Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Sound source location for reproduction of speech signal at local spot based on its decomposition into random signals

Shouichi Takane, Koji Abe, Kanji Watanabe and Sojun Sato

Faculty of Systems Science and Technology, Akita Prefectural University
84-4 Ebinokuchi, Tsuchiya, Yurihonjo, Akita, 015-0055 Japan
e-mail: takane@akita-pu.ac.jp

PACS: 43.60Dh, 43.60Ek

ABSTRACT

An approach for the reproduction of speech signal at local spot is introduced in this paper. It is based on signal decomposition using the orthogonal basis function made from random vectors. It has some difficulties, however, in the reproduction of speech signal at the desired local spot. One of them is that the information in speech can be perceived from the synthesized signal at the point except the desired spot, although its quality is degraded due to its decomposition into random signals. As far as the target of our study is focused on the reproduction of speech signals, location of the sound sources, by which the decomposed random signals are emitted, is related to the difficulty in understanding of the contents of speech. Considering the facility of the reproduction, the sound sources are located in the same distance from the spot. However, the performance is not appropriate in that case, meaning that the information in speech can be perceived at the point around the desired spot in this case. Location of the sound sources with their distance from the spot distributed has potential to solve that problem, bringing about the further distortion in the synthesized signal at the point except the spot.

INTRODUCTION

One of the purposes for sound emission in public space is to transfer the information involved in it. Since sound wave with the audible frequency has its wavelength comparable to the objects around us, it is difficult to avoid its propagation where it is not required, due to diffraction and reflection. If the information in sound can be conveyed at the desired local spot in the sound field, the communication with sound may yield new property beyond such physical limitation. Parametric loudspeaker based on ultrasound is useful in order to satisfy such need[1, 2, 3, 4]. This can generate very narrow beam of sound propagation. However, it can limit the "direction" of sound propagation, not the local "spot." Moreover, the special hardware must additionally be constructed and installed. It must be meaningful if the abovementioned need is satisfied by using the audible sound. Flat panel loudspeaker has some effective features due to its shape and structure[5]. It is the same as the parametric loudspeaker, however, in the point that the hardware with special structure must be manufactured. The method without special instruments and applicable to the ordinary public address equipments may widely contribute to practical use.

In this paper, another approach for the reproduction of speech signal at local spot is introduced. It is based on signal decomposition using the orthogonal basis function made from the random vectors. Representation of speech signal by using this decomposition was originally proposed by Togura *et al.*[6]. Negi *et al.* applied this decomposition method into transaural reproduction[7]. They used the sound of violin as the example source signal, decomposed it with random vectors, and reproduced each of the decomposed random signals with each of the loudspeakers distributed around the listener. They reported that the music signal was well reproduced at the listener's ears, and the distorted signal was obtained at the points except the listener's ears. Taking this property into account, this decomposition method is expected to bring about the reproduction of signal at the local spot, and the secrecy of information in signal except the spot.

However, this method has some difficulties in the reproduction of speech signal at the local spot. The synthesized speech signal is greatly distorted at the point except the desired local spot, but the contents of speech can be obtained from the signal synthesized there. This prevents the secrecy of information in signal except the spot from being insured. As far as the target of this study is focused on the reproduction of speech signals, location of the sound sources, with which the decomposed random signals are emitted, is related to the difficulty in understanding of the contents of speech. Hence the relation between the reproduction performance of speech signal and the location of sound sources is investigated via computer simulation in this paper.

SIGNAL DECOMPOSITION WITH RANDOM VECTORS

The method of signal decomposition with random vectors is briefly mentioned[6] in this section.

Signal representation with orthogonal basis functions

An arbitrary signal, $s(n)$ (sample length: N), can be expressed by using the linear combination of the orthonormal basis func-

tions as follows:

$$s(n) = \sum_{k=1}^{N} w_k \phi_k(n), \tag{1}$$

where $\phi_k(n)$ is the k-th orthonormal basis function, and w_k is a weight factor for $\phi_k(n)$. In matrix form, Eq. (1) is expressed as follows:

$$\mathbf{s} = \Phi \mathbf{w}, \tag{2}$$

where a matrix involving the set of orthonormal basis functions as the row vectors is expressed by Φ , and vectors $\mathbf{w}$ and $\mathbf{s}$ consists of w_k and $s(n)$ as their components, respectively. The Discrete Fourier Transform (DFT) and its inverse (IDFT) belongs to such representation if $\phi_k(n)$ is set to $(1/N)e^{j(2\pi/N)nk}$. Togura *et al.* stated that the set of orthonormal basis functions can also be made from the random vectors[6]. A random matrix of its size $N \times N$ is generated and its row vectors are expected to be linearly-independent, meaning that the matrix Φ is full rank. The orthonormal basis functions are easily obtained by various techniques[8]. The k-th orthonormal basis function is denoted as $v_k(n)$ and an orthonormal matrix involving $v_k(n)$ as its row vectors is expressed as $\mathbf{V}$, then the Eq. (1) is rewritten as

$$s(n) = \sum_{k=1}^{N} a_k v_k(n), \tag{3}$$

and

$$\mathbf{s} = \mathbf{V}\mathbf{a}, \tag{4}$$

where $\mathbf{a}$ is a weighting vector consisting of a_k as its components.

From the Eq. (4), the signal is decomposed into random vectors by computing the weighting vector $\mathbf{a}$ as follows:

$$\mathbf{a} = \mathbf{V}^{-1}\mathbf{s}. \tag{5}$$

Signal decomposition procedures

In order to execute the signal decomposition by calculating the Eq. (5), the set of orthonormal matrix $\mathbf{V}$ must be prepared. Choosing large value of N means that the size of $\mathbf{V}$ becomes large. However, the abovementioned principle can be adopted to the decomposition of long signal with relatively small N, regarding the value of N as the frame length. Source signal is divided into frames of length N, and the decomposition is repeated frame by frame. These procedures are illustrated in Fig. 1. In this figure, a block named "Computation of weighting coefficients" indicates the procedure to compute Eq. (5). Computed weighting factors are multiplied by each basis function $v_k(n)$. After the signal was decomposed into random vectors as stated in the previous section.

The value of N can be set arbitrary, but the smaller N sometimes leads to the matrix $\mathbf{V}$ not inversible or ill-conditioned, and the randomness of $v_k(n)$ is not sufficient. In this paper, the value of N was set to 200.

Reason for selecting random vectors as orthonormal basis functions

When the reproduction of sound at the local spot is aimed at, the selection of orthonormal basis functions is important. The complex sinusoidal function can also be selected, but the human perceives the pitch and/or the tonal component from the complex tone. That may make the listener annoyed. Moreover, Togura showed that small number of sinusoidal signals is sufficient to represent the sound signals since the ordinary sound signal, especially the speech signal has its frequency spectrum somewhat concentrated at the certain frequency range[6]. This

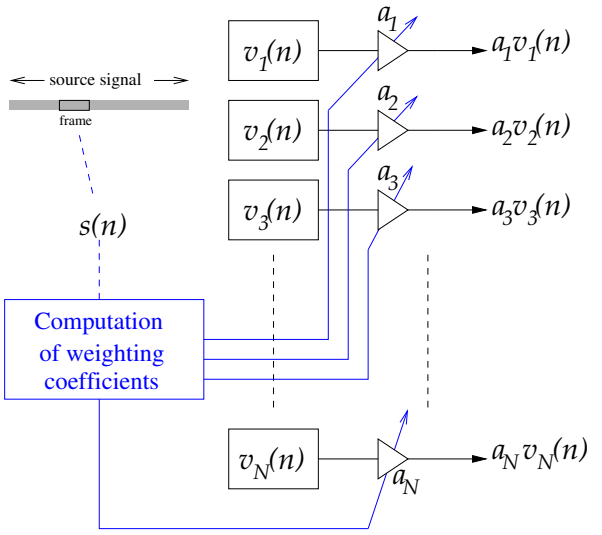


Figure 1: Illustration of signal decomposition procedures with random vectors.

means the distribution of the weighting coefficients is concentrated. On the other hand, each of the random vectors has its frequency spetrum widely spread and no concentration, hence the distribution of the weighting coefficients may have larger variance. The effect of masking is also expected so as not to hear the contents of speech.

SIGNAL SYNTHESIS WITH RANDOM VECTORS

The value of N corresponds to the number of vectors $v_k(n)$. If each of $v_k(n)$ is emitted from the loudspeaker with the corresponding weighting factor a_k, the value of N is equal to the number of sound sources. Therefore the large value of N is impractical. Negi *et al.* used four sound sources and mixed $N/4$ vectors for the input to each source[7]. In the computer simulations of this paper, the same procedures were adopted. Such processings are illustrated in Fig. 2.

