiments of BSS, which were made in a soundproof room, an office room, and a car. The results show that an appropriate configuration between the microphone set and the sound sources is very important to achieve satisfactory separation. Moreover, astonishingly, if the microphones are located appropriately, even using only two microphones considerably enhances a target sound from mixtures of more than two sounds.

Key-Words: - blind separation, independent component analysis, speech, voice, noise cancelling, minimal distortion

1 Introduction

Blind source separation (BSS) is a method for recovering a set of statistically independent signals from the observation of their mixtures without any prior knowledge about the mixing process. It has been receiving a great deal of attention from various fields as a new signal processing method. Among conceivable applications of BSS the most promising one might be separation of speech signals. For example, in a situation where a speaker is surrounded by a noisy environment, the target voice can be extracted by using a BSS technique. Indeed, in the literature of BSS one can see a lot of papers that report experiments on speech signal separation.

In our experience, however, although conventional methods for BSS can perform separation for artificially synthesized data, they do not necessarily work well for real-world data [4][5]. Separation accuracy is often unsatisfactory and, what is worse, they sometimes suffer from incomprehensible instability. The following can be considered as a reason.

The task of BSS is basically to find the inverse of the mixing matrix and to apply it to the observation. In practical applications the microphone set should be made compact, but then the mixing matrix becomes almost singular particularly for a low-frequency range. If the separator is constructed by an FIR filter for such a nearly singular mixing process, a large number of taps or parameters (say, some hundred or thousand taps) become necessary. It is not easy to determine such a large number of parameters in a reliable manner. Actually we sometimes face the following phenomenon. When applying an iterative BSS algorithm to a given data, in the beginning the algorithm appears to behave in a desired manner, but as the iteration goes on, some instability appears suddenly.

Investigating the result obtained in such a situation, we usually find that the following thing has occurred [4]. Some frequency components of a source appear at an output terminal of the separator while other frequency components originated from the same source appear at a different terminal. This phenomenon is probably due to a too high frequency resolution when a long-length filter is adopted. Time-domain BSS algorithms are usually thought to be relatively free from this kind of permutation problem as compared to frequency-domain approaches, but it is not necessarily the case.

If adversely the algorithm is performed with a shorter filter length, the stability can considerably be improved, but the accuracy of separation might be unsatisfactory. In this paper, by means of some experiments we show that this trade-off problem can be solved (relaxed) by devising an appropriate configuration of the microphones and the sound sources. That is, we can reduce the length of the separating filter without degrading the separation accuracy so much.

It is sometimes thought that a long filter length is required to cope with a long reverberation time. But the result shows that a separating filter with a considerably short filter length is able to achieve separation and it enhances robustness and stability of the algorithm.

In this paper we experimentally show what kind of configuration is appropriate for BSS of speech signals. The experiments were performed in a sound-proof room, a usual office room, and a car. The following parameters were changed:

(P1) Distance between the pair of microphones and the sound sources,
(P2) Distance between the two microphones,
(P3) Direction of the pair of microphones relative to the sound sources,
(P4) Number of sound sources.

As for the last point (P4), it is usually considered that the number of sensors needs to be greater than that of sources, but it is not necessarily the case. The result of the experiment shows that, if the microphone pair is appropriately located relative to the sound sources, a small number of microphones are enough to extract a single sound though separation is not complete, of course.

All the experiments described in this paper were made with a device using a digital signal processor (DSP), which can perform separation in a real-time manner. Although only an algorithm was tested in those experiments, we believe that the results suggested many things about actual implementation of BSS.

The paper is organized as follows. In the next section we describe a BSS algorithm and a device using a DSP. In section 3 we show the results obtained by the experiments performed various environments and settings. Section 4 concludes the paper.

2 A BSS Algorithm

In this section we describe the algorithm used for the experiments briefly. As usual, the mixing process is assumed to be given by the following linear equation:

$$\mathbf{x}(t) = \sum_{\tau=0}^{\infty} \mathbf{A}_\tau \mathbf{s}(t-\tau) = \mathbf{A}(z)\mathbf{s}(t), \qquad (1)$$

where $\mathbf{s}(t) = [s_1(t),\ldots,s_N(t)]^T$ represents a set of statistically independent signals, which is referred to as source signals, and $\mathbf{x}(t) = [x_1(t),\ldots,x_N(t)]^T$ is a set of observations. The process is represented by a transfer function matrix $\mathbf{A}(z)$. The task of BSS is to estimate the sources only from the observed signals in the absence of any knowledge about the sources $\mathbf{s}(t)$ and the mixing process $\mathbf{A}(z)$.

The process by which the sources are recovered is given by

$$\mathbf{y}(t) = \sum_{\tau} \mathbf{W}_\tau \mathbf{x}(t-\tau) = \mathbf{W}(z)\mathbf{x}(t), \qquad (2)$$

where $\mathbf{y}(t) = [y_1(t),\ldots,y_N(t)]^T$ represents a set of estimated source signals. We call the matrix $\mathbf{W}(z)$ the separating filter or the separator. If the mixing matrix $\mathbf{A}(z)$ is known beforehand, the problem is easy to solve; we have only to determine the demixing matrix as $\mathbf{W}(z) = \mathbf{A}^{-1}(z)$, leading to $\mathbf{y}(t) = \mathbf{s}(t)$.

