Speech Privacy for Sound Surveillance Using Super-Resolution Based on Maximum Likelihood and Bayesian Linear Regression

Ryouchi NISHIMURA^{†a)}, Member, Seigo ENOMOTO[†], and Hiroaki KATO[†], Nonmembers

SUMMARY Surveillance with multiple cameras and microphones is promising to trace activities of suspicious persons for security purposes. When these sensors are connected to the Internet, they might also jeopardize innocent people's privacy because, as a result of human error, signals from sensors might allow eavesdropping by malicious persons. This paper presents a proposal for exploiting super-resolution to address this problem. Super-resolution is a signal processing technique by which a high-resolution version of a signal can be reproduced from a low-resolution version of the same signal source. Because of this property, an intelligible speech signal is reconstructed from multiple sensor signals, each of which is completely unintelligible because of its sufficiently low sampling rate. A method based on Bayesian linear regression is proposed in comparison with one based on maximum likelihood. Computer simulations using a simple sinusoidal input demonstrate that the methods restore the original signal from those which are actually measured. Moreover, results show that the method based on Bayesian linear regression is more robust than maximum likelihood under various microphone configurations in noisy environments and that this advantage is remarkable when the number of microphones enrolled in the process is as small as the minimum required. Finally, listening tests using speech signals confirmed that mean opinion score (MOS) of the reconstructed signal reach 3, while those of the original signal captured at each single microphone are almost 1.

key words: sensor network, sound surveillance, maximum likelihood, Bayesian linear regression, Mean Opinion Score

1. Introduction

Surveillance of public streets and other public spaces using multiple cameras and microphones is promising to trace activities of suspicious persons for security of society. Such measures can record precise information including time and location so that the record is useful as reliable evidence when an incident happens. Moreover, once these sensor devices are connected to the Internet, they can immediately exchange the captured information and realize further advanced functions such as rapid information acquisition, target position tracing, multi-modal integration, and so on. Constructing such a sensor network might pave the way to a safer society by preparing for potential criminals. However, at the same time, personal privacy risks will become a matter of great concern because the sensor network gathers information blindly, including all information at a node. As the number of surveillance cameras grows, this problem is attracting greater attention.

The right to anonymity in public spaces has been thoroughly discussed with respect to government-driven public surveillance [1]. Such camera surveillance might prevent people from free-minded activities and might therefore undermine the general right to privacy. The topic of privacy problems on surveillance is dealt with sometimes in conjunction with research related to automatic face recognition [2]. Such technology simplifies the identification of individual people in a video, thereby creating a potential threat to privacy. Simultaneously, its ability to locate the face position in the image is useful to protect privacy by hindering the facial part. Many approaches to conceal regions of interest have been proposed [3], [4]. Some make it possible for authorized users to recover the original video content using watermarking technology [5]. It is insufficient to hinder facial parts to protect privacy because a video includes several personal features such as time of appearance, location, posture, fashion, action, and so on [6]. Therefore, it is also necessary to consider possible leaks of information related to privacy through features other than facial information. In this respect, speech information is delicate than facial information.

Speech privacy is sometimes discussed in the context of architecture, where speech intelligibility must become sufficiently low behind walls in meeting rooms and offices. B.N. Gover and J.S. Bradley proposed objective measures based on a signal-to-noise ratio for assessing architectural speech privacy [7]. Such an approach is possible because architectural speech privacy is passive, where the sound level of speech is reduced as it transmits through structural walls. Another approach to speech privacy is to encrypt or scramble the recorded speech [8], [9]. L. Lee introduced frequency domain asynchronous scrambling using filter bank analysis and synthesis. These methods are attractive but they present a potential threat: a malicious person might eavesdrop on the signal before it is scrambled. Therefore, the purpose of this paper is to propose a new method of speech privacy that renders the speech difficult to recognize at the moment it is sampled by an AD converter. We assume public surveillance using multiple microphones and use super-resolution to do it.

By super-resolution, a signal is obtainable with higher resolution than the actual measurement using multiple microphones. Reconstruction or estimation of the signal under such conditions is also regarded as super-resolution. In this perspective, signal processing for frequency-band expansion of a speech signal can be regarded as a kind of

Manuscript received April 4, 2017.

Manuscript revised August 25, 2017.

Manuscript publicized October 16, 2017.

[†]The authors are with the National Institute of Information and Communications Technology, Sendai-shi, 980–0812 Japan.

a) E-mail: ryouchi@nict.go.jp

DOI: 10.1587/transinf.2017MUP0003

super-resolution. Therefore, if the sampling rate of each AD converter is fixed sufficiently low, then the captured signal is difficult to recognize clearly as speech and only authorized persons who are eligible to access multiple microphone signals can expand the frequency band to recover its intelligibility. This scenario is promising when each microphone is individually connected to the Internet to send the data. In this case, some microphones might be attacked by a malicious person over the Internet. Moreover, unlike band expansion algorithms for single-channel speech signal, super-resolution can recover the signal no matter how little speech information each channel carries.

Super-resolution often requires some assumption or a priori knowledge about the signal features. For example, Roux proposed a method for reconstructing a complete audio signal from one lacking some components, using non-negative matrix factorization and the assumption of sparsity [10]. Keegan *et al.*, inspired by this idea, attempted to propose super-resolution for music signals using a sound dictionary and non-negative factorization [11]. Dong *et al.* developed the method further and proposed dictionary learning [12].

