Sound Processing Technologies for Realistic Sensations in Teleworking

Takashi Yazu Makoto Morito

In an office environment we usually acquire a large amount of information without any particular effort and unconsciously from sounds we hear without paying much attention to it. Who is doing what sort of work, is he or she busy, who is there or how is the person's health condition? What about the equipment, rather than just the people? The amount of information we obtain from nonverbal sounds is actually not so small. Sound communication with teleworking in general, however, is intended primarily for conversational vocal sounds and consideration does not extend to include the transmission of nonverbal information. When a microphone is used to record such a conversation and heard at a remote location via a communication channel, the clarity of such a conversation significantly deteriorates due to a variety of distorting factors. As a result, a lot of useful information is lost. In order to make teleworking truly effective, it is essential that a sound environment is realized with abundant realistic sensations, transmitting the natural atmosphere or mood of the setting.

In this paper an overview of the sound processing technologies intended to realize high quality and highly realistic sensations is provided, followed by an introduction to the sound source separation technology, one of the elemental technologies.

Sound processing technologies for realization of high quality sound and highly realistic sensations

(1) Acoustically realistic sensation generating technology

Let us consider communicating by sound between two remotely located points (**Fig. 1**). The sound of the other party, recorded with microphones installed at specified locations, is played back using a speaker, etc. The sense of direction, sense of distance, as well as the sound volume balance of individual sounds are lost, due to the limited number of available microphones or the positional relationship between the microphones and sound sources. In order to realize highly realistic sensations, it is necessary to use a stereophonic technology that can be used to reproduce the condition of the original site, including a spatial sense of direction or sense of distance, in order to reproduce the sound space to make it feel as if the spaces are connected, even if they are both at remote locations.

Binaural reproduction (**Fig. 2-a**) consists of hearing sounds recorded at the original sound location using a binaural recording*1) and a headphone. The recording incorporates acoustic effects (such as the reflections and diffractions of sounds by the head) that occur between the sound source, where the sound is emitted and both the listener's ears, when it arrives. The listener, therefore, can acquire a sort of realistic sensation that makes him or her feel as if the sound is being heard at the location

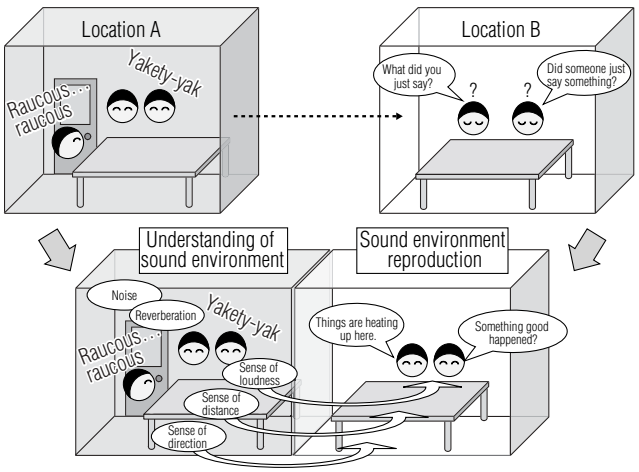


Fig. 1 Reproduction of realistic sounds for teleworking

*1) A binaural recording is a method of recording using a set of two microphones installed on the ears of a figure (dummy head) for the purpose of recording sounds that enter the left and right ears of a person.