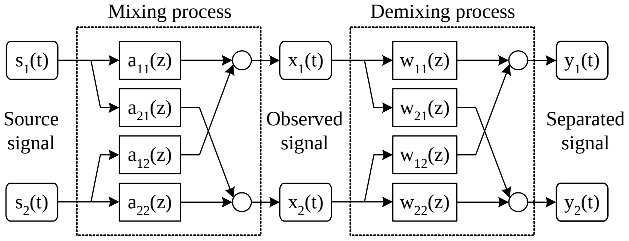


Fig.1 The structure of blind source separation (N = 2)

In real-world situations, however, it is usually difficult to find the mixing process beforehand. In BSS, mixing matrix $\mathbf{A}(z)$ or its inverse is identified by using the only assumption that the sources are statistically independent of each other. Namely, if one determines $\mathbf{W}(z)$ so that $y_1(t),\ldots,y_N(t)$ be statistically independent, then one can find the desired separating filter.

The algorithm used for the present experiment is the one propose by one of the authors [3]. It is a completely time-domain algorithm. The separator's output is given by

$$\mathbf{y}(t-L_1) = \sum_{\tau=-L_1}^{L_2} \mathbf{W}_\tau(t)\mathbf{x}(t-L_1-\tau), \qquad (3)$$

and the coefficient matrices, $\mathbf{W}_\tau(t)$, are updated by

$$\begin{aligned}\Delta\mathbf{W}_\tau(t) = &-\alpha\{\varphi(\mathbf{y}(t-L_3))\mathbf{u}^T(t-L_3-\tau)\\ &-\operatorname{diag}\varphi(\mathbf{y}(t-L_3))\mathbf{V}(t-L_3-\tau)\}\\ &-\beta\{\operatorname{diag}(\mathbf{y}(t-L_3)-\mathbf{x}(t-L_3))\mathbf{V}(t-L_3-\tau)\}\end{aligned} \qquad (4)$$

$$\mathbf{u}(t-L_0) = \sum_{r=-L_1}^{L_2} \mathbf{W}_r^T(t)\mathbf{y}(t-L_0+r) \qquad (5)$$

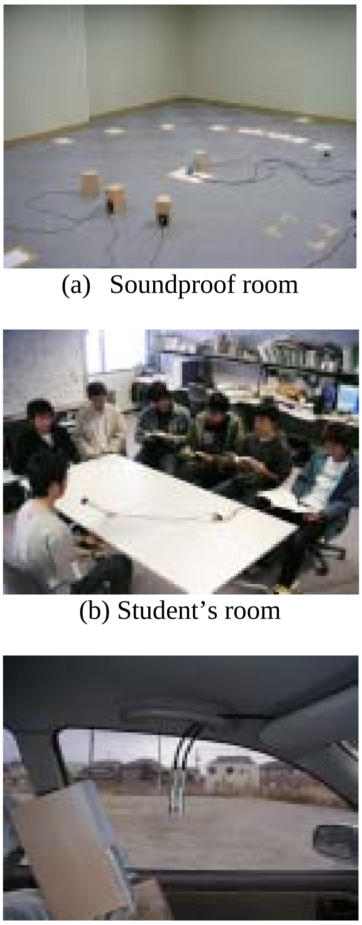
$$\mathbf{V}(t-L_0) = \sum_{r=-L_1}^{L_2} \mathbf{y}(t-L_0+r)\mathbf{W}_r(t), \qquad (6)$$

where $L_3 = 2L_1 + L_2$, $L_0 = L_1 + L_2$ and α, β are learning coefficients. In the experiments they were set as $L_1 = 8$, $L_2 = 28$, $\alpha = 1.0\text{x}10^{-3}$, and $\beta = 1.0\text{x}10^{-7}$. Function φ is defined as $\varphi(y) = 1$ for $y \ge 0$ and $\varphi(y) = -1$ for $y < 0$, taking into account that voice signals are super-Gaussian.

The first term in the right-hand side of eqn (4) is to achieve separation; it evaluates (non-Gaussian) cross-correlation between $y_1(t),\ldots, y_N(t)$ [1][2]. On the other hand, the second term is to satisfy the minimal distortion principle (MDP) proposed by Matsuoka et al.[3]. MDP is a particular constraint for $\mathbf{W}(z)$ to eliminate the scaling or filtering indeterminacy. It is derived based on the following idea.

As known well, in BSS the definition of the sources has certain indeterminacy; a source signal transformed by any linear filter can also be considered a source signal. Corresponding to the indefiniteness of the sources, the choice of the separator has certain arbitrariness. In this respect one of the authors proposed an idea: "Among the feasible separators, choose the one that best preserves the quality of the signals observed at the sensors (microphones)." If the separator is determined under this constraint, the output of the separator becomes equivalent to the signals that the microphones would observe in the absence of interfering.