In the image processing research area, many works described obtaining a higher-resolution image from lower-resolution images. These lower-resolution images differ slightly from one another in their view points and make it possible to create a higher-resolution image by combining these images in a sophisticated manner. A fundamental framework of super-resolution for image can be found in a paper [13]. This technique uses no a priori knowledge but uses physical laws and a sampling theorem. It enables us to extract higher frequency components implicitly contained as a form of aliasing. The extracted components are used to reconstruct the original signal with higher resolution. A direct translation of this technique to audio signal is that when using multiple microphones deployed at different positions, a signal at a higher sampling rate is obtainable from those with a lower sampling rate. That fact suggests that frequency components even higher than Nyquist frequency can be estimated using multiple measurements at different positions. Meanwhile, speech signals have primal energy approximately within the frequency range of 300 Hz to 3,400 Hz, as used in telephony. It means that when a speech signal is sampled with a sampling rate sufficiently lower than that required to capture this frequency range, its intelligibility becomes low enough to make it difficult for human listeners to recognize what is spoken. Combining this fact with the ability of previously described super-resolution, it is possible to make only those who can access all the audio signals recorded with multiple microphones be able to reconstruct an intelligible speech signal by the use of super-resolution. In other words, those who can access only some of the microphone signals are unable to obtain a meaningful signal. To show the feasibility of such a system, super-resolution of an audio signal is investigated, focusing especially on speech signals. Two methods are derived respectively based on the maximum likelihood (ML)

and Bayesian Linear Regression (BLR). Their performance is thoroughly investigated through extensive computer simulations and is also tested through listening tests using speech signals.

The remainder of this paper is organized as follows. In Sect. 2, fundamental mathematical formulation of super-resolution is presented, where two cases are discussed regarding microphone configurations. In Sect. 3, computer simulations are conducted while varying several parameters which might affect performance. Experiments are also conducted in Sect. 4 using actual speech signals captured in an anechoic room. Section 5 presents some related discussion. Section 6 gives conclusions and closing remarks.

2. Fundamental of Audio Super-Resolution

2.1 Super-Resolution Based on Maximum Likelihood

A fundamental framework of audio super-resolution was presented briefly by the authors [14]. This section overviews that framework and provides statistically more rigorous explanations. The aim of super-resolution under consideration is to estimate the signal that would be obtained with a specified sampling rate f_i from signals of a lower sampling rate f_s . The ratio of the specified sampling rate to the lower one is called *undersampling rate* (USR) hereinafter. For brevity, one might first consider a situation in which two microphones capture a signal at a certain sampling rate and try to estimate the signal that would be obtained if it was sampled at a twofold higher sampling rate. If the length of the estimated signal $\mathbf{u}$ is N , then that of the actual measurement $\mathbf{y}_i (i = 1, 2)$ should be half.

Therefore, they are expressed in a vector form as

$$\mathbf{u} = [u(0) \ u(1) \ \cdots \ u(N-1)]^T \quad (1)$$

$$\mathbf{y}_1 = [y_1(0) \ y_1(1) \ \cdots \ y_1(N/2-1)] \quad (2)$$

$$\mathbf{y}_2 = [y_2(0) \ y_2(1) \ \cdots \ y_2(N/2-1)]. \quad (3)$$

These variables are related each other using a matrix $\mathbf{C}$ as

$$\mathbf{y} = \mathbf{C}\mathbf{u} \quad (4)$$

$$\mathbf{y} = [\mathbf{y}_1 \ \mathbf{y}_2]^T, \quad (5)$$

where T represents the transpose operation of matrix. Now, we assume that the data are measured using error expressed by additive white Gaussian noise (AWGN) ϵ , which is of zero mean and variance β^{-1} , as

$$\mathbf{t} = \mathbf{y} + \epsilon. \quad (6)$$

Conditional probability can then be expressed as

$$p(\mathbf{t}|\mathbf{C}, \mathbf{u}, \beta) = \mathcal{N}(\mathbf{t}|\mathbf{C}\mathbf{u}, \beta^{-1}). \quad (7)$$

If all measurement data are independent, then the log likelihood becomes

$$\begin{aligned}\ln p(\mathbf{t}|\mathbf{u}, \beta) &= \sum_{n=0}^{N-1} \ln \mathcal{N}(t_n | ((\mathbf{C}\mathbf{u}))_n, \beta^{-1}) \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \beta E_D(\mathbf{u}),\end{aligned}\quad (8)$$

where $((\cdot))_n$ indicates the n -th component of the vector and where

$$\begin{aligned}E_D(\mathbf{u}) &= \frac{1}{2}(\mathbf{t} - \mathbf{C}\mathbf{u})^T(\mathbf{t} - \mathbf{C}\mathbf{u}) \\ &= \frac{1}{2}\{\mathbf{u}^T \mathbf{C}^T \mathbf{C} \mathbf{u} - 2\mathbf{u}^T \mathbf{C}^T \mathbf{t} + \mathbf{t}^T \mathbf{t}\}.\end{aligned}\quad (9)$$

Because it is a quadratic form of $\mathbf{u}$, the stationary point of (9) gives the minimum value. The stationary point can be derived by performing partial differential of (9) for $\mathbf{u}$ as

$$\begin{aligned}\mathbf{u}_{\text{ML}} &= (\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \mathbf{t} \\ &= \mathbf{C}^+ \mathbf{t},\end{aligned}\quad (10)$$

where $^+$ stands for the Moore–Penrose pseudo-inverse of the matrix. Consequently, the maximum likelihood estimation of the super-resolution signal is obtained by (10). If $\mathbf{C}$ is a square matrix, then the pseudo-inverse can be replaced with the simple inverse.