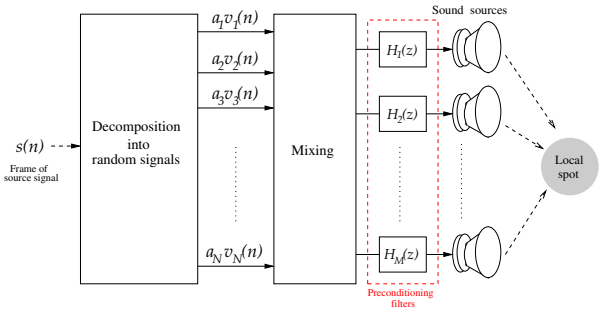


Figure 2: Illustration of reproduction system

The block named "Decomposition into random signals" in Fig. 2 is the processing depicted in Fig. 1. Number of the random vectors corresponds to the number of samples in a frame, usually exceeds the number of sound sources. Therefore the random vectors have to be mixed. The way of mixing is arbitrary, and it depends on which random vector is assigned to which sound source. In this paper, the random vectors are grouped from the first random vector $(a_1 v_1(n))$ according to the number of the sound sources (depicted as M in Fig. 2), since random vectors are almost random without correlation each other. The mixed signals generally pass preconditioning filters and then are output from the sound sources. Their characteristics

are determined so as to reproduce the original signal, $s(n)$, at the local spot. It is obvious that the sound field and the location of sound sources affects them, meaning that the reproduction accuracy is also affected by those filters.

From Eq. 5, the decomposition of a single frame into random vectors costs $O(N^2)$ operations (if the inverse matrix $\mathbf{V}^{-1}$ was formally computed), and the mixing costs $O(N \cdot M)$ operations. When the value of N and M is small, these operations are possible in real-time using conventional computing facilities such as DSP. Such a computing equipment is required to implement this processings, the further equipment is not required, implying that the introduced method can be implemented without any special hardware except the DSP.

COMPUTER SIMULATION OF SIGNAL SYNTHESIS WITH RANDOM VECTORS

In this paper, only the effect of sound source location is dealt with as the fundamental study. The sound field is assumed to be the free field. This simplification enables us to discuss the effectiveness of signal decomposition method itself.

Conditions on signal decomposition

As the starting point of the computer simulation, some conditions concerning the signal decomposition has to be determined.

Sample points in a frame

The first is the sample points in a frame (expressed as N in Fig. 1). The smaller value of N leads to the ill-condition of the matrix $\mathbf{V}$, hence N is set to 200 in this paper.

Number of sound sources

After the mixing of each random vector multiplying its corresponding weight factors as shown in Fig. 2, the authors listened to each input signal to each sound source. As a result of this simple hearing test, we concluded that the contents of speech can be understood unless the number of sound source is more than 20. Therefore the number of sound sources is determined as 20 in this paper.

Location of sound sources

In order to simplify the investigate on the effect of the sound field and the sound source location, two kinds of sound source location were considered.

Sound sources located with the same distance from the spot: This means no special preconditioning filters are required, *i. e.*, $H_m(z) = 1$ $(m = 1, \cdots, M)$, since all the output from the sound sources propagate to the local spot with the same delay, and simply summing the decomposed random vectors generates the original signal there. Two examples out of this type of sound source location are examined. They are the circular and the semicircular shapes.

Sound sources located with various distance from the spot: This type of location is termed "distributed" hereafter. In this case, the adjustment of gain and delay due to the difference in the distance from each sound source to the spot is required. When the position (or the center) of the spot, and the position of the i-th sound source are respectively expressed as $\mathbf{r}_0$ and $\mathbf{r}_i$, the procedure of this adjustment is summarized as follows:

1. The maximum values of the distance and the propagation delay from the spot to each sound source is obtained by calculating the following equations:

$$\Delta r_{\max} = \max_{1 \le m \le M} r_{m0}, \tag{6}$$

$$\Delta t_{\max} = \frac{\Delta r_{\max}}{c}, \tag{7}$$

where $r_{m0} = ||\mathbf{r}_m - \mathbf{r}_0||$, and c is the speed of sound.