The above algorithm was mounted on a DSP. A DSP is a high-speed processor for the real-time processing of sound, image, and others. It is good at a product-sum type operation required for the BSS calculation. The sampling rate and the quantization level were 8kHz and 16 bits, respectively.



(a) Soundproof room

(b) Student's room

(c) Inside of a car

Fig.2 Three environments for the experiments

3 Experiments and Results

The experiments were performed in three environments: see the views in Fig. 2 (A) a soundproof room, (B) an office room (student's room), (C) a car. In every experiment the number of the microphones used was two. Fig.3 is a sample data of voices used in an experiment. Since voice signals contain silent periods intermittently, it is easy to find from the separation data whether separation has been successful or not.

In what follows we describe many experiments made in various conditions, but we shall only show qualitative results. At present it seems to be neither important nor interesting to show quantitative data. Rather, we believe it is more important to see all the results as a whole and find some common tendencies.

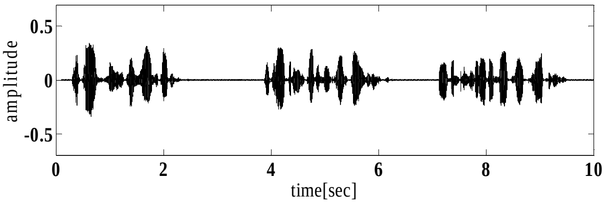


Fig.3 A sample of voice signals

The algorithm used for the present experiment is only one that is shown in the last section. However, we think similar properties will be found even if other algorithms are used.

A. Experiments in a soundproof room

The following experiments were performed in a soundproof room. Note that the reverberation time of this room is shorter than that of a usual office room.

(A1) Two sources: altering the distance between the microphones and the sound sources

In this experiment two microphones and two loudspeakers were put in line as shown in Fig.4. Loudspeaker 1 emitted a male voice and Loudspeaker 2 a female voice. The distances of the loudspeakers from the pair of microphones were either of (1) $d_1 = d_2 = 0.1$m, (2) $d_1 = d_2 = 1.0$m, (3) $d_1 = d_2 = 2.0$m. In every case separation was successful, but the separation accuracy became lower as the distance became longer. As an example, the result obtained in the worst case (3) is shown in Fig.5. The bottom in Fig.5(b) corresponds to a voice in Fig.4. [Henceforth we sometimes say "separation was successful." It means that the separation result was as good as or better than that in Fig.5. To save space, the source signals will not be shown.]

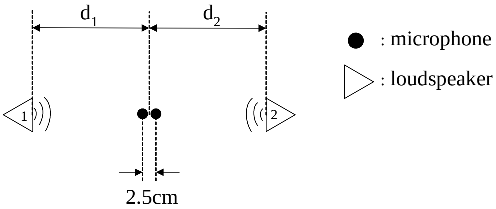
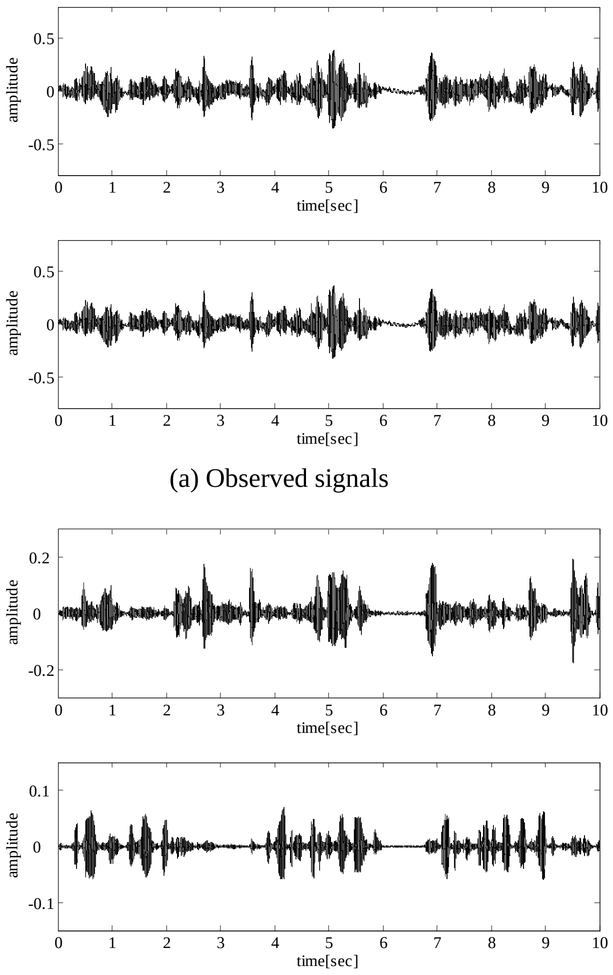


Fig.4 Setup of experiment (A1)

An important result of this experiment is that separation can be made with the pair of microphones placed a very short distance (2.5cm) apart. It implies that a small portable device for BSS can be made.