To derive the exact form of the matrix $\mathbf{C}$, one can consider a linear microphone array with equal interval and $\text{USR} = 1/2$. Assuming that the arrival time difference of sound between two adjacent microphones is τ , matrix $\mathbf{C}$ can be constructed as

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix} \quad (11)$$

$$\mathbf{C}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & \vdots & \cdots & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \quad (12)$$

$$\mathbf{C}_2 = \mathbf{C}_1 \mathbf{G} \mathbf{D} \mathbf{F} \quad (13)$$

$$\mathbf{D} = \begin{bmatrix} e^{-j\omega_s \tau \cdot 0/N} & 0 & \cdots & 0 \\ 0 & e^{-j\omega_s \tau \cdot 1/N} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{j\omega_s \tau \cdot 1/N} \end{bmatrix}, \quad (14)$$

where $\omega_w = 2\pi f_s$ and f_s is the sampling frequency. Matrices $\mathbf{F}$ and $\mathbf{G}$ respectively represent matrix operations of discrete Fourier transform and inverse discrete Fourier transform [15]. On the other hand, matrix $\mathbf{D}$ represents a delay operation in the frequency domain, assuming a plain wave without decay. Matrix $\mathbf{C}_1$ is a matrix operation of culling samples. This method can be extended easily to other undersampling rates as long as its inverse is an integer.

It is noteworthy that this method requires that a sampling rate of the signal restored by this method must be an integer multiple of that of the measurement so that matrix $\mathbf{C}_1$ can be constructed otherwise. Also depending on the incident angle of the signal, matrix $\mathbf{C}_2$ can consist of complex numbers. Therefore the operation of matrix transpose in (9)

and (10) should be replaced with a Hermitian transpose. In this case, the estimated data are expected to be complex. An approximation is expected to ignore imaginary parts.

2.2 Super-Resolution Based on Bayesian Linear Regression

This section specifically examines a case in which microphones are separated by unequal distances. Although one can apply the method explained in the previous section, it might cause a problem because an inverse operation of the matrix is involved in the process. Moreover, its stability depends on the microphone configuration. While microphones need to be close enough to capture the signal without large differences in level and frequency characteristics among them, it may cause a problem if they are too close. For example, when some microphones are located in close vicinity, then the matrix becomes ill-posed. A similar situation may arise when the sound is coming from a perpendicular direction to a linear microphone array even if the microphones are not too close. It means that the method would be prone to noise addition or configuration errors. To address this problem, a technique of super-resolution based on Bayesian linear regression (BLR) is proposed.

Using w_k as a model parameter, a linear regression of the observation $y(\mathbf{x}, \mathbf{w})$ is expressed as

$$y(\mathbf{x}, \mathbf{w}) = \sum_{k=0}^{N-1} w_k \phi_k(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}), \quad (15)$$

where N stands for the degree, or the number of model parameter, $\mathbf{x}$ represents a vector corresponding to the observation condition, and ϕ_k denotes basis functions. To realize audio super-resolution using a statistical approach to a linear regression problem, we begin with the formula of an inverse discrete Fourier transform. According to textbooks on digital signal processing, discrete Fourier transform (DFT) and inverse discrete Fourier transform (IDFT) are generally formulated respectively as

$$Y(k) = \sum_{n=0}^{N-1} y(n) e^{-j2\pi kn/N} \quad (16)$$

$$y(n) = \frac{1}{N} \sum_{k=0}^{N-1} Y(k) e^{j2\pi kn/N}, \quad (17)$$

where N must be a power of two. With a discrete time shift of ξ , it can be rewritten easily as

$$y(n + \xi) = \frac{1}{N} \sum_{k=0}^{N-1} Y(k) e^{j2\pi k(n+\xi)/N}. \quad (18)$$

Here, we introduce an approximation to accept non-integer ξ , and rewrite (18) as

$$y(n + \xi) \simeq \frac{1}{N} \left\{ \sum_{k=0, k \neq N/2}^{N-1} Y(k) e^{j2\pi k(n+\xi)/N} \right.$$

Fig. 1 Microphone array configuration.

$$+(-1)^n Y\left(\frac{N}{2}\right) \cos(\pi \xi)\}. \quad (19)$$

It is noteworthy that (19) is a circular shift and that special attention is needed. As a matter of fact, this issue can be mitigated if the length of the analytical window is sufficiently long compared with ξ and the data are discarded on both edges where data are affected by the circular shift.

