PAPER Special Section on Human Cognition and Behavioral Science and Technology

Enhancing Event-Related Potentials Based on Maximum a Posteriori Estimation with a Spatial Correlation Prior

Hayato MAKI[†a)], ***Nonmember***, **Tomoki TODA**[†,††], **Sakriani SAKTI**[†], ***Members***, **Graham NEUBIG**[†], ***Nonmember***, ***and*** **Satoshi NAKAMURA**[†], ***Member***

SUMMARY In this paper a new method for noise removal from single-trial event-related potentials recorded with a multi-channel electroencephalogram is addressed. An observed signal is separated into multiple signals with a multi-channel Wiener filter whose coefficients are estimated based on parameter estimation of a probabilistic generative model that locally models the amplitude of each separated signal in the time-frequency domain. Effectiveness of using prior information about covariance matrices to estimate model parameters and frequency dependent covariance matrices were shown through an experiment with a simulated event-related potential data set.

key words: *electroencephalogram (EEG), event-related potential (ERP), generative model, independent component analysis (ICA), Wiener filter, noise removal, Wishart distribution, spatial correlation prior*

1. Introduction

Electroencephalogram (EEG) is one of the useful tools to investigate human brain activity non-invasively, along with functional magnetic resonance imaging (fMRI), near infra-red spectroscopy (NIRS), and magnetoencephalography (MEG). Compared to other methods, EEG has advantages such as relatively low cost, high temporal resolution, and experimental flexibility. Event-related potential (ERP) is a brain response to an internal or external event such as visual stimuli or cognitive processes. ERP is a useful tool for cognitive neuroscience, and often used for manipulation of a brain-machine interface (BMI). However, analysis of ERP data usually suffers from its low signal-to-noise ratio (SNR) due to superimposition of task-unrelated brain activities as well as non-neural artifacts such as eye blinks, muscle movements, and heart beats.

Signal averaging in the time domain is often employed in order to improve SNR of ERP by attenuating task-unrelated background EEGs. In a usual ERP data analysis, the whole EEG recording during an experiment is cut into small intervals containing a single stimulus. Each of these intervals is called a trial or an epoch. ERP components are assumed to be identical over each trial, while background EEGs are assumed to be arbitrary phase with regards to the stimuli. Therefore, taking the average of all trials aligned to the onsets of stimuli of each trial will weaken the amplitude of background EEGs while preserving that of ERPs.

However, variabilities of ERP across trials exist in latency, amplitude, and scalp distribution [1]. This across-trial averaging procedure sacrifices all the information concerning across-trial variability of brain response. Moreover, to perform reliable analysis at least 20 trials for each experimental condition are required to be averaged, resulting in long experimental time and subject fatigue. Therefore, single-trial analysis is an important research topic to study event-related brain dynamics more precisely [2]–[4]. Throughout this paper, we refer to the ERPs as the *signal* to preserve and task-unrelated neural activities as well as non-neural artifacts as *noise* to be removed.

With regards to external artifact removal, especially eye blink removal, there is a large amount of previous research, and several effective methods [5]. However, background EEG removal is less studied, although there is some research proposing methods based on independent component analysis (ICA) [6], [7].

In this paper, a new approach to remove background activities from single-trial ERPs is addressed. The proposed method is an extension of a framework that has emerged recently in under-determined sound source separation [8], [9]. While ICA aims to separate an observed signal into its sources, this framework separates into signals elicited by each event such as ERP or eye blinks. We denote the group of source indexes that contributes to the k-th event by E_k and denote the EEG signal that is elicited by E_k and captured by $\mathbf{c}_k(t)$ and k-th event signal. In the time-frequency domain, short-time Fourier transform (STFT) coefficients of the event signal $\mathbf{c}_k(n, f)$ are locally modeled by a multivariate complex Gaussian distribution whose parameters are a function of (n, f), where n is the index of the time frame and f is the index of the frequency bin. One of the parameters is a full rank time-invariant covariance matrix that is called a *spatial correlation matrix* because it encodes the spatial spread of the event signal. The model parameters are estimated by the maximum likelihood criterion and used to separate an observed multi-channel signal into some event signal by a multi-channel Wiener filter. This framework was applied to EEG signal separation where it was used to remove external ocular artifacts from alpha band rhythmic activity [10] and background activities from an

Manuscript received October 31, 2015.
Manuscript revised March 11, 2016.
Manuscript publicized April 1, 2016.

[†]The authors are with Nara Institute of Science and Technology, Ikoma-shi, 630–0101 Japan.
[††]The author is with Nagoya University, Nagoya-shi, 464–8601 Japan.
a) E-mail: maki.hayato.lt3@is.naist.jp

DOI: 10.1587/transinf.2015CBP0008

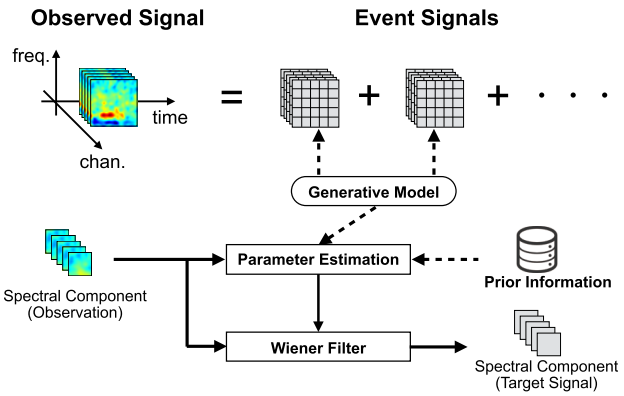

Fig. 1 Overview of the proposed method.

ERP signal [11]†.

This approach is extended using prior information. This paper provides two main contributions. First, the effectiveness of using prior information to estimate covariance matrix parameters is demonstrated. Second, the effectiveness of frequency dependent spatial correlation matrices are examined, in contrast to [10], which uses frequency independent ones. The overview of our proposed method is shown in Fig. 1.

2. Existing Techniques for Signal Enhancement

In this section, we describe two different types of signal separation techniques, independent component analysis in Sect. 2.1 and the maximum likelihood approach in Sect. 2.2.