(a) Observed signals

(b) Separated signals

Fig.5 Separation result of (A1-3).

(A2) Two sources: altering the distance between the two microphones

In this experiment the distance between the two microphones were changed as (1) d = 2.5cm, (2) d = 10cm, (3) d = 50cm, (4) d = 90cm, while the position of the loud speakers were fixed; see Fig.6.

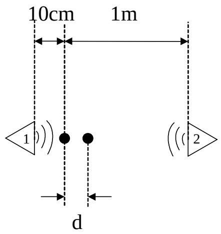


Fig.6 Setup of experiment (A2)

As the distance between the microphones became longer, separation performance became worse. Particularly in the cases of (3) and (4), separation was unsuccessful, i.e., no significant enhancement of the target sound was found. This result might be connected with spatial aliasing that could occur when the distance between the microphones is too long.

(A3) Two sources: altering the direction of a loudspeaker

In this experiment a loudspeaker was placed in various directions as (1) θ= 0deg, (2) θ= 90deg, (3) θ= 100deg, (4) θ= 120deg, (5) θ= 135deg; see Fig.7. When θ was around 100deg or larger, separation was unsuccessful.

Particularly in the cases of (4) and (5), the parameters of the separating filter and hence the output diverged to infinity. It is probably because the mixing matrix became nearly singular for large θ. The pair of microphones should be set so that θ be small, as long as it is possible.

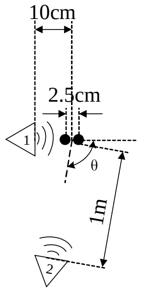


Fig.7 Setup of experimental (A3)

(A4) Three sources: altering the position of the loudspeakers

In this experiment the number of sound sources was three; see Fig.8. Henceforth we refer to the voice of loudspeaker 1 as the target voice and to the voices of other loudspeakers as interfering sounds. The task is to separate or enhance the target voice.

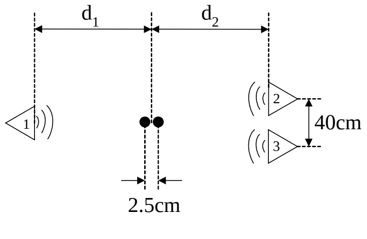


Fig.8 Setup of experiment (A4)

The positions of the loudspeakers were varied as (1) d_1 = d_2 = 0.1m, (2) d_1 = 0.1m, d_2 = 1.0 m, (3) d_1 = 1.0m, d_2 = 0.1m, (4) d_1 = 1.0m, d_2 = 1.0m. In every case, separation was succeeded though it was not complete, of course. Among the four conditions, case (1) gave the best result.

One might think that it is impossible to extract a voice signal because the inverse process of the mixing one does not exist in the case that the number of sources is larger than that of the microphones. The experiment shows a result to the contrary. The reason is as follows.

In the present case the mixing process can be expressed by

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} a_{11}(z) & a_{12}(z) & a_{13}(z) \\ a_{21}(z) & a_{22}(z) & a_{23}(z) \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \end{bmatrix} \quad (7)$$

If the locations of sources 2 and 3 are close to each other and therefore it can be considered to satisfy

$$a_{12}(z) \approx a_{13}(z) \triangleq \tilde{a}_1(z)\,,\ a_{22}(z) \approx a_{23}(z) \triangleq \tilde{a}_2(z)\,. \quad (8)$$

Then, eqn (6) can be approximated as

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \approx \begin{bmatrix} a_{11}(z) & \tilde{a}_1(z) \\ a_{21}(z) & \tilde{a}_2(z) \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t)+s_3(t) \end{bmatrix}. \quad (9)$$

This implies that the mixing process is actually a two-input, two-output process, enabling $s_1(t)$ to be separated though it is not complete, of course.

Another interesting finding was that in the other output of the separator the target voice is eliminated almost perfectly. This comes from the fact that there is a transformation eliminating one source component exactly (while there is no transformation extracting precisely one component). For instance, suppose the following form of separator:

$$\mathbf{W}(z) = \begin{bmatrix} * & * \\ a_{21}(z) & -a_{11}(z) \end{bmatrix} \quad (10)$$

Then the output of the separator becomes

$$\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} * & * \\ a_{21}(z) & -a_{11}(z) \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

$$= \begin{bmatrix} * \\ (a_{21}(z)a_{12}(z) - a_{11}(z)a_{22}(z))s_2(t) \\ +(a_{13}(z)a_{21}(z) - a_{11}(z)a_{23}(z))s_3(t) \end{bmatrix} \quad (11)$$

This implies that there exists a separator that cancels out the component of $s_1(t)$ perfectly.