Now we define a vector $\mathbf{x}$ as a combination of information of time n and position s as

$$\mathbf{x} = \{n, s\}, \quad (20)$$

where s is a three-dimensional vector, representing the position of each microphone. Because time shift ξ corresponds to the arrival time difference between the microphone under consideration and the reference microphone, for an unequally spaced linear microphone array such as depicted in Fig. 1, it can be expressed, under a plain wave assumption, as

$$\xi = (-1)^\eta |s - s_0| \sin(\theta) (f_s/r)/c, \quad (21)$$

where

$$\eta = \begin{cases} 0 & \tau_{\max} \geq 0 \\ 1 & \text{otherwise} \end{cases}, \quad (22)$$

$$\tau_{\max} = \arg \max_{\tau} \left[\sum_n y(n, s_0) y(\tau + n, s) \right] \quad (23)$$

and where c signifies the speed of sound, r stands for the under sampling rate, and f_s represents the sampling rate of the actual recording. The microphone array is not necessarily linear but the arrival time difference of sound should must be calculated somehow. If a function $\phi_k(\mathbf{x})$ is defined as

$$\phi_k(\mathbf{x}) \equiv \frac{1}{N} e^{j2\pi k(n+\xi)/N}, \quad (24)$$

then (18) can be rewritten in the same form as (15), using w_k instead of $Y(k)$. By this formulation, an estimation problem of DFT coefficients of the signal can be reduced to a problem of estimating the model parameters of linear regression. Thereby Bayesian linear regression can be applicable. The design matrix is then given as

$$\Phi = \begin{bmatrix} \phi_0(\mathbf{x}_0) & \phi_1(\mathbf{x}_0) & \cdots & \phi_{N-1}(\mathbf{x}_0) \\ \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{N-1}(\mathbf{x}_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_M) & \phi_1(\mathbf{x}_M) & \cdots & \phi_{N-1}(\mathbf{x}_M) \end{bmatrix}, \quad (25)$$

where M represents the number of observations. It is equal to the product of the numbers of microphones and the samples recorded at each microphone.

Bayesian linear regression requires knowledge about the a priori probability distribution function. Therefore, we assume that the model parameter, which is equivalent to DFT coefficients of the super-resolution signal, conforms to a normal distribution with mean of zero and a covariance matrix $\alpha^{-1}\mathbf{I}$ as

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I}) \quad (26)$$

Here, the parameter α represents the precision, serving as a hyperparameter. Similarly, for the observation signal, another hyperparameter β is introduced, assuming that there is an AWGN.

As explained in textbooks such as one by C.M. Bishop [16], hyperparameters which maximize the evidence function can be estimated using an iterative operation. However, the transpose operation of the matrix is replaced with a Hermitian in this paper to consider complex values. The marginal likelihood function is obtained as

$$p(\mathbf{t}|\alpha, \beta) = \int p(\mathbf{t}|\mathbf{w}, \beta) p(\mathbf{w}|\alpha) d\mathbf{w}. \quad (27)$$

Using (8), (9) and (26), this equation can be represented as

$$p(\mathbf{t}|\alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{M/2} \left(\frac{\alpha}{2\pi}\right)^{N/2} \int \exp\{-E(\mathbf{w})\} d\mathbf{w}, \quad (28)$$

where

$$E(\mathbf{w}) = \beta E_D(\mathbf{w}) + \frac{\alpha}{2} \mathbf{w}^* \mathbf{w}. \quad (29)$$

Note here that $E_D(\mathbf{w})$ in (29) is the same as (9), but $\mathbf{C}$ is replaced with Φ . Using (9), (28) and (29), the likelihood of (28) is represented as

$$\begin{aligned} \ln p(\mathbf{t}|\alpha, \beta) &= \frac{N}{2} \ln \alpha + \frac{M}{2} \ln \beta - E(\mathbf{m}_M) - \frac{1}{2} \ln |\mathbf{A}| - \frac{M}{2} \ln(2\pi), \end{aligned} \quad (30)$$

where

$$\mathbf{A} = \alpha \mathbf{I} + \beta \Phi^* \Phi \quad (31)$$

and $\mathbf{m}_M$ is the expectation of posterior probability distribution. As a result, the stationary point of (30) with respect to α is obtained by solving equation

$$\frac{N}{2\alpha} - \frac{1}{2} \mathbf{m}_M^* \mathbf{m}_M - \frac{1}{2} \sum_{i=0}^{N-1} \frac{1}{\lambda_i + \alpha} = 0, \quad (32)$$

where λ_i is the i -th eigenvalue of $\Phi^* \Phi$. Subsequently, if γ is defined as

$$\gamma \equiv \alpha \mathbf{m}_M^* \mathbf{m}_M = N - \alpha \sum_{i=0}^{N-1} \frac{1}{\lambda_i + \alpha} = \sum_{i=0}^{N-1} \frac{\lambda_i}{\alpha + \lambda_i}, \quad (33)$$

Table 1 Conditions of the computer simulation for validating the proposed method.

