Journal of Signal Processing, Vol.21, No.2, pp.63-71, March 2017

PAPER

Automatic Parameter Switching of Noise Reduction for Speech Recognition

Tomoko Kawase[1], Kenta Niwa[1], Yusuke Hioka[2] and Kazunori Kobayashi[1]

[1] NTT Media Intelligence Laboratories, NTT Corporation, 3-9-11 Midori-cho, Musashino, Tokyo 180-8585, Japan
[2] Department of Mechanical Engineering, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
E-mail: {kawase.tomoko,niwa.kenta,kobayashi.kazunori}@lab.ntt.co.jp[1], yusuke.hioka@ieee.org[2]

Abstract **Noise reduction has been extensively studied to prevent the accuracy of automatic speech recognition (ASR) being severely degraded due to various noise sources observed in practical use. The effect of noise reduction is usually dependent on the parameters used in the noise-reduction method, which need to be tuned for different types of noise to achieve the best accuracy. We propose a method for automatically switching noise-reduction parameters in such a way that ASR accuracy is maximized. The proposed method assigns the most suitable parameter set to each speech sentence by measuring and grouping the characteristics of noise observed in each sentence. Experiments using speech sentences contaminated by various types of noise showed that the proposed method can reduce the word error rate of ASR.**

Keywords: noise reduction, Wiener filter, microphone array, parameter switching, robust automatic speech recognition

1. Introduction

Automatic speech recognition (ASR) systems have been commonly used for interfacing between users and machines in various applications, e.g., smartphones, car navigation systems, and robotic automation systems. Despite their convenience, their accuracy is affected by the acoustic environments in which such applications are used, which are often very noisy. The characteristics of noise, such as power spectral distribution, sound pressure level, and stationarity, also vary depending on the chosen environment. For example, in running vehicles, speech will be contaminated by noise from the road, engine, wind, air conditioners and/or music from the audio system. Thus, extensive studies have been conducted to make ASR systems more robust against such varied noisy environments.

A common method of developing noise-robust ASR systems is making the classifier become robust such as by training an acoustic model using Gaussian mixture model (GMM)-based hidden Markov models (HMMs) (GMM-HMMs) [1–3]. It was also recently reported that the use of deep neural networks (DNNs) improves ASR accuracy compared with a GMM-HMM [4, 5]. With these methods, the differences in the input acoustic features for each phoneme still need to be maintained, even when speech signals are recorded in noisy environments. However, it becomes very difficult to identify the target speech from input acoustic features when the signal-to-noise ratio (SNR) of the observed speech is very low [6]. Therefore, ASR accuracy in very noisy environments cannot be sustained by simply revising the methods for acoustic model training.

To combat degradation in ASR accuracy, noise reduction is usually applied to the noisy speech signals as a pre-processing of ASR systems. Of those noise-reduction methods [7], those using microphone arrays can be seen more frequently in recent practical applications, and are able to reduce noise using the spatial characteristics of sound sources. Beamforming [8, 9] combined with post-filtering [10–14] is a framework that is known to be practically effective [15]; however, the power spectral densities (PSDs) of a target source and that of surrounding noise need to be estimated.

We previously investigated methods for estimating the PSDs of sound sources separately by looking into the temporal [16] and spatial [17,18] characteristics of each sound source. With these methods, several parameters that can only be determined empirically are needed to calculate the PSDs and coefficients of the post-filter. Since previous studies discovered that noise-reduction performance using estimated PSDs is highly dependent on the parameters selected, an additional method for automatically selecting the best parameters needed to be investigated [19, 20].

In this study, we propose a method for automatically *switching* noise-reduction parameters, providing the highest ASR accuracy. The method introduces the noise-power vector, which quantifies the characteristics of noise contaminating each speech sentence by measuring the spectral power of different frequency bands. The noise-power vector is then used to group speech sentences to assign the best noise-reduction parameter set. ASR systems will achieve

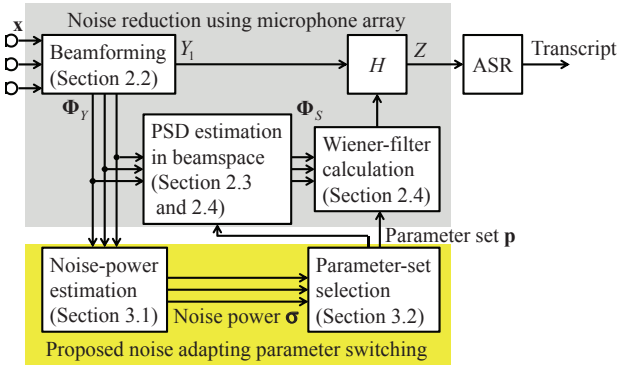


Fig. 1 Microphone-array noise reduction with proposed parameter-switching method for ASR

the lowest word error rate (WER), following noise reduction with the best parameter set. The WER is modelled by a function of the centroids of groups and parameter sets. The lowest WER value is searched for by adjusting the position of the centroids and parameter sets using the hill-climbing algorithm.