2.1 Independent Component Analysis

ICA is the most commonly used technique in the context of EEG signal separation. ICA generally models an observed EEG signal $\mathbf{x}(t)$ at time t as an instantaneous linear combination of M sources that is captured by L sensors as follows:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) = \sum_{i=1}^{M} \mathbf{a}_i s_i(t), \tag{1}$$

where $\mathbf{A}$ is a $L \times M$ mixing matrix whose i-th column is $\mathbf{a}_i$. ICA is an algorithm that estimates a mixing matrix $\mathbf{A}$ given only observation $\mathbf{x}$ such that sources $\mathbf{s}$ are as statically independent as possible. An estimation of sources $\mathbf{s}$ is obtained by multiplying the inverse of the mixing matrix with the observed signal as follows:

$$\hat{\mathbf{s}}(t) = \hat{\mathbf{A}}^{-1}\mathbf{x}(t). \tag{2}$$

We refer to estimated sources $\hat{\mathbf{s}} = (\hat{s}_1, \cdots, \hat{s}_L)^\top$ as independent components (ICs) and $\hat{\mathbf{A}}$ is an estimated mixing matrix whose i-th column is $\hat{\mathbf{a}}_i$. In order to obtain a noise-free ERP signal $\mathbf{c}$, it is necessary to select ICs corresponding to the ERP. We denote the group of source indexes corresponding to the ERP by E. Then, signal reconstruction is done as follows:

$$\mathbf{c} = \tilde{\mathbf{A}}\hat{\mathbf{s}} = \sum_{i=1}^{L} \tilde{\mathbf{a}}_i \hat{s}_i(t),\ \tilde{\mathbf{a}}_i = \begin{cases} \mathbf{0} & (i \notin E) \\ \hat{\mathbf{a}}_i & (i \in E) \end{cases} \tag{3}$$

In general, it is assumed that the number of sensors is equal to that of sources because the linear mixture is invertible in such a case. However, there is no reason to believe the number of EEG sources is equal to that of the channels of the EEG recording equipment. In fact, EEG source separation is often treated as an *under-determined problem*, implying the condition $M > L$ [7], [12]. In such a case, additional assumptions or prior knowledge about the underlying sources are usually exploited. For example, in [7], they used the following function $L_\lambda(\mathbf{x})$ to incorporate the prior information about an averaged ERP:

$$L_\lambda(\mathbf{x}^k) : \mathbf{x}^k \longrightarrow (1-\lambda)\mathbf{x}^k + \lambda\bar{\mathbf{x}},\ \lambda \in [0, 1] \tag{4}$$

where $\mathbf{x}^k$ is the k-th trial, $\bar{\mathbf{x}}$ is the averaged signal across trials and λ is the regularization parameter. They applied the ICA algorithm to $L_\lambda(\mathbf{x}^k)$ instead of $\mathbf{x}^k$ to estimate the mixing matrix, then decomposed each $\mathbf{x}^k$ using the inverse of the estimated matrix. Another problem of ICA is permutation ambiguity. The ICA algorithm doesn't tell us which ICs correspond to the signal of interest, which decreases the practical usability of ICA. The component classification is required to reconstruct a noise free signal that can be performed by visual inspection [3], [7], predefined [13], [14] or adaptive [15] thresholding.

2.2 Maximum Likelihood Approach

This approach has been proposed originally in the field of sound source separation [9] but its effectiveness for EEG signals has not been explored extensively. This approach has a virtue that, in contrast to ICA, it doesn't assume the number of sources because it doesn't estimate sources but contributions of each event signal to each EEG electrode.

2.2.1 Observation Model

The STFT coefficients of the k-th event signal in time-frequency slot (n, f) are expressed as $\mathbf{c}_k(n, f)$:

$$\mathbf{c}_k(n, f) = \sum_{l \in E_k} \mathbf{h}_l s_l(n, f), \tag{5}$$

$$\mathbf{c}_k(n, f) = [c_{k,1}(n, f), \cdots, c_{k,L}(n, f)]^\top,$$
$$\mathbf{h}_l = [h_{1l}, \cdots, h_{Ll}]^\top, \tag{6}$$

where h_{il} is the transfer function from the l-th source to the i-th channel assuming that $0 \le h_{il} \le 1$. The observed multichannel EEG signal $\mathbf{x}(n, f)$ is expressed as

$$\mathbf{x}(n, f) = [x_1(n, f), \cdots, x_L(n, f)]^\top = \sum_{k=1}^{K} \mathbf{c}_k(n, f), \tag{7}$$

†In this paper, we present further details of the proposed methods, more discussions, and more evaluations than those in our previous work [11].

where K is the total number of events that should be given as a hyper-parameter.

Recall that in ICA's observation model, observed signals are expressed as the sum of the product of time-variant scalars and time-invariant vectors in (1). Here, instead the observed signals are expressed as the sum of time-variant vectors, which means that signal separation of event signals that take this form can not be achieved by ICA.

2.2.2 Generative Model

To model the generation of observed signals probabilistically, we make the following two assumptions:

Assumption1
The amplitude of the l-th source that contributes to the k-th event in each slot (n, f) follows a complex normal distribution with mean 0. The variance is given by the product of the degree of the event activity $v_k(n, f)$ and the source activity λ_l:

$$\begin{aligned} p\left(s_l(n,f)\right) &= \mathcal{N}_c\left(s_l(n,f); 0, v_k(n,f)\lambda_l\right) \\ &= \frac{1}{\pi v_k(n,f)\lambda_l} \exp\left(\frac{|s_l(n,f)|^2}{v_k(n,f)\lambda_l}\right), \end{aligned} \tag{8}$$

where $l \in E_k$ and the group E_k contains source indexes associated with the k-th event signal.

Assumption2
Sources are not correlated with each other within each time-frequency slot (n, f):

$$E[s_{l_1}(n,f)s^*_{l_2}(n,f)] = 0 \text{ for } l_1 \neq l_2, \tag{9}$$

where $l_1, l_2 \in E_k$ and $\{\cdot\}^*$ indicates the complex conjugate of $s_{l_2}(n, f)$.