Sound signal	6 kHz pure tone
Number of microphones	2
Separation of microphones	5 cm
Sampling rate f_s	8 kHz, 16 kHz
Incident angle of sound	30°
Speed of sound	340 m/s

(a) $f_s = 16$ kHz (b) $f_s = 8$ kHz**Fig. 2** Waveform of the original signal, a 6 kHz sinusoid (dotted line) and the observed signal (solid and dashed lines).

then hyperparameter α is expressed as

$$\alpha = \frac{\gamma}{m_M^* m_M} \quad (34)$$

In a similar manner, the stationary point for β is obtained as

$$\frac{1}{\beta} = \frac{1}{M - \gamma} \sum_{m=0}^{M-1} \{t_l - \Phi(\mathbf{x}_m) \mathbf{w}_M\}^2 \quad (35)$$

Consequently, hyperparameters α and β are obtained using an iterative method with (34), (35) and the following two equations:

$$\mathbf{S}_M^{-1} = \alpha \mathbf{I} + \beta \Phi^* \Phi \quad (36)$$

$$\mathbf{w}_M = \beta \mathbf{S}_M \Phi^* \mathbf{t}_M. \quad (37)$$

3. Computer Simulations

3.1 Fundamental Validation

A simple computer simulation is first conducted to validate the method under the conditions listed in Table 1. Waveform and power spectra of the sinusoidal signal of 6 kHz sampled at different sampling rates are presented respectively in Figs. 2 and 3. It is apparent that in the case of a sampling rate of 8 kHz, the signal appears at 2 kHz because of aliasing. Furthermore, Fig. 4 shows the waveform and

Fig. 3 Power spectrum of (a) original signal and (b) observed signal. **Fig. 4** Waveform and power spectrum of the estimated signal.

power spectrum of the signal estimated using the proposed method. It is apparent that the original signal of 6 kHz is reconstructed. Although the power spectrum is broadened to some extent in Fig. 4 probably because of errors in numerical computation, the level of restored signal is comparable to the original shown in Fig. 3 (a).

We also conducted another simulation where the signal consists of two sinusoidal components of 2 kHz and 6 kHz. When this signal is captured at a sampling rate of 4 kHz, 6 kHz component overlaps 2 kHz component due to aliasing. However, it was confirmed that the method can separately restore each component under such condition.

3.2 Microphone Configuration

Another computer simulation is conducted to illustrate how the performance varies depending on the microphone configuration. For simplification, let us assume that a sound source and a linear microphone array exist, consisting of four microphones in a free field. The sound source is sufficiently distant from the array. Therefore, the sound propagation is regarded as a plane wave. Conditions other than that are presented in Table 2. The signal to be observed is a sinusoidal of frequency higher than the Nyquist frequency determined by the actual sampling rate f_s but less than that of the super-resolution signal f_t . Although the conditions are the same as those presented in Table 2, the microphone configuration was varied in three ways. The modes of variation were the following:

Table 2 Conditions of the computer simulation for examining the effect of microphone configuration.

Desired sampling frequency f_t	8 kHz
Actual sampling frequency f_s	4 kHz
Frequency of the sinusoidal signal	3 kHz
Number of microphones K	4

Fig. 5 Variation of microphone positions.

1. Microphones #2 and #4 were moved in parallel towards #1.
2. Microphone #2 was moved towards #1 and #4 was moved towards the opposite by the same amount.
3. Only microphone #2 was moved towards #1.

Please refer to Fig. 5 for the microphones to be moved and their moving directions, where the microphone separation is normalized by the distance corresponding to a single sampling period. Additionally, it is assumed that the observed signal is distorted by AWGN. The variance of the noise was set as $S/N = 0$ dB. Goodness of estimation was calculated using the maximum of the cross-correlation function between the original and the signal obtained by performing inverse discrete Fourier transform of the estimated model parameter because the estimated model parameter is DFT coefficients of the super-resolution signal.

Figure 6 presents results as a function of the position of microphone #2. The three conditions are depicted respectively in (a), (b), and (c) of Fig. 6. It is noteworthy that in (a) and (b), microphone #4 changed its position as well as microphone #2, but it was fixed in (c). The line with upside-down triangles shows results of the method based on ML. The line with normal triangles shows those on BLR. Also, the line with circle marks, using the right vertical axis, shows condition numbers of matrix $\Phi^* \Phi$. It is apparent that the condition number increases rapidly because some of the microphones are in proximity. It is noteworthy that the right vertical axis is in the logarithmic scale. Whereas the goodness of estimation of the method based on ML worsens as the condition number increases, that on BLR remains greater than 0.6, irrespective of the microphone configuration. This result implies that the method based on BLR can be regarded as more robust than that on ML against additive noise and various system configurations.

3.3 Signal Frequency and Additive Noise

Effects of the signal-to-noise ratio and the frequency of the signal restored are also investigated. Figure 7 presents results as a function of frequency of the signal reconstructed, where the actual sampling rate and USR are fixed

Fig. 6 Estimation accuracy of sinusoidal signal ($S/N = 0$ dB): (a) #2 and #4 in parallel, (b) #2 and #4 in opposite, (c) #2 only.

as 1000 Hz and $1/4$. The *nominal* sampling rate after super-resolution becomes 4000 Hz. The number of microphones is four. Because the reconstruction accuracy relies greatly on the microphone configuration as described in the previous section, microphone positions are randomized every trial. Trials are performed 50 times for each condition with different microphone configurations. Mean values are shown. Standard deviations are shown by error bars. It is shown in Fig. 7 that the signal is restored as long as its frequency is lower than the Nyquist frequency of the *nominal* sampling frequency, and that the method based on BLR outperforms that on ML. Results also show that, as anticipated by Fig. 6, the method based on ML is considerably affected by microphone configuration, leading to longer error bars than that on BLR.