The rest of this paper is organized as follows. We give an overview of a previously developed noise-reduction method using a microphone array in Section 2. In Section 3, we explain the details of the proposed selection method of a noise-reduction parameter set. We present the experimental results obtained using speech sentences in various noisy environments along with discussion in Section 4 and conclude the paper with some remarks in Section 5.

2. Noise Reduction Using Microphone Array

Fig. 1 is an overview of the framework considered in the paper. In this section, we review a well-known noise-reduction framework using microphone arrays, also known as beamforming with a Wiener post-filter, followed by a brief summary of the PSD-estimation method for designing the Wiener post-filter.

2.1 Problem setup

Assuming that an M-sensor microphone array is surrounded by K spatially *coherent* sound sources $S_k(\omega,\tau)$ and spatially *incoherent* background noise $V_m(\omega,\tau)$, let $X_m(\omega,\tau)$ be the signal observed with the m-th microphone of the array given by

$$X_m(\omega,\tau) = \sum_{k=1}^{K} A_{m,k}(\omega)\, S_k(\omega,\tau) + V_m(\omega,\tau) \qquad (1)$$

Here, $A_{m,k}(\omega)$ denotes the transfer function between the m-th microphone and k-th sound source; ω and τ are the indices of frequency bins and frames, respectively. It is also assumed that each sound source is uncorrelated with other sources, i.e.,

$$E\left[S_k(\omega,\tau)\, S^*_{k'}(\omega,\tau)\right]_\tau = 0 \quad \forall k, k \neq k'$$
$$E\left[V_m(\omega,\tau)\, V^*_{m'}(\omega,\tau)\right]_\tau = 0 \quad \forall m, m \neq m'$$
$$E\left[S_k(\omega,\tau)\, V^*_m(\omega,\tau)\right]_\tau = 0 \quad \forall k, \forall m$$

where $E[\cdot]_\tau$ denotes the average across frames. In this study, we aimed at extracting the signal of target source $S_1(\omega,\tau)$ provided its arrival direction θ_1 is known.

2.2 Beamforming with Wiener post-filter

Beamforming with Wiener post-filtering [11] is a well-known approach to noise reduction using microphone arrays. Beamforming is first applied to microphone observation, the output signal of which is given by

$$Y_l(\omega,\tau) = \mathbf{w}_l^{\mathrm{H}}(\omega)\, \mathbf{x}(\omega,\tau) \qquad (2)$$

where

$$\mathbf{w}_l(\omega) = \left[W_{l,1}(\omega), \ldots, W_{l,M}(\omega)\right]^{\mathrm{T}} \qquad (3)$$
$$\mathbf{x}(\omega,\tau) = \left[X_1(\omega,\tau), \ldots, X_M(\omega,\tau)\right]^{\mathrm{T}} \qquad (4)$$

and $^{\mathrm{H}}$ and $^{\mathrm{T}}$ denote the Hermitian conjugation and transposition, respectively. The weights of the beamforming $\mathbf{w}_l(\omega)$ are designed to emphasize sound arriving from θ_l with, e.g., the minimum variance distortionless response (MVDR) method [9].

To boost noise-reduction performance, the Wiener filter $H(\omega,\tau)$ is then applied to the output of the beamforming given by

$$Z(\omega,\tau) = H(\omega,\tau)\, Y_1(\omega,\tau) \qquad (5)$$

The Wiener filter gain $H(\omega,\tau)$ is calculated using the PSD of target source $\phi_S(\omega,\tau)$ and that of other interfering sources and background noise $\phi_N(\omega,\tau)$ (referred to as noise hereafter). Details on estimating $\phi_S(\omega,\tau)$ and $\phi_N(\omega,\tau)$ and calculating $H(\omega,\tau)$ are given in the rest of this section.

Finally, the output signal in the time-domain is obtained by applying the inverse Fourier transform to $Z(\omega,\tau)$.

2.3 PSD estimation in beamspace

To calculate the Wiener filter gain, the PSD of the target sound source and noise needs to be estimated from microphone array observation. As modeled in (1), the noise consists of $K-1$ spatially *coherent* interfering sources and spatially *incoherent* background noise. Due to the differences in their spatial propagation properties, the noise-reduction method [16] estimates the PSD of these noise components separately.