From the assumptions above, the probability density function that each event signal follows is derived, namely the multivariate complex normal distribution with zero mean.

$$\begin{aligned} p\left(\mathbf{c}_k(n,f)\right) &= \mathcal{N}_c(\mathbf{c}_k(n,f); \mathbf{0}, \mathbf{R}_{c_k}(n,f)) \\ &= \frac{\exp(-\mathbf{c}_k(n,f)^H \mathbf{R}^{-1}_{c_k}(n,f)\mathbf{c}_k(n,f))}{\pi^L \det(\mathbf{R}_{c_k}(n,f))}. \end{aligned} \tag{10}$$

For simplicity, we restrict the covariance matrix $\mathbf{R}_{c_k}(n, f)$ to the product of the time-invariant covariance matrix $\mathbf{R}_k$ that encodes spatial spread of the k-th event signal and the time-frequency variant scalar $v_k(n, f)$ that encodes time-frequency power of the signal as follows:

$$\begin{aligned} \mathbf{R}_{c_k}(n,f) &= E\left[\mathbf{c}_k(n,f)\mathbf{c}_k(n,f)^H\right] \\ &= E\left[\sum_{l_1 \in E_k} \mathbf{h}_{l_1} s_{l_1}(n,f) \left(\sum_{l_2 \in E_k} \mathbf{h}_{l_2} s_{l_2}(n,f)\right)^H\right] \\ &= \sum_{l_1 \in E_k} \sum_{l_2 \in E_k} E[s_{l_1}(n,f)s^*_{l_2}(n,f)]\mathbf{h}_{l_1}\mathbf{h}^\top_{l_2}] \\ &= \sum_{l \in E_k} E[|s_l(n,f)|^2]\mathbf{h}_l\mathbf{h}_l^\top \\ &= \sum_{l \in E_k} v_k(n,f)\lambda_l \mathbf{h}_l\mathbf{h}_l^\top \\ &= v_k(n,f)\mathbf{R}_k, \end{aligned} \tag{11}$$

where $\{\cdot\}^H$ is the complex conjugate transpose. We call $\mathbf{R}_k$ the *spatial correlation matrix* of the k-th event signal.

2.2.3 Event Sparseness

To simplify the generative model, we assume that the events in each time frequency slot are sparse. In other words, we assume that we observe only one event signal in each slot. In order to express this sparseness mathematically, we introduce the latent variables $z_k(n, f)$, which take a 1-of-K representation, in which one particular element $z_k(n, f)$ is equal to 1 and all other elements are equal to 0. In other words, the value of $z_k(n, f)$ satisfies

$$z_k(n,f) \in \{0,1\}, \ \sum_{k=1}^{K} z_k(n,f) = 1. \tag{12}$$

If we observe the l-th event signal in the slot (n, f), we can express the observed signal as:

$$\begin{aligned} \mathbf{x}(n,f) &= [x_1(n,f), \cdots, x_L(n,f)]^\top \\ &= \sum_{k=1}^{K} z_k(n,f)\mathbf{c}_k(n,f) = \mathbf{c}_l(n,f), \end{aligned} \tag{13}$$

where $z_l(n, f) = 1$.

2.2.4 Observation Likelihood

From all assumptions above, finally we derive the likelihood of the observed signals $\mathbf{x}$ in the time-frequency slot (n, f), which is modeled by a Gaussian mixture model as follows:

$$\begin{aligned} p(\mathbf{x}|\theta) &= \prod_{n,f} p(\mathbf{x}(n,f)|\theta) \\ &= \prod_{n,f} \sum_{k=1}^{K} \alpha_k \mathcal{N}_c\left(\mathbf{x}(n,f); \mathbf{0}, v_k(n,f)\mathbf{R}_k\right) \end{aligned} \tag{14}$$

$$p(z_k(n,f) = 1) = \alpha_k \text{ for all } (n,f), \tag{15}$$

where α_k is the prior probability that the k-th event signal is observed assuming that their values are shared among all slots. The model parameter set θ consists of α_k, $v_k(n, f)$, and $\mathbf{R}_k$ of each mixture component.

2.3 Event Separation

The optimal model parameter set θ is estimated by maximizing the observation likelihood described in Eq. (14). This maximization process can also be effectively solved with the expectation-maximization (EM) algorithm. θ is chosen by maximizing the following auxiliary function, which is the conditional expectation of the complete-data log likelihood:

$$Q = \sum_{n,f,k} m_k(n,f)\Big(\log(\alpha_k) - L\log(v_k(n,f)) + \log(|\mathbf{R}_k^{-1}|) - \frac{1}{v_k(n,f)}\mathbf{x}(n,f)^H\mathbf{R}_k^{-1}\mathbf{x}(n,f)\Big) + \text{Const}, \tag{16}$$

where in the E-step, the posterior probability $m_k(n,f)$ is calculated at each time-frequency slot:

$$m_k(n,f) = \frac{\alpha_k \mathcal{N}_c\left(\mathbf{x}(n,f); \mathbf{0}, v_k(n,f)\mathbf{R}_k\right)}{\sum_{k'=1}^{K} \alpha_{k'} \mathcal{N}_c\left(\mathbf{x}(n,f); \mathbf{0}, v_{k'}(n,f)\mathbf{R}_{k'}\right)}. \tag{17}$$

In the M-step, $\hat{\alpha}_k$ and $\hat{v}_k(n,f)$ are updated as follows:

$$\hat{\alpha}_k = \frac{\sum_{n,f} m_k(n,f)}{\sum_{n,f,k'} m_{k'}(n,f)}, \tag{18}$$

$$\hat{v}_k(n,f) = \frac{1}{L}\mathbf{x}(n,f)^H\mathbf{R}_k^{-1}\mathbf{x}(n,f), \tag{19}$$

$$\hat{\mathbf{R}}_k = \frac{\sum_{n,f} \frac{m_k(n,f)}{\hat{v}_k(n,f)}\mathbf{x}(n,f)\mathbf{x}(n,f)^H}{\sum_{n,f} m_k(n,f)}, \tag{20}$$

where $\hat{v}_k(n,f)$ and $\hat{\mathbf{R}}_k(f)$ are iteratively updated because they depend on each other.

Finally, the target event signal is extracted from the observed EEG signals using a multi-channel Wiener filter. A multi-channel Wiener filter $\hat{\mathbf{M}}_k \in \mathcal{R}^{L\times L}$ is a linear filter that minimizes the square errors as follows:

$$\hat{\mathbf{M}}_k = \arg\min_{\mathbf{M}_k} \|\mathbf{M}_k\mathbf{x}(n,f) - \hat{\mathbf{c}}_k(n,f)\|^2 \tag{21}$$

where $\hat{\mathbf{c}}_k(n,f)$ is the k-th event signal separated from $\mathbf{x}(n,f)$. Solving this minimization problem, we obtain

$$\hat{\mathbf{M}}_k(n,f) = \mathbf{R}_{\mathbf{c}_k\mathbf{x}}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f), \tag{22}$$

where $\mathbf{R}_{\mathbf{ab}} = \mathrm{E}[\mathbf{a}\mathbf{b}^\top]$ and $\mathbf{R}_{\mathbf{a}} = \mathrm{E}[\mathbf{a}\mathbf{a}^\top]$. Assuming event signals are independent from each other, $\mathbf{R}_{\mathbf{c}_k(n,f)\mathbf{x}}(n,f)$ can be expressed as:

$$\begin{aligned}\mathbf{R}_{\mathbf{c}_k(n,f)\mathbf{x}}(n,f) &= [\mathbf{c}_k(n,f)\mathbf{x}(n,f)^\top] \\ &= \left[\mathbf{c}_k(n,f)\mathbf{c}_k(n,f)^\top\right] \\ &= \mathbf{R}_{\mathbf{c}_k}(n,f) \\ &= m_k(n,f)\hat{v}_k(n,f)\hat{\mathbf{R}}_k.\end{aligned} \tag{23}$$