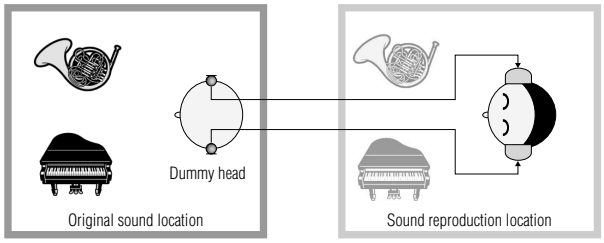


Fig. 2-a Binaural reproduction

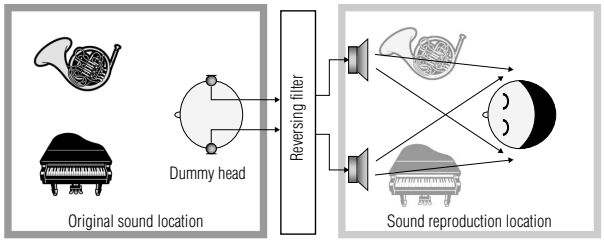


Fig. 2-b Transaural system

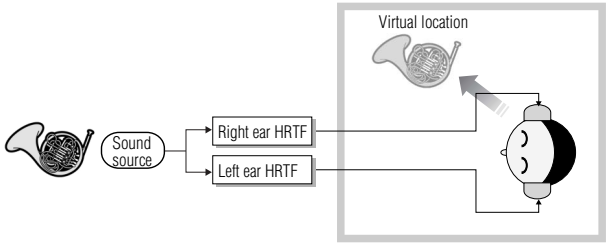


Fig. 3 Sound image positioning using HRTF

where the sound originates. On one hand the system is simple and is not subject to the influences of the environment in the room where the sound is reproduced, on the other hand problems and limitations arise from the use of the headphones, which often fix the sound image within the head, inhibiting ordinary communications. However, a system that uses multiple speakers, which control the acoustic pressure of the ear at the point of hearing instead of headphones, is also being proposed. This system is referred to as the "transaural system" (**Fig. 2-b**) in order to distinguish it from binaural systems. Measurements for the transmission characteristics of sound reaching both ears of the listener from both speakers are taken in advance at the location where the sound is reproduced and the reverse of these characteristics are implemented to the recorded sound in order to realize a binaural reproduction of sound at the point of hearing. Such a method involves controlling a fixed point within the space and thus the realistic sensations and sense of direction are lost if the listener moves his or her head or relocates. A method for controlling a region, rather than a point, is also being proposed.

So far the technology known as acoustic field reproduction, which stores the acoustic information of the original sound location before reproducing it at another location, has been described. Sound image positioning technologies for creating stereophonic sounds by assigning positional sensations to each individual sound source is also being widely researched. Humans recognize the direction and distance of sound based on the differences in the acoustic characteristics (head related transfer function, HRTF) of the sound traveling from the sound source and arriving at both ears. When measurements of HRTF are taken from a variety of angles and if binaural signals can be manipulated to make them suitable for the HRTF of sound emitted from the sound source located at a specific location, then it would be possible to position sound as if the sound source actually existed at a particular location (**Fig. 3**). In order to adapt this to teleworking with realistic sensations, it is necessary to specify the location of each individual sound source and to have sound separation for each sound source. For this reason, it would also be necessary to acquire technologies to estimate the location of a sound source or to separate sound sources.

(2) Quality level raising technology

The original sound emitted by a sound source is transmitted with air as the medium and reaches the microphones or ears. In transit, the sound undergoes a variety of deformations by mixing with conversational voices other than the intended voice, as well as environmental noise and reverberations. In order to realize communications with a superior sound quality, not only are stereophonic acoustics necessary, but also a means to implement deformations.

Noises are relatively constant, as with air conditioning sounds, there are diffused noises for which sound sources cannot be specified in one direction and directional noises that have directivity, such as voice or music, which significantly change over time. Strategies to deal with these noises differ, due to their different characteristics. Noise reduction methods, such as spectral subtraction (commonly referred to as the "SS method") or the Wiener filter, are used to deal with diffused noises, whereas noise canceller methods can be used if a reference microphone is available for observing noise alone, aside from the main microphone. It is however difficult to utilize such noise reduction methods to remove directional noises. A sound source separation technology is used for such a purpose, as it separates and extracts only the intended sound (primarily voice) from a mixture of sounds originating from multiple sound sources. Sound source separation using a beam former, with a microphone array to create a strong directivity to the direction of the intended sound and sound source separation, utilizing the independent component analysis (ICA), are well known.

Reverberation not only deteriorates the intelligibility of sound, but also causes a significant negative impact on sound source separation and acoustic field control, described later. Elimination of reverberation, therefore, is an essential topic that cannot be overlooked. Other than these, preventative action must be implemented to deal with the acoustic echo that is caused by the sound generated by a speaker located at a remote location entering the microphone and returning to the location of the party emitting the sound. These echoes present difficult issues, such as longer delays in comparison with