2. According to the maximum values calculated above, the gain and the delay of each sound source are normalized. The analog impulse response of the filter are determined as follows:

$$h_m(t) = \frac{r_{\max}}{r_{m0}} \delta\left(t - \frac{r_{\max} - r_{m0}}{c}\right), \tag{8}$$

where $\delta(t)$ is the Dirac's Delta function. The z-tranform of digitally-sampled $h_m(t)$ corresponds to $H_m(z)$ in Fig. 2.

The filters are characterized only by the gain and the delay, resulting the flat transfer functions.

Used source signal and SDR

The source signal was an announcement of female talker in Japanese sampled from CD (sampling frequency: 44.1 kHz) as an example. The reproduction accuracy was evaluated by modified SDR (Signal-to-Distortion Ratio) defined by the following equation:

$$\mathrm{SDR}(\mathbf{r}) = 10\log_{10} \frac{\sum_{n=0}^{N-1} s_r^2(n)}{\sum_{n=0}^{N-1} \{s_r(n) - x(\mathbf{r}, n)\}^2} \ [\mathrm{dB}]. \tag{9}$$

$x(\mathbf{r}, n)$ denotes the signal acquired at the point $\mathbf{r}$, and $s_r(n)$ indicates the original signal, but its energy is normalized to that acquired at the desired local spot. When the acquired signal is close to the original one, SDR has large value.

Results and discussion

Although the simulated sound field is assumed to be three-dimensional, the positions of the sound sources and the local spot were also assumed in a certain two-dimensional plane. Therefore it is noted that the plot of the sound sources and the local spots are two-dimensional.

Sound sources located in circular and semicircular shapes

Positions of sound sources in circular shapes are shown in Fig 3. In this figure, the circles indicate the positions of sound sources, number of which is 20 on the circle of 3 m radius. The local spot is set to the origin. The $\times$ points show the listening points, at which the acquired signals were computed. SDR was calculated at the crossing points of the horizontal and the vertical broken lines. The distribution of SDR is shown in Fig. 4. It can be found out that the value of SDR is high at the origin, and is around 0 dB at all points except the spot. This relects that the original signal is well reproduced at the spot, and the acquired signal is somewhat far from the original except there. The case of semicircular sound source location are shown in Fig 5. The meaning of each point is the same as that in Fig. 3. The distribution of SDR is shown in Fig. 6. While the slight change of SDR can be seen at around the source positions and the spot, the tendency is almost the same as Fig. 4.

Although the low SDR was obtained at the point except the desired spot in both sets of sound source positions, but there is not a huge difference between the acquired waveform at the spot and that at the positions except the spot. As an example, the acquired waveforms at the origin and at the point of (2.5,2.5) [m]

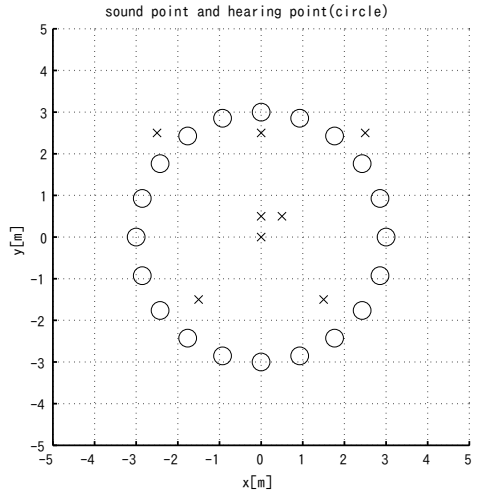


Figure 3: Sound source positions and listening points (circular shape)