(A5) Four sources: altering the positions of the loudspeakers

In this experiment four loudspeakers were put as shown in Fig.9; (1) d_1 = d_2 = 0.3m, (2) d_1 = 0.3m, d_2 = 1.0m, (3) d_1 = 1.0m, d_2 = 0.3m. The orientations of loudspeakers 2, 3, 4 with respect to the pair of microphones are 45deg, 0deg, -45deg. In all experiments separation was successful. The reason of why separation is possible may be the same as shown in (A4).

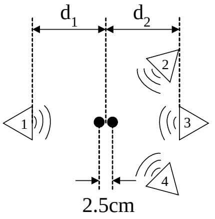


Fig.9 Setup of experiment of (A5)

(A6) Six sources

Six loud speakers were put as shown in Fig.11. The orientation of five microphones on the right side was ± 45deg, ± 22.5deg, 0deg. The result of separation is shown in Fig.11. The separation was, of course, not complete but a clear enhance of the target voice was observed.

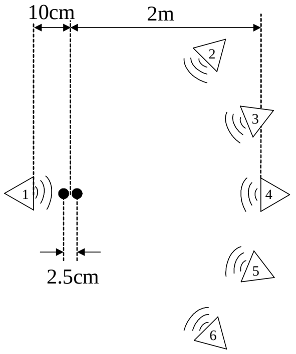
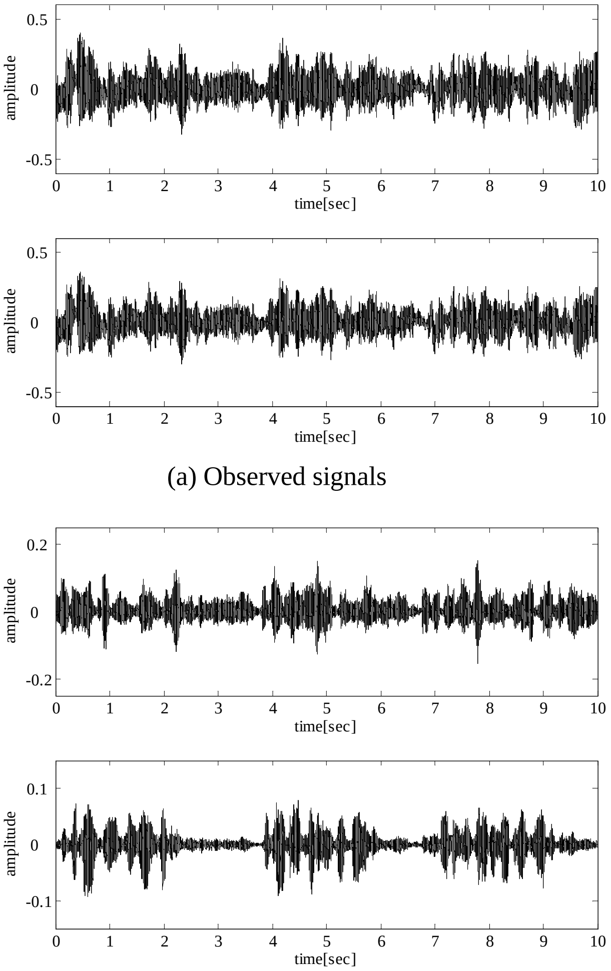


Fig.10 Setup of experiment (A6)



(a) Observed signals

(b) Separated signals. The bottom corresponds to the target voice.

Fig.11 Separation result of (A6)

B. Experiments in a student's room

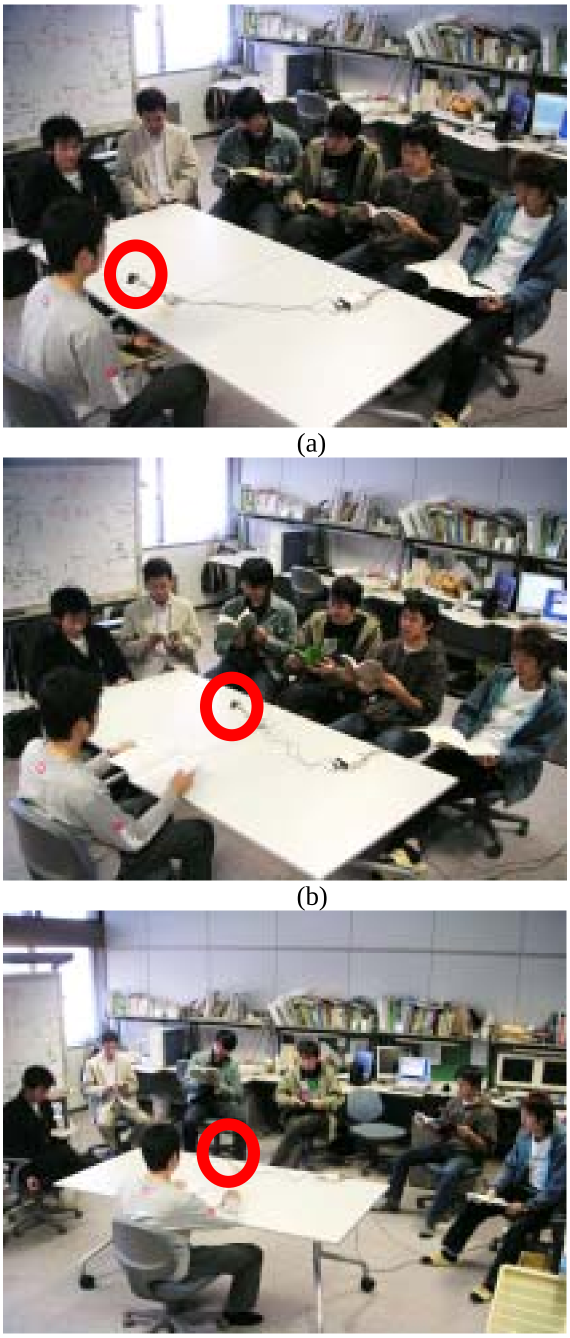
In this experiment totally seven people spoke simultaneously in a student's room; one target voice and six interfering voices. The room has a longer reverberation time than the soundproof room. A pair of microphones adjacently located 1.0cm apart was put between the target source and the interfering sources. The experiment was made under four conditions; see Fig.13.