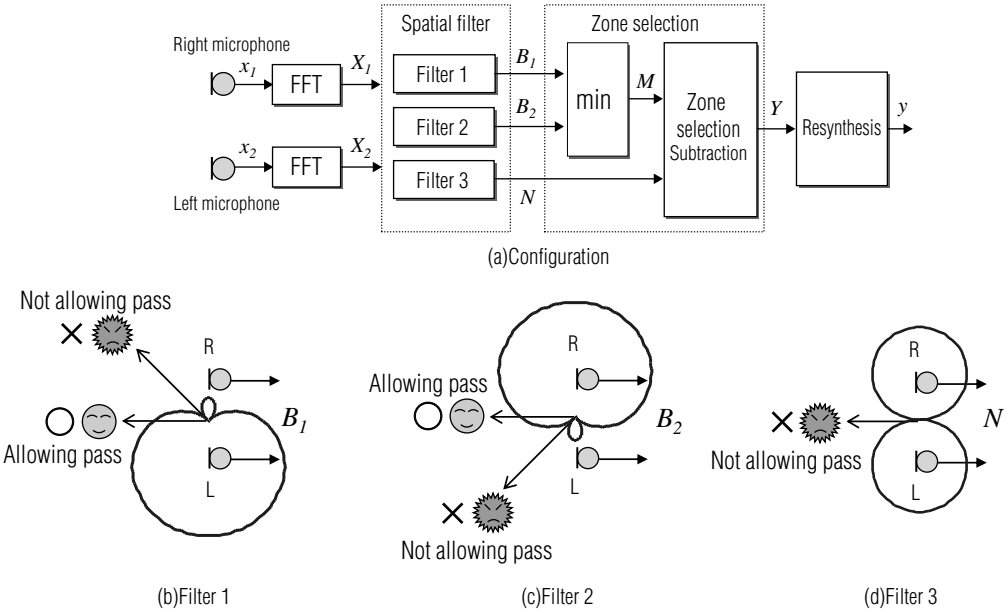


Fig. 4 Sound source separation method

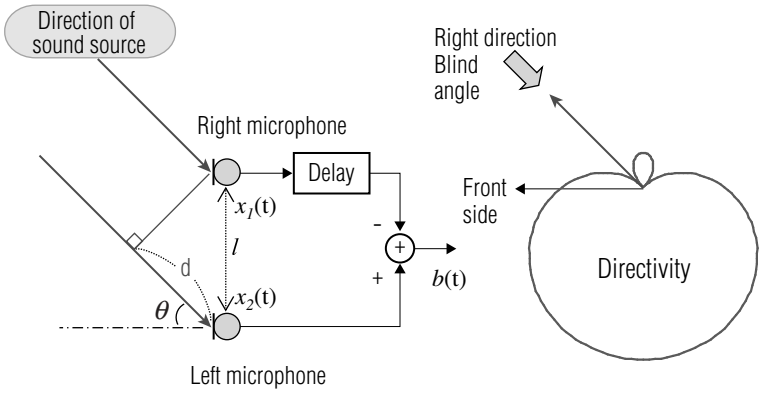


Fig. 5 (a) Principle of spatial filter **Fig. 5 (b) Directivity**

line echoes or the fluctuation of echo channel characteristics.

Sound source separation

We have been researching and developing a sound source separation technology, which is one of the elemental technologies for sound processing in teleworking. This method proposed by Kobayashi and his associates[1)] is comprised of a compact arrangement of microphones and can be realized with lower calculation costs. The configuration for a basic method involving the use of two microphones is shown in **Fig. 4** (on the next page), while the principle of a spatial filter used for this method is shown in **Fig. 5** .

A description of the principle of a spatial filter is provided first. Let us consider a situation when plane waves arriving from direction θ are received by two microphones, which are separated by distance I, as shown in **Fig. 5 (a)**. The sound wave arriving from direction θ is received first by the microphone on the right, which is closer to the sound source. Next, the sound wave proceeds by distance d, before arriving at the microphone on the left. The distance d, is then expressed as:

$$d = I \sin\theta \quad (1)$$

This means that the signal received by the microphone on the left, $x_2(t)$, is a signal that is delayed by τ, which is the amount of time required for the sound wave to travel the distance, d, in comparison with the signal received by the microphone on the right, $x_1(t)$. This means that the following relationship is established:

$$x_2(t)=x_2(t-\tau) \quad (2)$$

$$\tau=d/c=l \sin\theta / c, \text{ where c is speed of sound} \quad (3)$$

This means that both signals are cancelled out and a blind angle is created in the specific direction of θ, if the delay that equals t is added to the signal $x_1(t)$ and subtracted from signal $x_2(t)$ (added in reversed phase).