To explore the difference between the methods based

Fig. 7 Estimation accuracy as a function of the frequency of the signal to be reconstructed.

on ML and BLR further, Fig. 7 is rearranged as a function of the condition number of $\Phi^* \Phi$ and is shown in Fig. 8. Two upper panels are for a sinusoidal signal under Nyquist frequency. The lower two are over Nyquist frequency. It is apparent that no discrepancy exists in estimation accuracy between ML and BLR when the condition number is low, but it expands as the condition number increases. This tendency is apparent only when the frequency of the signal is less than the Nyquist frequency.

To investigate the origin of this difference, Fig. 8 is again rearranged as a function of hyperparameter β and is shown in Fig. 9. The hyperparameter β here is not given as a priori knowledge but is estimated from the observed signal to maximize the evidence function. Because the method based on ML uses no hyperparameter β , data are shown with respect to β of the BLR method for the same microphone configuration. It is shown clearly in Fig. 9 that, for conditions where the hyperparameter β converges to a small value, the method on BLR exhibits moderate performance whereas the method on ML fails to restore the signal.

Fig. 8 Estimation accuracy as a function of the condition number of $\Phi^* \Phi$.

Fig. 9 Estimation accuracy as a function of hyperparameter β estimated for maximizing evidence function.

Fig. 10 Estimation accuracy as a function of the number of microphones used.

3.4 Number of Microphones

In this test, the number of microphones was varied up to 50 microphones. Trials were repeated 50 times. The microphone configuration was randomized every trial. The signal frequency was fixed to 1800 Hz for an actual sampling frequency of 500 Hz. Additive white Gaussian noise was added so that signal-to-noise ratio became -10 and 0 dB. Results for up to 25 microphones are depicted in Fig. 10. Beyond this number of microphones, the result increased monotonically but no remarkable behavior was observed. Results show that, as the number of microphones increases, the difference between the ML and BLR becomes small. Moreover, the amount of difference varies in accordance with the condition number. When the condition number is large, the difference also becomes large.

4. Experiments

In the previous section, fundamental characteristics of the method is examined through a sinusoidal signal. Considering that the method is also applicable to a complex signal as described in Sect. 3.1, it is expected that the method can reconstruct a speech signal, too. Therefore, the method is examined through speech signals under a more practical condition in this section.

4.1 Experimental Setup

The method was tested on data that were actually captured using a microphone array in an anechoic room. Signals were captured in the anechoic room of ATR, the size of which is $5.4 \text{ m} \times 4.8 \text{ m} \times 3.8 \text{ m}$. A linear microphone array consisting of 24 microphones, each of which is separated equally with a constant distance of 7.25 cm , was set at almost the center of the room. Speech signals are played through a loudspeaker placed so that the incident angle of the sound to the microphone array becomes 35.97° , as shown in Fig. 11. Assuming the speed of sound as 340.65 m/s , this incident angle produces a special case in which the time of sound traveling between adjacent two microphones equals one sampling period when the sampling rate

Fig. 11 Configuration of the system recording sound signals.

Fig. 12 Maximum of cross-correlation function between signals of each microphone and microphone #1 and the corresponding time delay.

is 8 kHz . Therefore, under the assumption of a plain wave, when sound signals are captured at a sampling rate of 8 kHz , the same waveform is observed at each microphone but at a different time. The difference becomes an integer multiple of the sampling period. Therefore, super-resolution can be achieved using a simple method explained in Sect. 2.1.

Figure 12(a) presents the maximum of the cross-correlation function between signals captured at each microphone and microphone #1. If the assumption of a plain wave is valid, then the line should be completely flat, but it is not flat because of distortion and reflections. Figure 12(b) shows the time shift which provides the maximum of the cross-correlation function in Fig. 12(a). An ideal response should be a linearly declining line, as shown by the broken line in this figure. Results show that the measured data do not hold the ideal relation because the loudspeaker is located rather close to the microphone array as a result of the size limitation of the room. However, to some degree, it shows a desirable property for large microphone numbers, or in other words, as the microphone position becomes distant. Considering this point, only signals captured at 16 microphones from #9 to #24 were used in the processing of super-resolution.

4.2 Mean Opinion Score

Subjective tests were conducted to ascertain that the proposed method can significantly improve sound quality with respect to speech intelligibility. The speech quality was assessed using Mean Opinion Score (MOS) ITU-T P.800 rec-

Table 3 Grade used for mean opinion score.

Score	Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 4 Test signals.

Label	Freq.[Hz]	Description
a	500	Signal captured by a single microphone
b	2k	Reconstructed using 4 microphones
c	8k	Reconstructed using 16 microphones
d	48k	Ideal signal as anchor

ommendation [17]. Because speech intelligibility is our interest, referring to QRK in Q code [18], we instructed participants to assess the sound quality, especially with respect to speech intelligibility. Eight participants, four males and four females, participated in the listening tests. They listened to the test speech signal through a pair of headphones (HDA-200; Sennheiser Electronic GmbH and Co. KG), after which they were asked to score the speech quality using the five grades shown in Table 3. The speech signal was spoken by one male speaker and two female speakers. Its durations were 6 s to 8 s each. Signal levels were adjusted to be equal in terms of the rms value. The absolute sound level was measured using a white Gaussian noise, which has the same rms value as the test signals. It was 56 dB(L_{Aeq}) when measured using an artificial ear wearing the headphone. Because usual consumer products of audio interface have no ability to sample audio signals at a sampling rate of less than 8 kHz, we captured audio signals at a rate of 48 kHz and then thinned them out by 1/96 to create a signal that would be obtained at a sampling rate of 500 Hz. Therefore, although the sampling rate was 500 Hz, the cut-off frequency of the anti-aliasing filter was yet around 24 kHz. Consequently, aliasing components remain in the resultant signals.