PSD estimation in beamspace [17] is first applied to separate the coherent interfering sources. Let L ($\geq K$)

beamformers, which focus their directivity on different angles, be applied for microphone array observation. Given that each source signal is assumed to be uncorrelated with other source signals and background noise, the PSD of the l-th beamforming output $\phi_{Y_l}(\omega)$ can be approximated by an affine transformation of the PSDs of each source $\phi_{S_k}(\omega)$ with the directivity gain of the l-th beamformer to the k-th source direction $|D_{l,k}(\omega)|^2$ and noise PSD, as shown in

$$\underbrace{\begin{bmatrix}\phi_{Y_1}\\ \vdots \\ \phi_{Y_L}\end{bmatrix}}_{\mathbf{\Phi}_Y(\omega)} \approx \underbrace{\begin{bmatrix}|D_{1,1}|^2 & \dots & |D_{1,K}|^2\\ \vdots & \ddots & \vdots \\ |D_{L,1}|^2 & \dots & |D_{L,K}|^2\end{bmatrix}}_{\mathbf{D}(\omega)} \underbrace{\begin{bmatrix}\phi_{S_1}\\ \vdots \\ \phi_{S_K}\end{bmatrix}}_{\mathbf{\Phi}_S(\omega)} + \underbrace{\begin{bmatrix}\phi_{V_1}\\ \vdots \\ \phi_{V_L}\end{bmatrix}}_{\mathbf{\Phi}_V(\omega)} \tag{6}$$

where the indices of ω and τ are omitted.

If the transfer function $A_{m,k}(\omega,\tau)$ can be modeled by any means such as using array manifold vectors [8,13], the directivity gain is given by

$$D_{l,k}(\omega) = \sum_{m=1}^{M} W_{l,m}^{*} A_{m,k}(\omega) \tag{7}$$

where * denotes the complex conjugate.

Thus, the PSD of each coherent sound source can be separated by

$$\begin{aligned}\hat{\mathbf{\Phi}}_{S+N}(\omega,\tau) &= \left[\phi_{S_1+N}(\omega,\tau),\dots,\phi_{S_K+N}(\omega,\tau)\right]^{\mathrm{T}}\\ &:= \mathbf{G}(\omega)\,\mathbf{\Phi}_Y(\omega,\tau)\\ &\approx \mathbf{\Phi}_S(\omega,\tau) + \mathbf{G}(\omega)\,\mathbf{\Phi}_V(\omega)\end{aligned} \tag{8}$$

where

$$\mathbf{G}(\omega) = \begin{cases}\mathbf{D}^{-1}(\omega) & L = K\\ \mathbf{D}^{+}(\omega) & L > K\end{cases} \tag{9}$$

and $^+$ denotes pseudo inverse.

2.4 PSD estimation of incoherent background noise and Wiener-filter calculation

As can be seen in the second term of (8), components originating from the spatially incoherent background noise are still included in the estimated PSD $\hat{\mathbf{\Phi}}_{S+N}(\omega,\tau)$. These background noise components should be removed to estimate the PSD of coherent sound sources. To this end, the noise-reduction method uses the temporally stationary property of the background noise. Provided all spatially coherent sources including the target source are nonstationary, the PSD of background noise can be estimated by measuring the level of stationary components. The stationary components in $\phi_{S_k+N}(\omega,\tau)$ can be estimated by taking the minimum value of a temporarily smoothed PSD in a given time interval Υ, that is,

$$\phi_{N_k}(\omega,\tau) = \min_{\tau\in\Upsilon}\left\{\bar{\phi}_{S_k+N}(\omega,\tau)\right\} \tag{10}$$

Here, $\bar{\phi}_{S_k+N}$ is calculated by applying a recursive update algorithm, that is,

$$\begin{aligned}\bar{\phi}_{S_k+N}(\omega,\tau) =& \alpha\phi_{S_k+N}(\omega,\tau)\\ &+ (1-\alpha)\,\bar{\phi}_{S_k+N}(\omega,\tau-1)\end{aligned} \tag{11}$$

where α denotes the forgetting coefficient parameter, which is set so that its time constant is around 150 ms.

Given that the directivity of the first beamformer ($l = 1$) points to the angle of the target source, the PSD of target source $\phi_S(\omega,\tau)$ is calculated by

$$\phi_S(\omega,\tau) = \phi_{S_1+N}(\omega,\tau) - \phi_{N_1}(\omega,\tau) \tag{12}$$

Likewise, the PSD of noise can be calculated using the PSD of other coherent sources and background noise, that is,

$$\begin{aligned}\phi_N(\omega,\tau) =& \underbrace{p_1(\omega)\left\{\sum_{k=2}^{K}\left(\phi_{S_k+N}(\omega,\tau) - \phi_{N_k}(\omega,\tau)\right)\right\}}_{\text{PSD of interference sources}}\\ &+ \underbrace{\phi_{N_1}(\omega,\tau)}_{\text{PSD of background noise}}\end{aligned} \tag{13}$$

where $p_1(\omega)$ is a weighting parameter.