(B1) The target source is 'near' the microphone pair; the interfering sources are 'far' from it;

(B2) 'intermediate', 'intermediate';

(B3) 'intermediate', 'far';

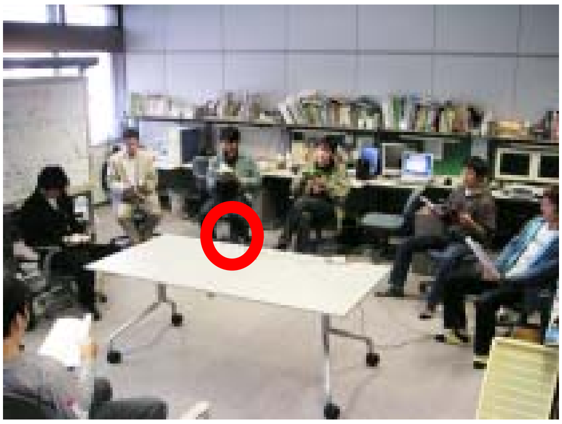
(B4) 'far', 'far'.



(a)

(b)

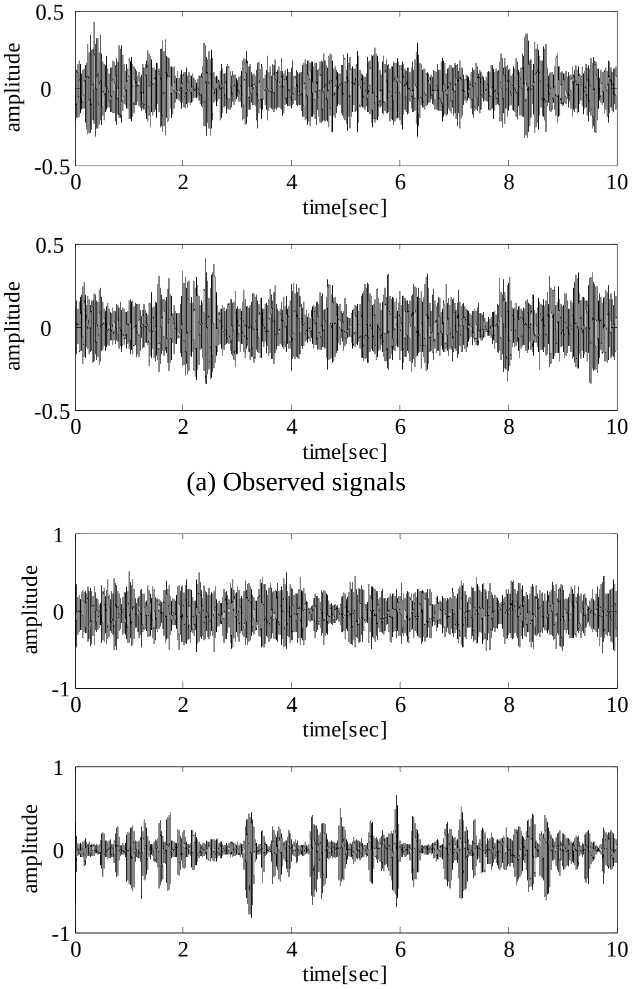
(c)



(d)

Fig.12 Scenes of four experiments in a student's room

Separation performance was the best in (B1) and the worst in (B4); the results of (B2) and (B3) were similar. In Fig.13 the result of (c) is shown; it can be seen that the target voice was enhanced considerably though only two microphones were used. Also it should be noted that the two microphones were located very closely (1cm).



(a) Observed signals

(b) Separated signals: the bottom corresponds to the target voice.

Fig.13 The result of (B3)

C. Experiments in a car

This experiment was made in a car running at about 60km/h. A loudspeaker was set on the passenger seat and a pair of microphones was put in front with a variety of configuration. An A-weighting filter was applied to sound signals detected by the microphones.