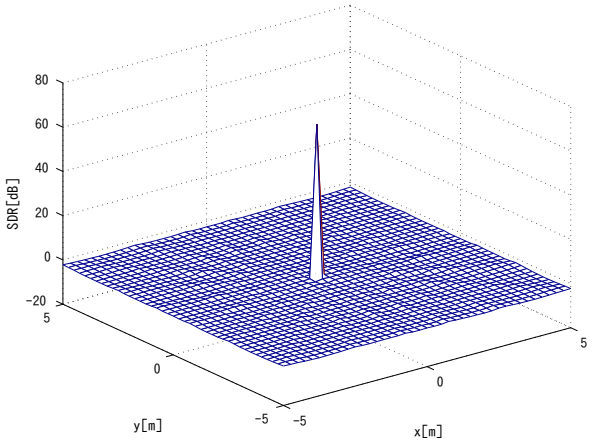


Figure 4: SDR in the case of sound source positions on circle

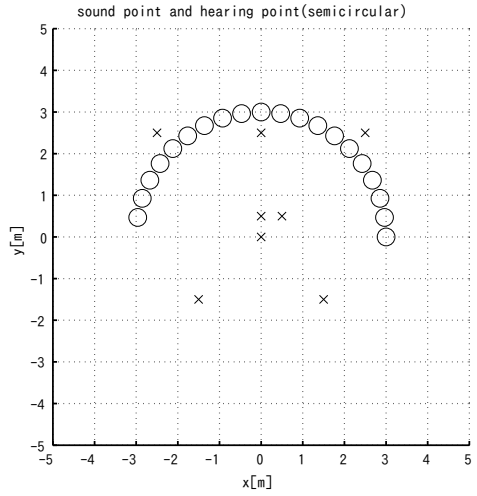


Figure 5: Sound source positions and listening points (semicircular shape)

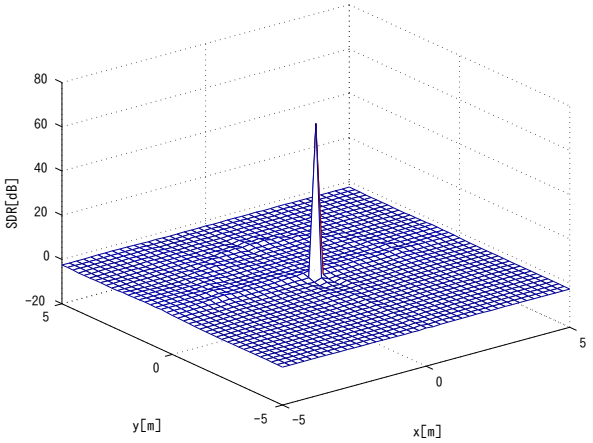


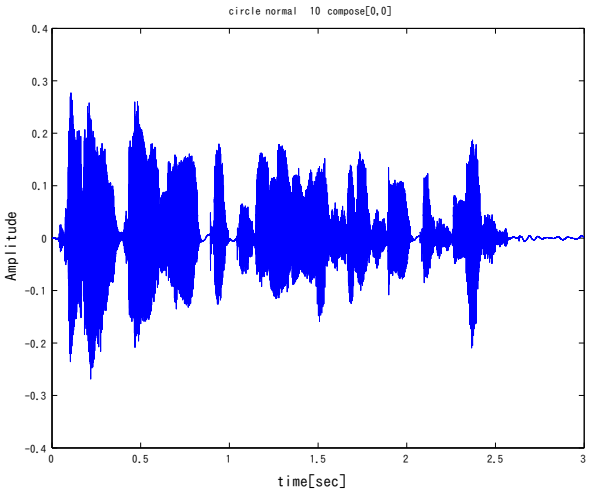
Figure 6: SDR in the case of sound source positions on semicircle

are compared in Fig. 7. Figure 7(a) corresponds to the waveform acquired at the spot, this is almost the same as the original speech signal. It can be shown from Fig. 7(b) that the macroscopic form is similar to each other while the acquired waveform is somewhat distorted comparing to Fig. 7(a). When the authors listened to the signal depicted in Fig. 7(b), a part of the speech contents could be heard. This may reflect that the SDR does not reflect the secrecy of information in speech. Investigation on the subjective evaluation of information in speech may be one of the future works.