In order to focus on fundamental effects of super-resolution, the microphone condition was set not to be ill-conditioned. In this special case, which corresponds to the right part of each panel of Fig. 6, no remarkable difference exists between ML and BLR. Therefore, the method based on ML was used to exclude undesirable factors such as estimation of hyperparameters and validity of assumption in BLR. These factors may also affect the reconstructed speech quality to some extent, and validity of the assumption will be discussed later in Sect. 5. Tested signals are presented in Table 4. These signals were once again up-sampled to 48 kHz and were played through a pair of headphones using an audio interface (MOTU 4PRE). Case (a) in Table 4 corresponds to the case where one microphone signal is eavesdropped. By contrast, case (d) is the case where the original signal is obtained. Cases (b) and (c) are signals that would be obtained by a legitimate user of the system. Test results are presented in Fig. 13. Results show that the super-resolution can significantly improve sound quality with re-

Fig. 13 MOS of signals reconstructed using the conditions shown in Table 4. Asterisks denote a significant difference ($p < 0.01$).

spect to intelligibility as the number of microphone signals increases.

5. Discussion

The proposed method has similar properties to the secret sharing scheme (SSS) [19], where a person who collects a greater number of pieces than a certain threshold can recreate the original data. Inspired by this feature, a privacy preserving video surveillance was also proposed by M. Upmanyu *et al.* [20]. The secret sharing scheme is accompanied by a considerable increase in data size to scramble the original information completely. By contrast, the proposed method requires no increase in data size. This difference derives from the fact that the purpose of the secret sharing scheme is to achieve bit-wise security, whereas that of speech privacy is to make each piece of data useless from the perspective of intelligibility. Therefore, this property is an important benefit of the proposed method compared with SSS.

A shortcoming is that it needs information of the incident angle of sound to be restored. As described in this paper, it is assumed to be known because numerous research works related to direction-of-arrival (DOA) estimation have been published in the relevant literature [21], [22]. It is noteworthy that estimation error in the incident angle brings about deterioration in the sound quality of the reconstructed signal. This problem demands further investigation in a future study.

It should be noted that the number of microphones needs to be greater than USR to make (10) a minimum norm solution. Otherwise, it will become a least squares solution, and complete reconstruction can not be achieved. It is also noteworthy that the microphone configuration more or less affects the resultant performance, even for the method based on BLR. Therefore, particular attention should be devoted to the microphone configuration for better performance in a practical use.

Samples of speech signals are generally described by a Laplacian distribution rather than a Gaussian distribution. For short frame length, such as 2.5 ms, a Gaussian distribution describes speech samples better than a Laplacian distribution [23]. However, for frame lengths from 5 ms

to 200 ms, a Laplacian distribution exhibits better performance. For longer frames, a gamma distribution or generalized Gaussian distribution describes it better. For this study, the frame length was set to 512 samples long, which corresponds to 128 ms for the sampling rate of 8 kHz. Therefore, the Laplacian distribution is the best to describe the probability distribution function of the signal. However, an assumption that DFT coefficients of speech signal have Gaussian probability distribution is used here, as in other papers [24].

As shown in Fig. 7, when the signal frequency is an integer multiple of the actual sampling rate, performance behaves differently, probably because observed discrete signals become constant in this condition. Therefore, for some microphones, the signal-to-noise ratio might be consistently low.

6. Conclusion

We proposed methods of speech privacy for public surveillance using multiple microphones and super-resolution based on maximum likelihood and Bayesian linear regression. A remarkable feature of the method is that the speech is unintelligible at the moment it is captured by an AD converter. Therefore, even if a single node of sensor is hacked by a malicious person through the Internet, speech privacy is protected. Computer simulations conducted to test practicality reveal that both methods can realize the fundamental ability of restoring higher frequency components than Nyquist frequency. Extensive simulations have revealed that the performance can considerably vary depending on the microphone configuration. In general, the method based on Bayesian linear regression is likely to show better performance than the ML method. This difference seems attributable to the behavior of hyperparameters. Moreover, because of fundamental ability, it is possible that speech intelligibility is recoverable even if a signal captured at each microphone is completely unintelligible. Consequently, this method could be used for speech privacy during audio surveillance. This point is also verified through listening tests using the mean opinion score, where an increase in MOS score by 1 to 2 points from 1 is achieved.

Acknowledgments

The authors would like to thank Prof. Suzuki and Sakamoto of the Advanced Acoustic Information Systems Laboratory of the Research Institute of Electrical Communication, Tohoku University for their help in conducting listening tests, and thank Dr. Owada for his support with computer simulations. This work is partly supported by JSPS KAKENHI Grant number 25280060 and 16KT0100.