Consequently, an event signal is estimated as follows:

$$\hat{\mathbf{c}}_k(n,f) = m_k(n,f)\hat{v}_k(n,f)\hat{\mathbf{R}}_k\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f) \tag{24}$$

3. Proposed Method

We often know which event signal we would like to enhance, such as ERP, and we can record EEG signals related to the target event beforehand and use them as prior knowledge for enhancement. In such a case, we need to enhance the target event signal from the observed EEG signal, rather than blindly separating it into multiple EEG event signals described in the previous section. Hence, we introduce prior informations of event signals to estimate the spatial correlation matrices utilizing the property of the Wishart distribution.

3.1 Spatial Correlation Prior

The Wishart distribution is a type of probability distribution, whose probability density function is given by

$$\mathcal{W}_p(\mathbf{W}, q) = B(\mathbf{W}, p, q)|\mathbf{A}|^{\frac{q-p-1}{2}}\exp\left(-\frac{\mathrm{Tr}(\mathbf{W}^{-1}\mathbf{A})}{2}\right) \tag{25}$$

where $\mathbf{W}$ is a $p \times p$ positive definite matrix and q is called the *number of degrees of freedom* and is restricted to $q \geq p$, and $B(\mathbf{W}, p, q)$ is a normalizing factor defined by

$$B(\mathbf{W}, p, q) = |\mathbf{W}|^{-\frac{q}{2}}\left(2^{\frac{pq}{2}}\pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma\left(\frac{q+1-i}{2}\right)\right)^{-1}. \tag{26}$$

where $\Gamma(\cdot)$ is the Gamma function.

The Wishart distribution is the conjugate prior distribution of a multivariate Gaussian random variable's covariance matrix. Assume that we observe b-dimensional random vectors that follow the normal distribution with the known mean vector μ and the unknown precision matrix Λ, and want to infer Λ. In such a case, the conjugate prior distribution for Λ is the Wishart distribution with a hyper parameter of $b \times b$ matrix [16].

3.2 Maximum a Posteriori Estimation of Spatial Correlation Matrices

From the Bayes' theorem, the posterior probability density function of the time-frequency invariant spatial covariance matrices $\mathbf{R} = \{\mathbf{R}_1, \cdots, \mathbf{R}_K\}$ is proportional to the product of the likelihood of the observed data and their prior as follows:

$$p(\mathbf{R}|\mathbf{x}, \theta_{\backslash\mathbf{R}}, \mathbf{\Psi}_k, q_k, w_k) \tag{27}$$

$$\propto \prod_{n,f} p\left(\mathbf{x}(n,f)|\theta\right)\prod_{k=1}^{K}\mathcal{W}_L\left(\mathbf{R}_k^{-1}|\mathbf{\Psi}_k^{-1}, q_k\right)^{w_k} \tag{28}$$

where $\mathbf{x}$ is the M-dimensional observation signal, $\theta_{\backslash\mathbf{R}}$ is the model parameter set except for $\mathbf{R}$, and $w_k \geq 0$ is the weighting parameter that expresses our belief about the strength of the prior. If we have little confidence in the prior of $\mathbf{R}_k$, we set a small value for w_k and vice versa. When $w_k = 0$, this formula reduces to maximum likelihood (ML) estimation.

3.3 Hyper Parameter Learning

We denote the j-th pre-recorded signal that contains the k-th event at time t by $\mathbf{c}_k^{(j)}(t)$, although we can't observe the single event signal itself because EEG signals always contain multiple event signals. However, we can design an experiment that elicits a specific ERP component, and we can

record EEG signals intentionally contaminated by a specific artifact e.g., eye blinks. Thus we can record EEG signals in which a specific event signal is supposed to be observed and also record EEG signals that are not supposed to be observed to establish a contrast between them. In addition, we can assume that we can obtain a noise-reduced event signal by synchronous averaging if the event signal is time-locked e.g., ERPs.

The hyper parameters are calculated by following steps:

1: {Initialization}

$$\bar{v}_k(n,f) \leftarrow 1 \tag{29}$$

2: **repeat**

3: {Iterative update}

$$\boldsymbol{\Psi}_k(f) \leftarrow \left(\sum_{n,f} \frac{\bar{\mathbf{c}}_k(n,f)\bar{\mathbf{c}}_k^H(n,f)}{\bar{v}_k(n,f)} \right) / N$$

$$\hat{v}_k(n,f) \leftarrow \left(\bar{\mathbf{c}}_k(n,f)^H \boldsymbol{\Psi}_k^{-1} \bar{\mathbf{c}}_k(n,f) \right) / L$$

4: **until** convergence

where $\bar{\mathbf{c}}_k(n,f)$ is a signal that contains the k-th event signal, N is the total number of time frames, $\bar{v}_k(n,f)$ is an estimation of the scaling parameter $v_k(n,f)$. If the k-th event signal is time-locked e.g., ERP, we use the averaged signal $\mathrm{E}_j[\hat{\mathbf{c}}_k^{(j)}(t)]$ as the template of the k-th event signal to calculate hyper parameters where $\hat{\mathbf{c}}_k^{(j)}(t)$ is the j-th trial signal that contains the k-th event signal:

$$\bar{\mathbf{c}}_k(t) = \mathrm{E}_j[\hat{\mathbf{c}}_k^{(j)}(t)]. \tag{30}$$

If the k-th event signal is not time-locked e.g., background EEGs or eye blinks, we use the concatenated signal as follows:

$$\bar{\mathbf{c}}_k = [\hat{\mathbf{c}}^{(1)}(1) \cdots \hat{\mathbf{c}}^{(1)}(t), \cdots, \hat{\mathbf{c}}^{(j)}(1) \cdots \hat{\mathbf{c}}^{(j)}(t)]^\top. \tag{31}$$