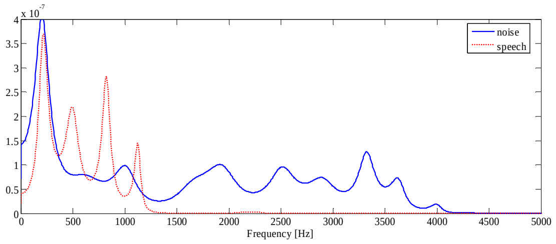


Fig.14 Power spectra of a voice and noise in a car

In the car there were a lot of noises: road noise, engine noise, wind noise, noise from the air conditioner, sounds from a radio, etc. Fig.14 shows the power spectra of a voice signal and noise in a moving car. The frequency ranges of the two spectra overlap with each other, implying that it is impossible to separate the target voice by just applying a band-pass filter.

(C1) Putting the microphones closely to the loudspeaker in line

In this experiment the microphones were put as shown in Fig.15. Noise environment was either of (1) voice + road noise (including engine noise); (2) voice + road noise + air conditioner noise; (3) voice + road noise + wind noise (in which a window was opened); (4) voice + road noise + radio sound.

In condition (1), separation was almost perfect. Also in (2) and (4), separation was successful, but accuracy was somewhat worse than in (1). On the other hand, in condition (3) separation was totally unsuccessful. It is probably because the mixing process fluctuates very rapidly and therefore the adaptive BSS algorithm could not follow it.

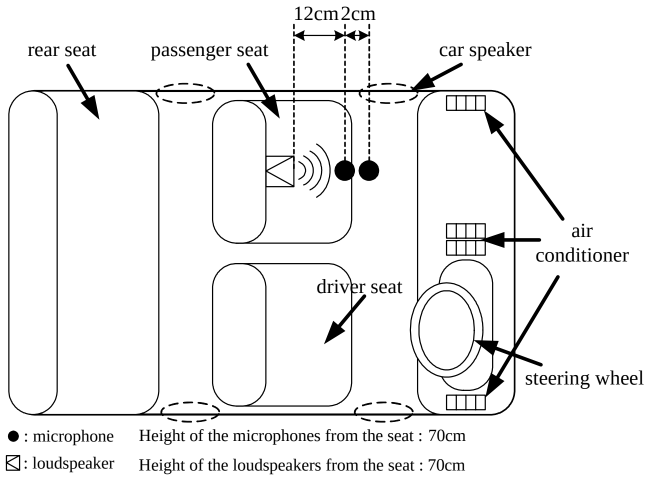


Fig.15 Setup of experiment (C1)

(C2) Putting the microphones closely to the loudspeaker out of line

The experiment was made under the same four conditions as before, but with a different direction of the microphone pair. In this case separation was unsuccessful in every condition. It indicates that how to set the direction of the microphone pair is very important for separation.

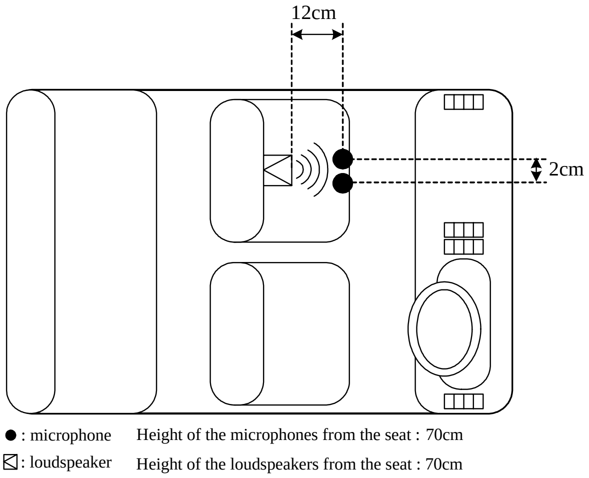

12cm
2cm
● : microphone
Height of the microphones from the seat : 70cm
: loudspeaker
Height of the loudspeakers from the seat : 70cm

Fig.16 Setup of experiment (C2)

(C3) Putting the microphones far from the loudspeaker in line

The experiment was made under the same four conditions as in (C1) but with a different position of the microphone pair. In this setting separation performance was similar in (C1). A remarkable thing is that even in the case (3) separation was done to some extent though the reason is unknown.