References

- [1] S. Christopher, "Public privacy: Camera surveillance of public places and the right to anonymity," *Mississippi Law Journal*, vol.72, pp.213–299, 2002.
- [2] K.W. Bowyer, "Face recognition technology: security versus privacy," *IEEE Technol. Soc. Mag.*, vol.23, no.1, pp.9–19, Spring 2004.
- [3] F. Dufaux and T. Ebrahimi, "Scrambling for privacy protection in video surveillance systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol.18, no.8, pp.1168–1174, Aug. 2008.
- [4] X. Yu, K. Chinomi, T. Koshimizu, N. Nitta, Y. Ito, and N. Babaguchi, "Privacy protecting visual processing for secure video surveillance," *Proc. International Conference on Image Processing (ICIP)*, pp.1672–1675, Oct. 2008.
- [5] W. Zhang, S.C.S. Cheung, and M. Chen, "Hiding privacy information in video surveillance system," *Proc. International Conference on Image Processing (ICIP)*, pp.868–871, Sept. 2005.
- [6] F. Dufaux and T. Ebrahimi, "Smart video surveillance system preserving privacy," *Proc. Image and Video Communications and Processing 2005*, pp.54–63, April 2005.
- [7] B.N. Gover and J.S. Bradley, "Measures for assessing architectural speech security (privacy) of closed offices and meeting rooms," *J. Acoust. Soc. Am.*, vol.116, no.6, pp.3480–3490, 2004.
- [8] K. Kasparian and J. Ide, "Variable time inversion algorithm controlled system for multi-level speech security," June 26 1990. US Patent 4,937,867.
- [9] L.-S. Lee, "A speech security system not requiring synchronization," *IEEE Commun. Mag.*, vol.23, no.7, pp.42–55, July 1985.
- [10] J.L. Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama, "Computational auditory induction by missing-data non-negative matrix factorization," *Statistical and Perceptual Audition (SAPA)*, pp.1–6, Sept. 2008.
- [11] B.P. Keegan, S.K. Tjoa, and K.J.R. Liu, "Super-resolution of musical signals using approximate matching pursuit," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp.81–84, Oct. 2011.
- [12] J. Dong, W. Wang, and J. Chambers, "Audio super-resolution using analysis dictionary learning," *International Conference on Digital Signal Processing (DSP)*, pp.604–608, July 2015.
- [13] S.C. Park, M.K. Park, and M.G. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Process. Mag.*, vol.20, no.3, pp.21–36, May 2003.
- [14] R. Nishimura and Y. Suzuki, "Reconstruction of a high-sampling audio signal from low-sampling audio signals using super-resolution," *Proc. Western Pacific Acoustics Conference (WESPAC)*, pp.369–375, June 2006.
- [15] A.V. Oppenheim and R.W. Schaffer, *Digital Signal Processing*, Prentice Hall, 1975.
- [16] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, 2006.
- [17] "Methods for subjective determination of transmission quality," Aug. 1996.
- [18] "Miscellaneous abbreviations and signals to be used in radiotelegraphy communications in the maritime mobile service," 2012.
- [19] A. Shamir, "How to share a secret," *Commun. ACM*, vol.22, no.11, pp.612–613, Nov. 1979.
- [20] M. Upmanyu, A.M. Namboodiri, K. Srinathan, and C.V. Jawahar, "Efficient privacy preserving video surveillance," *Proc. International Conference on Computer Vision (ICCV)*, pp.1639–1646, Sept. 2009.
- [21] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Process.*, vol.33, no.4, pp.823–831, Aug. 1985.
- [22] R.O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol.34, no.3, pp.276–280, March 1986.
- [23] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett.*, vol.10, no.7, pp.204–207, July 2003.
- [24] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol.33, no.2, pp.110–116, Feb. 1985.

[1] S. Christopher, "Public privacy: Camera surveillance of public places and the right to anonymity," *Mississippi Law Journal*, vol.72,

tor,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol.32, no.6, pp.1109–1121, Dec. 1984.

Ryouichi Nishimura Ryouichi Nishimura was conferred a B.S. degree in Information Engineering in 1993 and M.Sc. and Ph.D. degrees in Information Science in 1995 and 1998, respectively, all from Tohoku University, Sendai Japan. During 1998–2000, he was a Visiting Researcher at ATR, Media Information and Communications Research Laboratories, Kyoto, Japan. He was a Research Associate at Tohoku University during 2000–2004, and then an Associate Professor until November 2006. He is currently a Research Expert at the National Institute of Information and Communications Technology (NICT), Sendai Japan, where he has been working mostly on audio signal processing.

Seigo Enomoto received his B.E. degree from Kinki University in 1997, his M.E. degree from Nara Institute of Science and Technology (NAIST) in 1999, and his Ph.D. degree from Kyoto University in 2005. During 2005–2009, he worked as a researcher at the Spoken Language Communication Lab., Advanced Telecommunications Research Institute International (ATR-SLC). He is currently a Research Expert at the National Institute of Information and Communications Technology (NICT), where he is working on 3D sound reproduction and spatial hearing.

Hiroaki Kato received B.A. and M.A. degrees in Instrumentation Engineering and a Ph.D. in Information and Media Science from Kobe University, in 1986, 1988, and 1999, respectively. He has been working at ATR Laboratories, Kyoto, Japan, since April 1988. In 2006 he joined NICT as a research expert, where he is engaged in research on spatial hearing and three-dimensional sound reproduction.