3.4 Event Separation

The optimal model parameter set θ is estimated by maximizing the posterior probability described in Eq. (27) instead of observation likelihood described in Eq. (14). This maximization process can also be effectively solved with the EM algorithm. The following auxiliary function is maximized with respect to θ.

$$\begin{aligned} Q = \sum_{n,f,k} m_k(n,f) \Bigg(& \log(\alpha_k) - L\log(v_k(n,f)) \\ & + \log(|\mathbf{R}_k^{-1}|) - \frac{1}{v_k(n,f)} \mathbf{x}(n,f)^H \mathbf{R}_k^{-1} \mathbf{x}(n,f) \Bigg) \\ & + \sum_{k=1}^{K} w_k \left(\frac{q_k - L - 1}{2} \log|\mathbf{R}_k^{-1}| - \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\Psi}_k \mathbf{R}_k^{-1}\right] \right) \\ & + \mathrm{Const.} \end{aligned} \tag{32}$$

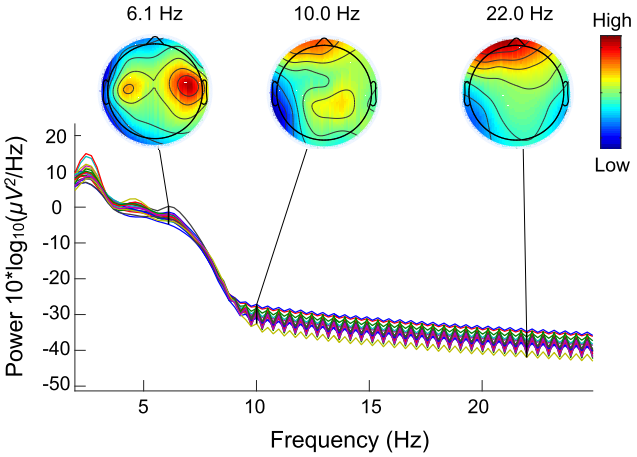

Fig. 2 Scalp topographies within the different frequency bins. Each line stands for an EEG channel.

In the E-step, the posterior probability $m_k(n,f)$ is calculated at each time-frequency slot as shown in Eq. (18). In the M-step, $\hat{\alpha}_k$ and $\hat{v}_k(n,f)$ are updated as shown in Eqs. (18) and (19). The spatial correlation matrices are updated as follows:

$$\hat{\mathbf{R}}_k = \frac{\left(w_k \boldsymbol{\Psi}_k + 2\sum_{n,f} \frac{m_k(n,f)}{\hat{v}_k(n,f)} \mathbf{x}(n,f)\mathbf{x}(n,f)^H \right)}{w_k(q_k - L - 1) + 2\sum_{n,f} m_k(n,f)}. \tag{33}$$

where $\hat{v}_k(n,f)$ and $\hat{\mathbf{R}}_k$ each other. Finally, the target event signal is extracted from the observed EEG signals using a multi-channel Wiener filter constructed according to Eqs. (21), (22), and (23).

3.5 Frequency Dependent Spatial Correlation Matrix

EEG signals usually have different spatial spread over the scalp according to the frequency as shown in Fig. 2. Hence, another generative model with frequency dependent spatial correlation matrices $\mathbf{R}_k(f)$ and $\boldsymbol{\Psi}_k(f)$ instead of $\mathbf{R}_k$ and $\boldsymbol{\Psi}_k$ is proposed. In this manner, the estimation of spatial covariance matrices in the EM algorithm is done as follows:

$$\hat{\mathbf{R}}_k(f) = \frac{\left(\Psi_k(f) + 2\sum_{n} \frac{m_k(n,f)}{\hat{v}_k(n,f)} \mathbf{x}(n,f)\mathbf{x}(n,f)^H \right)}{(q_k(f) - L - 1) + 2\sum_{n} m_k(n,f)}. \tag{34}$$

Note that the range of summation is over only time frame n.

4. Experiments

To validate the performance of the proposed methods, we applied the methods to pseudo-ERP data, which were made by the superposition of measured background EEGs (*noise*) and the measured and averaged ERP of P300.

4.1 Subjects and EEG Recording

EEG data were collected from three healthy male volunteers aged from 23 to 26 years. All participants gave written informed consent and the Ethical Review Board of the Nara

Institute of Science and Technology approved all experimental procedures.

All EEG signals were recorded from 22 scalp electrodes at locations based on a modified International 10-20 system [17], digitally sampled at 1000 Hz, downsampled to 250 Hz referred to the average of the both sides of ears, and filtered between 0.01 Hz and 30 Hz.

4.2 Experimental Paradigm

We conducted two sessions for each subject. In the first session, we conducted an auditory oddball paradigm experiment [18] using 1 kHz and 2 kHz sound. A random sequence of auditory stimuli including 2 kHz and 1 kHz sine waves was presented to the subject with an earphone. All the stimuli had the duration of 0.2 seconds and were presented at intervals of 1.4 seconds. 1 kHz sounds were presented 200 times as non-target stimuli and 2 kHz sine wave were presented 50 times as target stimuli. Subjects sat in the front of a computer display and were told to look at the fixation mark in the display and count the number of target stimuli without any body movement during the session. At the onset of each stimulus, the two small areas corresponding to two types of stimuli in the corner of the display flashed. The photo sensor sent a trigger marker that marked the type of stimulus presented and its onset time to the EEG recording computer when it caught the flash. The flashes were hidden from the subjects using a thick paper to avoid distracting them.

In the second session, the subjects were told to relax without any body movement to record resting state EEG for two minutes.

4.3 Parameter Settings

We implemented regularized ICA based on [7]. We decomposed each raw single-trial multi-channel EEG signal into two event signals with time-frequency modeling methods. Hyper-parameters of the prior distribution for the first event signal was calculated from the signal that were given by averaging the validation dataset as described in Sect. 3.3. Hyper-parameters for the second event signal were calculated from the resting state EEG signal recorded in the second session. All hyper-parameters are calculated individually for each subject.

4.4 Creation of Validation Dataset

The multi-channel EEG data obtained from session 1 of each type of stimulus were cut into trials of 1.2 second length, then averaged the trials across the trials for each channel, subject and type of stimulus. Next, the resting state EEG obtained from session 2 was also cut into 40 signals of 1.2 second length. The averaged target trial ERP of each corresponding channel that stands for *signal* was added to the each of 40 resting state trial that stands for *noise* with randomly shifted phase from −60 ms to 60 ms to make target-trial pseudo-ERP data set. The averaged non-target trial ERP was also added to resting state signals as same as the target trial ERP to make non-target trial pseudo-ERP data set. The averaged and embedded ERP for each type of stimulus and an example of resting state trial were shown in Fig. 3.