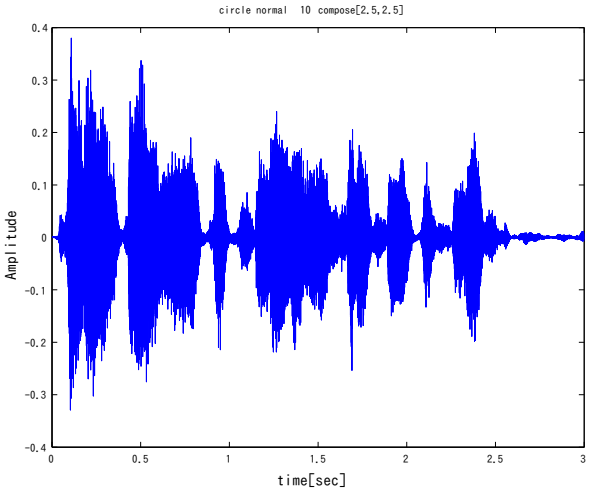
Case of distributed sound sources

As an example of sound sources located with various distance from the desired spot, the location depicted in Fig. 8 was adopted. The meaning of each symbol is the same as that in Figs. 3 and 5. Since the distance from each sound source to the spot is different, the previously mentioned processings were executed so as to normalize the gain and the delay. The resulting distribution of SDR is shown in Fig. 9. It is clear that the tendency of SDR is the same as Figs. 4 and 6, meaning that the original signal is accurately reproduced at the spot, and the acquired signal is distorted at the points except the spot. It seems to be stated that the distributed sound source location has almost no effect comparing to the case of all sound sources at the same distance.

Figure 10(a) and (b) show the acquired waveforms at the position of (0,0) [m] and (2.5,2.5) [m], respectively. Comparing the waveforms to those in Fig. 7, scale of the vertical axis is different. This comes from the variation in distance of each sound source to the spot. As shown in Fig. 9, the waveform is almost the same as the original one at (0,0) [m] (Fig. 10(a)). Seeing Fig. 10(b), there still remains the macroscopic feature of the original signal. However, the fine structure of the waveform seems to be distorted more than that in Fig. 7(b). This feature may come from the distribution in the gain of each sound source given due to the distributed location. The variance in the gain generates the variance in the acquired sound waves from each sound source, resulting the change in frequency spetrum of the acquired signal from that of the original one. However, this difference is not represented in the value of SDR as shown in Fig. 9(b). Instead of that, it can be said that the information in the speech is difficult to be heard when the signal depicted in Fig. 10(b) was presented as sound. Although this feature can not be indicated in quantitative manner, the authors consider that the distributed sound source location has the potential to satisfy both the accurate reproduction of speech signal at the desired spot and the secrecy of information in speech at any



(a) at (0,0) [m]



(b) at (2.5,2.5) [m]

Figure 7: Acquired waveforms at various positions (sound source location: circular shape)

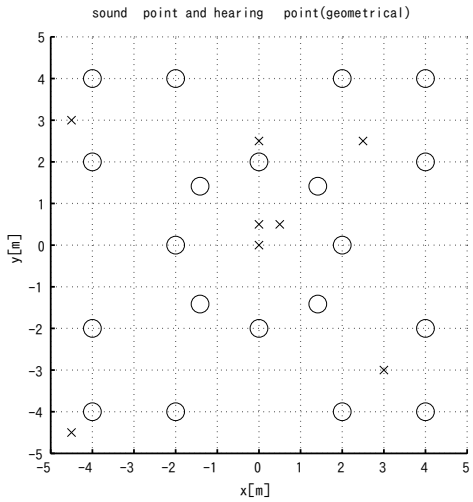


Figure 8: Sound source positions and listening points (distributed)

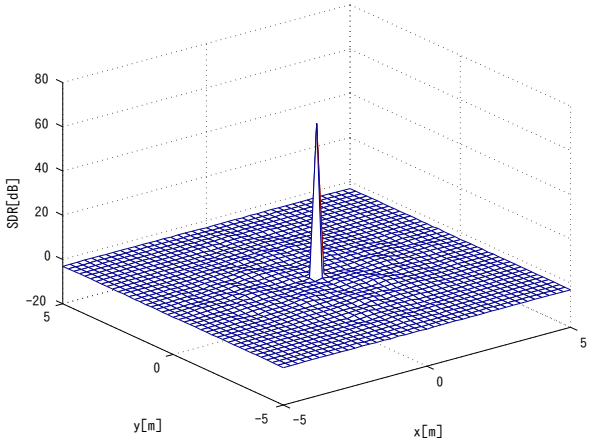


Figure 9: SDR in the case of distributed sound source positions

point except the spot.

In order to clearly confirm the effectiveness of the distributed sound source location, the relation between the distortion of speech and the ability of information perception in speech is one of the subjects of the future works. Moreover, the macroscopic structure (envelope) of the original speech is still remained even in the synthesized signal with the distributed sound source location. The introduced method is based on repeating the processing within a frame, and that spanning the multiple frames is not taken into account. Applying such a processing leads to the generation of the distortion in the synthesized signal at the point except the local spot. Investigation on this is also considered to be the future works.