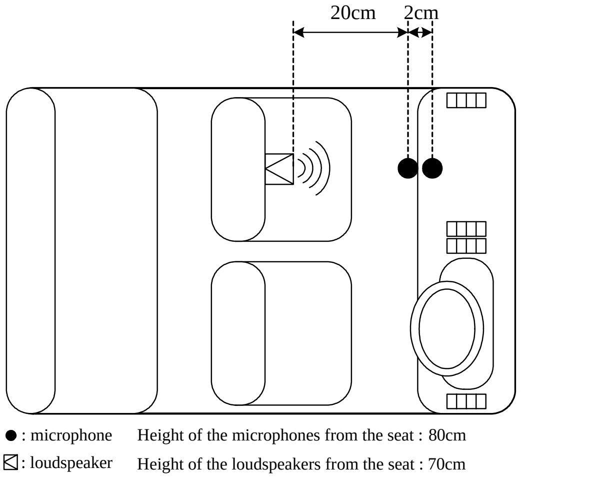

20cm
2cm
● : microphone
Height of the microphones from the seat : 80cm
: loudspeaker
Height of the loudspeakers from the seat : 70cm

Fig.17 Setup of experiment (C3)

(C4) Putting the microphones far from the loudspeaker out of line

The experiment was again made under the same four conditions as before. Only in the case of (1) separation was successful.

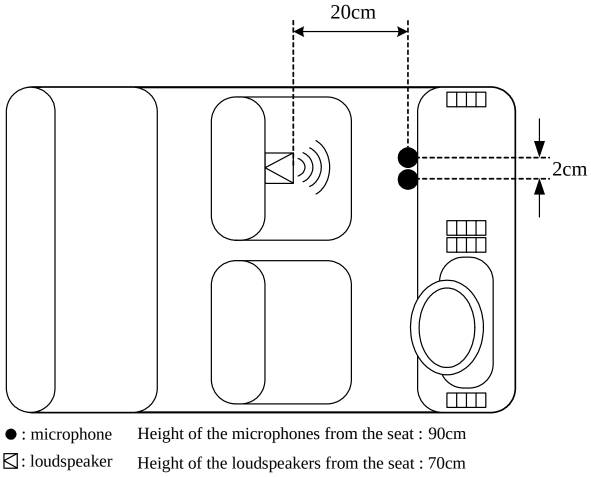

20cm
2cm
● : microphone
⧅: loudspeaker
Height of the microphones from the seat : 90cm
Height of the loudspeakers from the seat : 70cm

Fig.18 Setup of experiment (C4).

(C5) Putting the microphones far from each other
The microphones were put far from each other as shown in Fig.20; (1) d = 25cm, (2) d = 90cm. In this case only the road noise was considered. The separation was completely unsuccessful. One might think that to make longer the distance between the microphones helps improve the separation performance, but this result shows that it is not so.

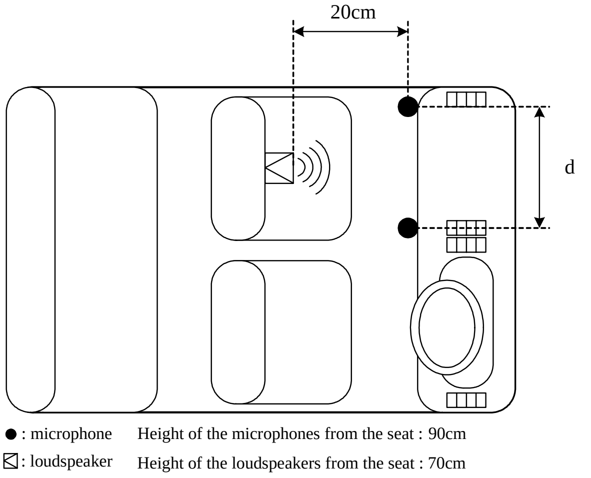

20cm
d
● : microphone
⧅: loudspeaker
Height of the microphones from the seat : 90cm
Height of the loudspeakers from the seat : 70cm

Fig.19 Setup of experiment (C5).

4 Conclusion

We have developed a device for blind separation of voice signals, using a DSP, and tested its performance in various situations (a sound-proof room, an office room, and a car). The main result is that, if the microphone pair is placed appropriately, a relatively short length of the separating filer can achieve separation. It reduces the computation time and improves stability of the algorithm. That makes it possible to adopt a larger learning coefficient, making the convergence speed higher. Another interesting finding is that even for a larger number of the sound sources than that of the microphones, source separation can be done to a certain degree. We believe these findings are very important in practical applications of BSS to sound separation.

Acknowledgement: The authors thank Kyushu Industrial Technology Centre for financial support in 2003 and 2003.

References

[1] S. Amari, S. C. Douglas, A. Cichocki and H. H.Yang, Multichannel blind deconvolution and equalization using the natural gradient, *Proc. IEEE International Workshop on Wireless Communication*, 1997, pp. 101-104.

[2] S. Choi, S. Amari, A. Cichocki, and R. Liu, Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels, *Proc. International Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, 1999, pp. 371-376.

[3] K. Matsuoka and S. Nakashima, Minimal distortion principle for blind source separation, *Proc. International Workshop on Independent Component Analysis and Blind Signal Separation (ICA2001)*, 2001, pp. 722-727.

[4] K. Matsuoka, Y. Ohba, and S. Isogai, Blind separation of sound sources in real-world situations, *18th Congress on Acoustics*, 2004.

[5] K. Matsuoka, Independent component analysis and its application to sound signal separation, *Eighth International Workshop on Acoustic Echo and Noise Control*, 2003, pp. 15-18.