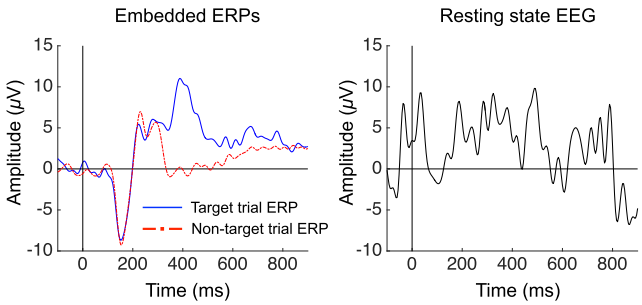

Fig. 3 The target trial (left, solid blue line) and non-target trial (left, dashed red line) ERPs that are embedded into each of resting state signal (right), respectively.

Table 1 Compared methods.

	Prior	Frequency dependency
Conventional method	No	No
Proposed method 1	Yes	No
Proposed method 2	Yes	Yes

4.5 Evaluation

We compared five denoising methods, regularized ICA [7] and time-frequency modeling methods that have a spatial correlation matrix with or without prior information and frequency dependency (See Table 1).

We evaluate these methods for each subject using amplitude deviation (AD), latency deviation (LD) and root mean square errors (RMSE) defined as follows because ERPs are often quantified by their peak amplitude and peak latency,

$$\mathrm{AD} = \frac{1}{J}\sum_{j=1}^{J} |\bar{\mathbf{x}}(\bar{t}_{peak}) - \hat{\mathbf{x}}^{(j)}(t_{peak})| \tag{35}$$

$$\mathrm{LD} = \frac{1}{J}\sum_{j=1}^{J} |\bar{t}_{peak} - \hat{t}^{(j)}_{peak}| \tag{36}$$

$$\mathrm{RMSE} = \frac{1}{J}\sum_{j=1}^{J} \sqrt{\sum_{t}^{T} \bar{\mathbf{x}}(t) - \hat{\mathbf{x}}(t)} \tag{37}$$

where $\bar{\mathbf{x}}$ is the averaged ERP signal obtained from session 1 added to the resting state EEGs and, $\hat{\mathbf{x}}$ is the j-th single-trial denoised EEG signal, J is the total number of trials, T is the signal length, and $\bar{t}_{peak}$ is the time when the averaged ERP signal $\bar{\mathbf{x}}$ has the largest amplitude and $\hat{t}^{(j)}_{peak}$ is the time when the j-th single-trial denoised EEG signal $\hat{\mathbf{x}}^{(j)}$ does.

5. Results

Figure 4 shows the plots of a single-trial EEG signal of

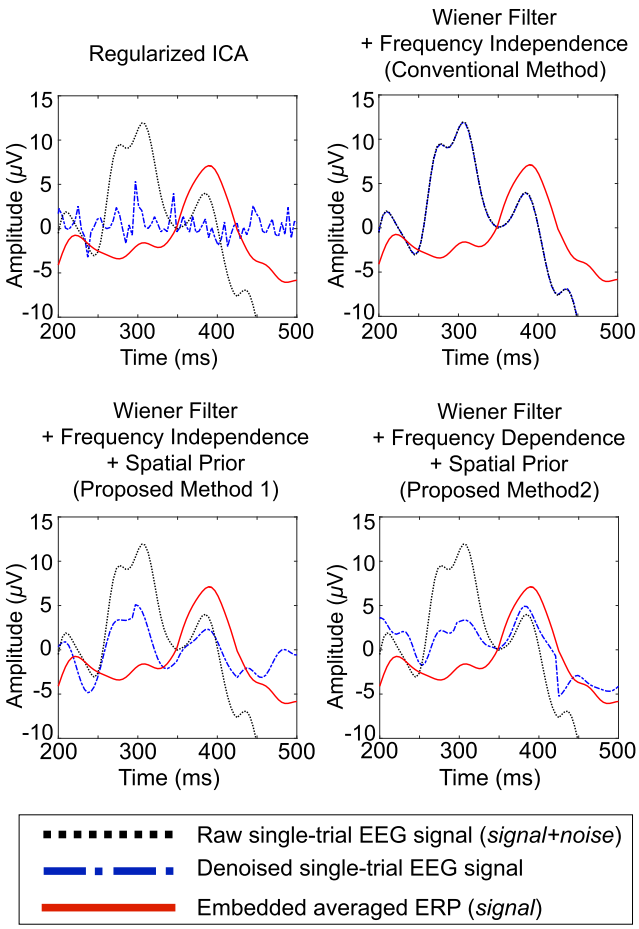

Fig. 4 Plots of a raw signal, a denoised signal, and the embedded averaged ERP: Comparison of regularized ICA (upper left), Winer Filter (Conventional Method, upper right), Winer Filter + Spatial Prior (Proposed Method 1, lower left), and Winer Filter + Spatial Prior + Frequency Dependent please describe the (Proposed Method 2, lower right).

the validation dataset, a single-trial denoised signal, and the averaged ERP signal that is embedded into the validation dataset. The raw signal has two main peaks about 300ms and 380ms. Compared to the averaged ERP, the former peak is eliminated by averaging while the latter peak still exists in the averaged ERP. Hence, we can think that the former peak is made by chance due to background EEG to be removed by the denoising methods while the latter peak is due to embedded ERP to be preserved. First, looking at the ICA results, we can see that it removed the ERP component as well as the background EEG. The conventional method, Wiener filtering without spatial correlation prior, on the other hand, preserved the noise as well as the ERP component. Both of the proposed method 1, which uses prior information and frequency independent spatial correlation matrices, and the proposed method 2, which uses prior information and frequency dependent spatial correlation matrices removed the noise from the ERP. However, noise removal done by the proposed method 1 was not enough to make the first peak smaller than the second, resulting in worse estimation of the peak latency. The proposed method 2 managed to remove the first peak by an amount sufficient to make the first peak smaller than the second, allowing for better estimation of the peak amplitude and latency, although it was not able to remove the first peak entirely.

Table 2 The result of the multiple comparison test (Target trials).