$$b(t)=x_2(t)-x_2(t-\tau) \quad (4)$$

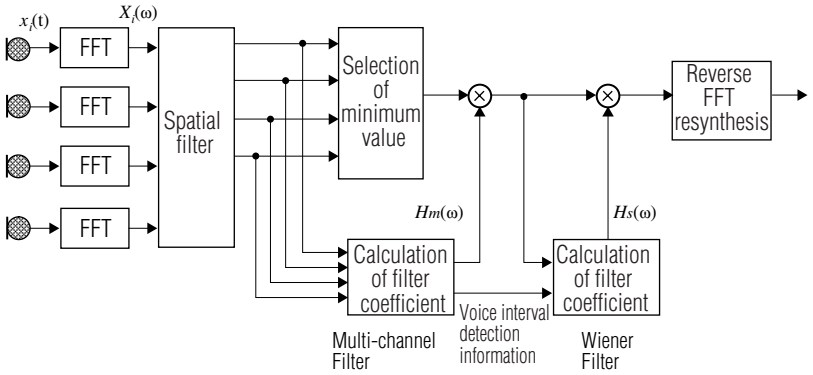


Fig. 6 Configuration of sound source separation with consideration for diffused noise

An example of this directivity is shown in **Fig. 5 (b)**.

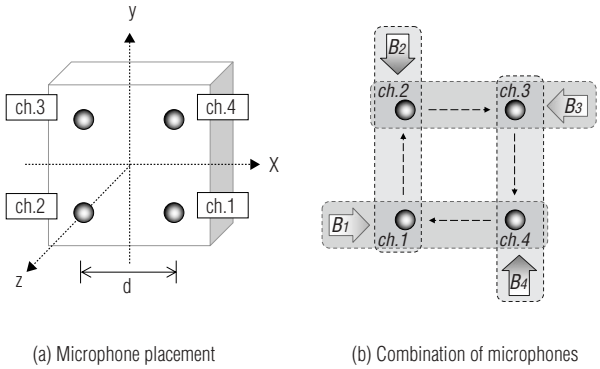
Operations similar to the spatial filter forming operation along the time axis can also be performed in the frequency domain. Iis known that the Fourier transformation of a signal with a time axis delayed by τ, is equal to the result of the performing Fourier transformation on the original signal and multiplied by $e^{-j\omega\tau}$. The equation (4) along the time axis is expressed as equation (5) along the frequency axis using the short time Fourier transformation $X_1(\omega)$ and $X_2(\omega)$ of $x_1(t)$ and $x_2(t)$.

$$B(\omega) = X_2(\omega) - e^{-j\omega\tau} X_1(\omega) \quad (5)$$

The sound source separating method is described next. In this method, input from two microphones is used to form three spatial filters, as shown in **Fig. 4 (a)**. A blind angle is set on the right direction of spatial filter 1, which suppresses the arrival of any interfering noise from the right direction. The intended sound is output with a certain amount of gain. This output is referred to as $B_1(\omega)$ [**Fig. 4 (b)**]. The spatial filter 2 has blind angles set to the left direction, which suppresses the arrival of any interfering noise from the left direction. Similarly to spatial filter 1, an intended sound or sound output with a certain amount of gain is referred to as $B_2(\omega)$ [**Fig. 4 (c)**]. The spatial filter 3 has a blind angle set in front of it [**Fig. 4 (d)**] and extracts the components other than the intended sound. The output is referred to as $N(\omega)$. Select the smaller amplitude component, $|B_1(\omega)|$ of the output from spatial filter 1 and the amplitude component, $|B_2(\omega)|$ of the output from spatial filter 2.

$$M(\omega) = min\,[\,|B_1(\omega)|, |B_2(\omega)|\,] \quad (6)$$

If a sound source with interfering noise exists in the right direction, the output $B_1(\omega)$ of spatial filter 1, with a blind angle on the right direction, suppresses the interfering noise and reduces its amplitude. On the other hand, the output $B_2(\omega)$ of spatial filter 2, with a blind angle in the direction where no interfering noise exists, does not result in a significant change in amplitude. Reversely, if a sound source of an interfering noise is located in the left direction, then $B_2(\omega)$ becomes smaller but the change with $B_1(\omega)$ will remain small. Thus the selected minimum value M is a candidate component of an intended sound, which suppresses the largest interfering noise. Finally,



(a) Microphone placement (b) Combination of microphones

Fig. 7 Microphone placements

the output $Y(\omega)$ is determined by selecting the zone and spectral subtraction using $M(\omega)$ and $N(\omega)$

$$Y(\omega) = \begin{cases} \sqrt{|M(\omega)|^2 - \alpha|N(\omega)|^2} & \text{if } |M(\omega)| > \alpha|N(\omega)| \\ 0 & otherwise \end{cases} \quad (7)$$