CONCLUDING REMARKS

In this paper, an alternative approach was introduced for the reproduction of signal at a certain local spot based on the signal decomposition of signal into random vectors. The decomposition method was originally proposed by Togura *et al.*[6], and applied to sound reproduction by Negi *et al.*[7]. However, some problems arises when the reproduction of speech signal was intended by using this approach. One of them was that the information in speech can be heard at the point except the desired spot. The simple computer simulation was carried out in order to investigate the effect of sound source location. The results showed that the original signal was precisely synthesized at the spot, and the acquired signal was distorted as long as the value of SDR is concerned. However, the macroscopic waveform was still remained in the acquired signal at the point except the spot. For solving this difficulty, the distributed sound source location may contribute to some extent.

ACKNOWLEDGMENT

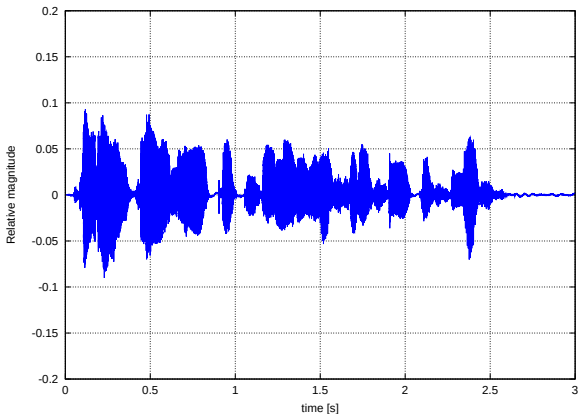
The authors would like to thank Mr. Takashi Kumon for his help in the computer simulation. A part of this research was supported by the Grant-in-Aid for Exploratory Research (No. 21656097).

REFERENCES

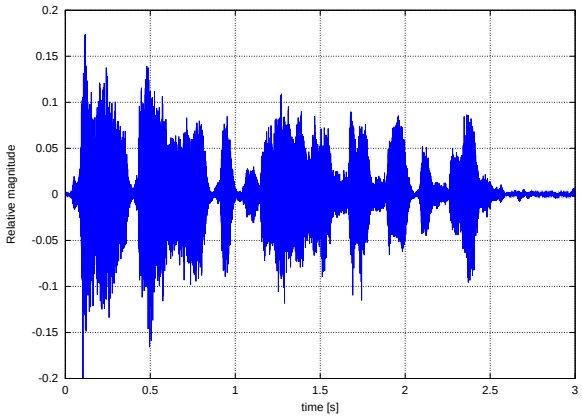
[1] P. J. Westervelt, “Parametric acoustic array,” J. Acoust. Soc. Am., **35**, 535-537(1963).

[2] M. B. Moffett and R. H. Mellen, “Model for parametric acoustic sources,” J. Acoust. Soc. Am., **61**, 325-337(1977).

[3] T. Kamakura, M. Yoneyama and K. Ikegaya, “De-



(a) at (0,0) [m]



(b) at (2.5,2.5) [m]

Figure 10: Acquired waveform at various positions (sound source location: distributed)

velopments of parametric loudspeaker for practical use," Proc. 10th International Symposium on Nonlinear Acoustics, 147-150(1984).

[4] J. F. Pompei, "The use of airborne ultrasonics for generating audible sound beams," J. Audio Eng. Soc., **47**(9), 726-731(1999).

[5] K. Ogino, Y. Ouchi and Y. Yamasaki, "Flat panel loudspeaker: multi-cell dynamic type and flexible electrostatic type," J. Acoust. Soc. Jpn., **62**(11), 802-807(2006) (in Japanese).

[6] A. Togura, O. Miura and M. Tohyama, "Speech signal representation using random vector linear combination," Trans. Autumn Research Meeting of Acoust. Soc. Jpn., 497-498(1999) (in Japanese).

[7] N. Negi, Y. Oikawa, H. Hattori and Y. Yamasaki, "Acoustic signal reproduction in a real field using random vectors synthesis," Trans. Spring Research Meeting of Acoust. Soc. Jpn., 497-498(2000) (in Japanese).

[8] H. Anton and R. C. Busby, *Contemporary linear algebra* (John Wiley & Sons, 2003).