Method 1	Method 2	p-val (AD)	p-val (LD)	p-val (RMSE)
Raw	ICA	**0.000	**0.000	0.994
Raw	Conv	0.952	1.000	0.955
Raw	Pro 1	**0.004	0.449	0.252
Raw	Pro 2	**0.000	1.000	*0.045
ICA	Conv	**0.000	**0.000	0.796
ICA	Pro 1	**0.001	0.075	0.105
ICA	Pro 2	*0.011	**0.000	*0.013
Conv	Pro 1	*0.039	0.410	0.679
Conv	Pro 2	**0.004	1.000	0.238
Pro 1	Pro 2	0.960	0.533	0.947

Table 3 The result of the multiple comparison test (Non-target trials).

Method 1	Method 2	p-val (AD)	p-val (LD)	p-val (RMSE)
Raw	ICA	*0.019	**0.000	**0.000
Raw	Conv	0.102	1.000	0.349
Raw	Pro 1	**0.000	*0.047	**0.000
Raw	Pro 2	0.227	0.594	0.864
ICA	Conv	0.975	**0.000	**0.000
ICA	Pro 1	0.105	**0.000	0.122
ICA	Pro 2	0.870	**0.000	**0.000
Conv	Pro 1	*0.020	*0.028	*0.031
Conv	Pro 2	0.996	0.479	0.911
Pro 1	Pro 2	**0.006	0.691	**0.002

In addition to this qualitative analysis, we also objectively measure the effectiveness of the proposed method quantitatively by measuring AD, LD, and RMSE, respectively. Figure 4 shows the averages and standard deviations of these values over all trials for each type of stimulus. We performed ANOVA to analyze difference among compared methods and we found significant difference between methods in either of AD, LD and RMSE of target and non-target trials ($p < 10^{-13}$ for AD, $p < 10^{-4}$ for LD, and $p < 0.01$ for RMSE of target trials, and $p < 10^{-5}$ for AD, $p < 10^{-21}$ for LD and $p < 10^{-12}$ for RMSE of non-target trials). We also performed multiple comparison test with Tukey's honestly significant difference (HSD) criterion [19] and the result is shown in Table 2 for target trials and Table 3 for non-target trials where Pro stands for the proposed method and Conv stands for the conventional method and the asterisks * and ** indicate statically significant difference at the 5 and 1 percent level, respectively. With regards to target trials, we can see that regularized ICA resulted in significantly smaller AD than raw signals and all the other denoising methods. However, it had also significantly larger LD than others. ICA constantly outputted a denoised signal with small amplitude regardless of how large amplitude the raw signal had. Both of the proposed methods made AD significantly smaller than the raw signal and the conventional method in target trials. The proposed method 2 made RMSE significantly smaller than raw signals and regularized ICA. With regards to non-target trials, the proposed method 1 had significantly smaller AD and RMSE compared to raw signals and the conven-

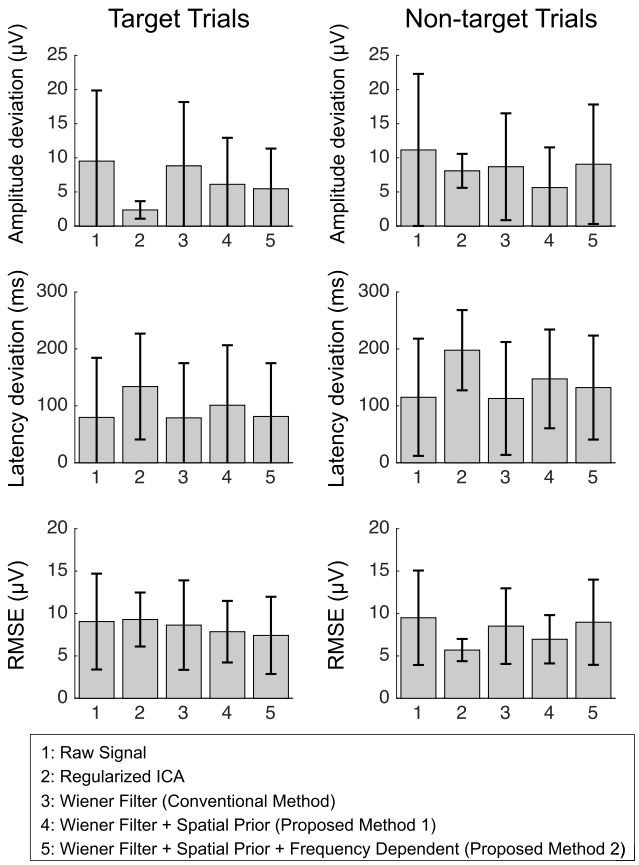

Fig. 5 Amplitude deviation of target trials (upper left), latency deviation of target trials (middle right), root mean square errors of target trials (lower right), amplitude deviation of target trials (upper right), latency deviation of target trials (middle right), and root mean square errors of non-target trials (lower right). The bar length are the means and the lines are standard deviations of 40 trials for each type of stimuli.

tional method. The result of AD and RMSE was consistent in each of type of stimulus.

6. Discussion

6.1 Effectiveness of Using Prior Information

All the compared denoising methods are statistical methods that require sufficiently sized data. However, we face a shortage of data in the context of estimation of single trial ERPs because trial length is usually only about 1 second or less. In fact, we tried to compare one more time-frequency modeling method that does not use prior information with frequency dependent spatial correlation matrices. However, the number of time frames of STFT was only eight, which was too small to estimate 22×22 matrices. As a result, the EM algorithm never finished properly, forcing us to discard this method. Utilizing prior information effectively compensated for the shortage of data size and led to the better denoising performance of the proposed methods, as well as allows us to avoid the permutation ambiguity of event signals because the target signal corresponds to the mixture component with the target prior distribution.

6.2 Frequency Dependency

We could not find the proposed method 2 (frequency dependent covariance model) always significantly better than the proposed method 1 (frequency independent covariance model) although the former could have been able to more effectively represent a frequency dependent spatial spread of a multi-channel EEG signal as stated in Sect. 3.5. In the frequency dependent covariance model, the number of samples $\mathbf{x}(n, f)$ to estimate a covariance matrix is reduced to $1/F$ compared to the frequency independent covariance model, as the samples that belong to only one frequency bin are involved in estimating the corresponding covariance matrix as shown in Eq. (34) while all samples $\mathbf{x}(n, f)$ in the time-frequency grid are involved in the frequency independent model as shown in Eq. (33). This has an adverse effect on robust covariance matrix estimation and could deteriorate signal separation performance placing a trade off between frequency specific covariance matrix modeling and robust covariance matrix estimation.