This is a spatial filter gain correction coefficient. Zone selection is performed to determine whether or not components of an intended sound is included in the signal $M(\omega)$. Since $N(\omega)$ is considered to be an ambient noise from a direction other than that of the intended sound, and if $N(\omega)$ is larger than $M(\omega)$, then this segment is considered to be a location where none of the components belonging to the intended sound exist and thus it is discarded. If it is determined that a component of the intended sound does exist in the signal $M(\omega)$, then subtraction is performed and acute directionality is directed to the frontal direction, separating the intended sound.

For the sake of simplicity a configuration comprised of two microphones is shown here. It becomes possible to deal with directional noise from various directions if the space can be dealt with by placing microphones not only in the left-to-right horizontal directions but also up-and-down vertical directions.

Sound source separation with consideration for diffused noise

It is extremely rare to find only directional noise in any environment where this sound processing technology is used and both directional and diffused noises are a mixture in actual environments. A description of the sound source separating system, which is capable of suppressing not only directional noises but also diffused noises, is provided here[2]. The system is comprised of a directional noise suppressing section, diffused noise suppressing section and residual noise suppressing section, as shown in **Fig. 6**. Four non-directional microphones are placed in a square formation over a flat surface in this system, as shown in **Fig. 7 (a)**. The intended sound is expected to arrive from the front (Z-axis direction).

(1) Suppression of directional noise

The suppression of directional noise by the system is described first. The principle of the spatial filter described earlier is used and, of the four microphones, two microphones are paired up as shown in **Fig. 7 (b)**, to comprise four sets of microphone pairs, in order to configure spatial filters in four directions. These individual spatial filters are made possible by equations (8) to (11), with four directionalities in the directions of up, down, left and right.

$$B_1(\omega) = X_1(\omega) - e^{-j\omega\tau} X_4(\omega) \quad (8)$$

$$B_2(\omega) = X_2(\omega) - e^{-j\omega\tau} X_1(\omega) \quad (9)$$

$$B_3(\omega) = X_3(\omega) - e^{-j\omega\tau} X_2(\omega) \quad (10)$$

$$B_4(\omega) = X_4(\omega) - e^{-j\omega\tau} X_3(\omega) \quad (11)$$

The smallest of all amplitude components among the outputs of these four spatial filters is selected and output to obtain the minimum output of the component of a directional noise.

$$|B_{min}| = \min\,[\,|B_i|\,] \ (i = 1,2,3,4) \quad (12)$$

(2) Suppression of diffused noise

Diffused noise suppression is realized using a Multi-Channel Wiener Filter, which uses four spatial filters the same as used for the suppression of directional noise. The voice of a speaker, the intended sound, exhibits a high correlation with the signal observed by the respective microphones, but the diffused noise exhibits a low correlation between the individual signals. Using this characteristic, signals with directionalities in opposite directions are paired up (B_1 with B_3 and B_2 with B_4) to comprise filters with a coefficient that reflects the extent of mutual correlation.

$$H_m(\omega) = \frac{|B_1(\omega)B_3^*(\omega)| + |B_2(\omega)B_4^*(\omega)|}{\frac{1}{2}\sum_{i=1}^{4} |B_i(\omega)|^2} \quad (13)$$

The equation shown above normalizes the cross spectrum of the numerator with the power spectrum of the denominator. The characteristic of this equation is such that when the correlation is high, the resulting value is one, whereas if the correlation is low, then the value approaches zero. Diffused noise is reduced by suppressing those components with a lower correlation as well as by multiplying this filter with the signal $|B_{min}|$ and the directional noise suppressed as described earlier.

$$\hat{S}_m(\omega) = H_m(\omega)\,|B_{min}(\omega)| \quad (14)$$

(3) Suppression of residual noise

Residual constant noise is suppressed by applying the Wiener filter with a single channel on the signal $\hat{S}m(\omega)$, which suppresses the directional noise and diffused noise. The Wiener filter considers signals and noises to be a stochastic process and minimizes the mean square error. If signals and noises are assumed to not have a correlation, then the gain function is given by the following equation:

$$H_s(\omega) = \frac{SNR_{prio}(\omega)}{SNR_{prio}(\omega) + 1} \quad (15)$$

The post signal to noise ratio $SNR_{post}(\omega)$ and ante-signal to noise ratio $SNR_{prio}(\omega)$ are defined respectively in the following manner:

$$SNR_{post}(\omega) = \frac{|\hat{S}_m(\omega)|^2}{E[\,|N(\omega)|^2\,]} \quad (16)$$

$$SNR_{prio}(\omega) = \frac{E[\,|S(\omega)|^2\,]}{E[\,|N(\omega)|^2\,]} \quad (17)$$