6.3 Target and Non-Target Trials

Denoising performance of all compared methods on non-target trials were mostly worse than target trials. As shown in Fig. 3, the amplitude of P300 that was embedded into resting state EEGs was smaller in non-target trials than target trials as is often the case with an oddball paradigm experiment [18], which made SNR worse in non-target trials and could make this noise removal problem harder than in target-trials.

7. Conclusion

We developed a new noise removal method from ERP data extending a conventional method that models generation of each event signal in the time frequency domain using prior information of ERP. Both of the proposed methods showed effectiveness in removing background EEG signals from single-trial EEG signals to quantify ERP data more precisely and lead us to better understanding of event-related brain dynamics.

There are a number of avenues for future work. First, in this study we used EEG data to obtain prior information of each event signal recorded in the same session and the same subject. Ideally, we would like to use prior information collected from other days or subjects, which is a future challenge. Second, in this paper we used a carefully controlled experimental paradigm to avoid severe noise such as body movements. Hence, we hope to further validate the effectiveness of our proposed methods to remove such severe noise. Moreover, we used prior information about spatial spread over the scalp by setting a prior distribution of spatial correlation matrix $\mathbf{R}_k$. However, we did not use any prior information about a timing of event signals. We can set a prior

distribution of the scaling parameter $v_k(n, f)$ to exploit a timing prior, which may be effective to do better estimation of single-trial ERP because ERP components have characteristic timing of their amplitude shifts.

Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 26540117.

References

[1] M. Kutas, G. McCarthy, and E. Donchin, "Augmenting mental chronometry: The P300 as a measure of stimulus evaluation time," Science, vol.197, no.4305, pp.792–795, 1977.
[2] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," IEEE Signal Process. Mag., vol.25, no.1, pp.41–56, 2008.
[3] T.-P. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, and T.J. Sejnowski, "Analysis and visualization of single-trial event-related potentials," Human Brain Mapping, vol.14, no.3, pp.166–185, 2001.
[4] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.R. Müller, "Single-trial analysis and classification of ERP components — A tutorial," NeuroImage, vol.56, no.2, pp.814–825, 2011.
[5] J.A. Urigüen and B. Garcia-Zapirain, "EEG artifact removal—State-of-the-art and guidelines," Journal of Neural Engineering, vol.12, no.3, 031001, 2015
[6] D.M. Groppe, S. Makeig, and M. Kutas, "Independent component analysis of event-related potentials," Cognitive Science Online, 2008.
[7] S. Lemm, G. Curio, Y. Hlushchuk, and K.R. Müller, "Enhancing the signal-to-noise ratio of ICA-based extracted ERPs," IEEE Trans. Biomed. Eng., vol.53, no.4, pp.601–607, 2006.
[8] C. Févotte and J.F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp.78–81, 2005.
[9] N.Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," IEEE Trans. Audio Speech Language Process., vol.18, no.7, pp.1830–1840, 2010.
[10] Y. Kurihana, S. Miyabe, T.M. Rutkowski, Y. Matsumoto, T. Yamada, and S. Makino, " Signal separation of EEG using multivariate probabilistic model," IEICE Technical Report, MBE2012-119, 2013 (in Japanese).
[11] H. Maki, T. Toda, S. Sakti, G. Neubig, and S. Nakamura, "EEG signal enhancement using multi-channel Wiener filter with a spatial correlation prior," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp.2639–2643, 2015.
[12] M. Zibulevsky and B.A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," Neural Comput., vol.13, no.4, pp.863–882, 2001.
[13] Y. Okada, J. Jung, and T. Kobayashi, "An automatic identification and removal method for eye-blink artifacts in event-related magnetoencephalographic measurements," Physiological Measurement, vol.28, no.12, pp.1523–1532, 2007.
[14] F.C. Viola, J. Thorne, B. Edmonds, T. Schneider, T. Eichele, and S. Debener, "Semi-automatic identification of independent components representing EEG artifact," J. Clinical Neurophysiology, vol.120, no.5, pp.868–877, 2009.
[15] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features," Psychophysiology, vol.48, no.2, pp.229–240, 2011.
[16] C.M. Bishop, Pattern recognition and machine learning, Springer New York, 2006.
[17] American Clinical Neurophysiology Society, "Guideline Thirteen: Guidelines for standard electrode position nomenclature," J. Clinical Neurophysiology, vol.11, no.1, pp.111–113, 1994.
[18] S.J. Luck, An introduction to the event-related potential technique, MIT Press, 2014.
[19] J.W. Tukey, "Comparing individual means in the analysis of variance," Biometrics, vol.5, no.2, pp.99–114, 1949.

Hayato Maki received his B.S. from University of Tsukuba, Ibaraki, Japan in 2009, and recieved his M.E. from Nara Institute of Science and Technology (NAIST), Nara, Japan in 2015. He is currently a Ph.D. student at NAIST. He is a graduate school member of IEEE SPS and BME. His research interests include cognitive science and signal processing with a focus on biological signal processing.

Tomoki Toda earned his B.E. degree from Nagoya University in 1999 and his M.E. and D.E. degrees from Nara Institute of Science and Technology (NAIST) in 2001 and 2003, respectively. He was a Research Fellow of JSPS from 2003 to 2005. He was an Assistant Professor at NAIST from 2005 to 2011, and an Associate Professor from 2011 to 2015. He is currently a Professor at Nagoya University. His research interests include statistical approaches to speech processing. He received more than 10 paper awards including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (Speech Communication Journal).

Sakriani Sakti received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999, and her M.Sc. degree from University of Ulm, Germany in 2002. While working in Japan at ATR (2003–2009) and NICT (2006–2011), she continued her study (2005–2008) with Dialog Systems Group University of Ulm, Germany, and received her Ph.D. degree in 2008. Currently, she is an assistant professor of Nara Institute of Science and Technology, Japan. Her research interests include statistical pattern recognition, speech recognition, spoken language translation, cognitive communication, and graphical modeling framework.

Graham Neubig received his B.E. from University of Illinois, Urbana-Champaign, U.S.A, in 2005, and his M.E. and Ph.D. in informatics from Kyoto University, Kyoto, Japan in 2010 and 2012 respectively. He is currently an assistant professor at the Nara Institute of Science an Technology, Nara, Japan. His research interests include speech and natural language processing, with a focus on machine learning approaches for applications such as machine translation, speech recognition, and spoken dialog.

Satoshi Nakamura is Professor of Graduate School of Information Science, Nara Institute of Science and Technology, Japan, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007-2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009-2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampoli Award 2012. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011, IEEE Signal Processing Magazine Editorial Board Member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016.