The $E[\cdot]$ represents the expected value, while $S(\omega)$ represents the intended sound signal. The ante- signal to noise ratio $SNR_{prio}(\omega)$ cannot be directly measured, since it also includes $E[|S(\omega)|^2\,]$. Thus the post signal to noise ratio and the estimated signal $S_{-1}(\omega)$ are used to calculate an approximation.

$$\hat{SNR}_{prio}(\omega) = \beta \frac{|\hat{S}_{-1}(\omega)|^2}{E[\,|N(\omega)|^2\,]} + (1-\beta)P[SNR_{post}(\omega) - 1] \quad (18)$$

The $P[\cdot]$ here represents a half wave rectification, while β represents the obliteration coefficient.

An estimation of the noise level, however, is performed in an oblivious manner from the signal in a non-conversational segment:

$$|N(\omega)|^2 = (1-\lambda)\,|Sm(\omega)|^2 + \lambda\,|N_{-1}(\omega)|^2 \quad (19)$$

The forgetting coefficient λ is selected to about 0.95 to 0.99. Furthermore, in order to prevent components of the intended sound from mixing in, noise learning is suspended during the segment in which speech vocalization takes place.

(4) Sound source separation device prototype

In order to evaluate the developed sound source separating method in an actual environment, a compact terminal with four channels, consisting of MEMS microphones, a CPU board and an AD conversion board,

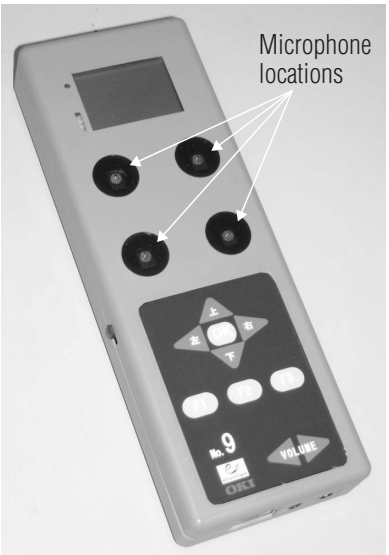


Photo 1 Sound source separator prototype

was prototyped (**Photo 1**). Every calculation process was conducted with fixed point values as well as calculation processes, such as FFT, square roots and double length divisions, furthermore, the processes described have all been incorporated into the prototype. The distance between the microphones was three centimeters, making this a very small piece of equipment that could be mounted on compact devices, such as remote controls or mobile phones.

Conclusion

An overview of sound processing technologies pertaining to teleworking with realistic sensations has been provided as well as descriptions on a sound source separation technology, which is an elemental technology. Realistic sensation generating technologies using sound have been studied as acoustic field reproducing technologies for fixed content, such as musical performances, as well as sound image positioning technologies to create virtual sound images by adding the feel of the position to the sound sources themselves. There are however numerous issues to be resolved in order to generate realistic sensations in real-time remote communications. We will continue with our research into sound processing technologies with the aim of realizing a sound environment that would make one feel as if he or she is actually in the office, while at home. The development and prototyping of the sound source separating method was implemented as a contracted work from Waseda University, using part of the budget provided by the Ministry of Economy, Trade and Industry for the strategic technology development contract of fiscal 2006 and 2007, “Development of Basic Technologies for Voice Recognition”.

References

1) Shintaro Takada et al: Sound Source Separation Using Small Number of Microphones for Mobile Terminals, 3-1-8, Collection of Seminar Papers, Acoustical Society of Japan, September 2006.
2) Shintaro Takeda et al: Considerations of Voice Enhancement for Mobile Terminals in Directional Noise and Diffused Noise Mixed Environments, 3-P-3, Collection of Seminar Papers, Acoustical Society of Japan, September 2007.

Authors

Takashi Yazu: Corporate Research and Development Center, Human Communication Lab., Specialist

Makoto Morito: Corporate Research and Development Center, Human Communication Lab., Senior Specialist