Finally, the Wiener filter gain is calculated as follows. On the basis of the minimum mean square error criterion, the Wiener filter gain is basically calculated by

$$H_{\text{base}}(\omega,\tau) = \frac{\phi_S(\omega,\tau)}{\phi_S(\omega,\tau) + \phi_N(\omega,\tau)} \tag{14}$$

Then, $H_{\text{base}}(\omega,\tau)$ is shaped to reduce the musical noise in the output signal of the Wiener filter. It is smoothed in the time domain, that is,

$$\begin{aligned}H_{\text{smooth}}(\omega,\tau) =& p_2(\omega)\,H_{\text{base}}(\omega,\tau)\\ &+ (1-p_2(\omega))\,H_{\text{smooth}}(\omega,\tau-1)\end{aligned} \tag{15}$$

and floored in the time-frequency domain, that is,

$$H = \begin{cases}1 & (H_{\text{smooth}}(\omega,\tau) \geq 1)\\ H_{\text{smooth}}(\omega,\tau) & (p_3(\omega) \leq H_{\text{smooth}}(\omega,\tau) < 1)\\ p_3(\omega) & (H_{\text{smooth}}(\omega,\tau) < p_3(\omega))\end{cases} \tag{16}$$

where $p_2(\omega)$ and $p_3(\omega)$ are a forgetting coefficient and flooring parameter, respectively.

Since noise-reduction performance is highly dependent on the values given to parameters, the parameters $\mathbf{p}_j = [p_1(\omega), p_2(\omega), p_3(\omega)]^{\mathrm{T}}$ $(j = 1,\dots,J)$ need to be manually pre-adjusted for each noise environment, and one of them needs to be selected to suit the given environment, where j denotes each noise environment. This process is heuristic since the relationship between the quality of *denoised*, i.e., post noise reduction, speech and the parameters is not explained deterministically. The following section introduces our parameter-switching method that adapts to the noise environment, which eventually contributes to noise-robust ASR providing a lower WER.

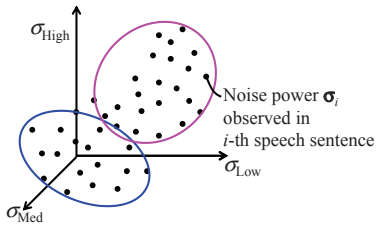


Fig. 2 Group of noise-power vectors

3. Noise Adapting Parameter Switching

3.1 Noise characteristic measurement

Because the characteristics of a noisy environment vary depending on the scene where the ASR is used, one can hypothesize that the WER of ASR can be reduced by switching the values set for the parameters of noise reduction depending on the noisy environment. The proposed parameter-switching method selects a parameter set from J pre-adjusted sets for each frequency $\mathbf{p}_j(\omega)$ $(j = 1, \ldots, J)$ that minimizes the WER. Given that a dataset consisting of I speech sentences with various types of noise being superimposed and their correct word labels are provided, assume the noise-reduction parameter sets $\mathbf{p}_j(\omega)$ are manually pre-adjusted using parts of the dataset. The ASR is applied to the denoised speech using every $\mathbf{p}_j(\omega)$ $(j = 1, \ldots, J)$, and the parameter set that minimizes the WER is then selected as the optimal parameter set.

To quantify the characteristics of a noisy environment, a noise-power vector, $\boldsymbol{\sigma}_i$, composed of noise power in a different frequency band is introduced as

$$\boldsymbol{\sigma}_i = \left[\sigma_{\text{Low},i}, \sigma_{\text{Med},i}, \sigma_{\text{High},i}\right]^{\mathrm{T}} \tag{17}$$

where $\sigma_{\text{Low},i}$, $\sigma_{\text{Med},i}$, and $\sigma_{\text{High},i}$ denote the power of noise observed in the i-th speech sentence in low (0.0–0.1 kHz), medium (0.1–3.3 kHz), and high (3.3–8.0 kHz) frequency bands, respectively. The noise-power vectors are given by the average of the PSDs in the relevant frequency band estimated by applying a previous method [21] to the first element of the PSD of the background noise, i.e., $\frac{1}{\mathcal{T}\omega_{\text{band}}}\sum_{\tau}\sum_{\omega_{\text{band}}}\phi_{N_1}(\omega,\tau)$, where $\mathcal{T}$ is the number of frames to be averaged across and ω_{band} is each of the low, medium, and high frequency bands, respectively. Measuring the noise power in different frequency bands allows the use of the spectral distribution of the noise in addition to its loudness.

3.2 Parameter selection by grouping noise-power vectors

Fig. 2 shows a scatter plot of the noise-power vector calculated from I noisy speech sentences, i.e., $\boldsymbol{\sigma}_i$ $(i = 1, \ldots, I)$. Noise-power vectors are grouped into U $(I \gg U \geq 2)$ groups; then, different noise-reduction parameter sets are applied to the noisy speech sentences that belong to each group in such a way that the overall WER averaged

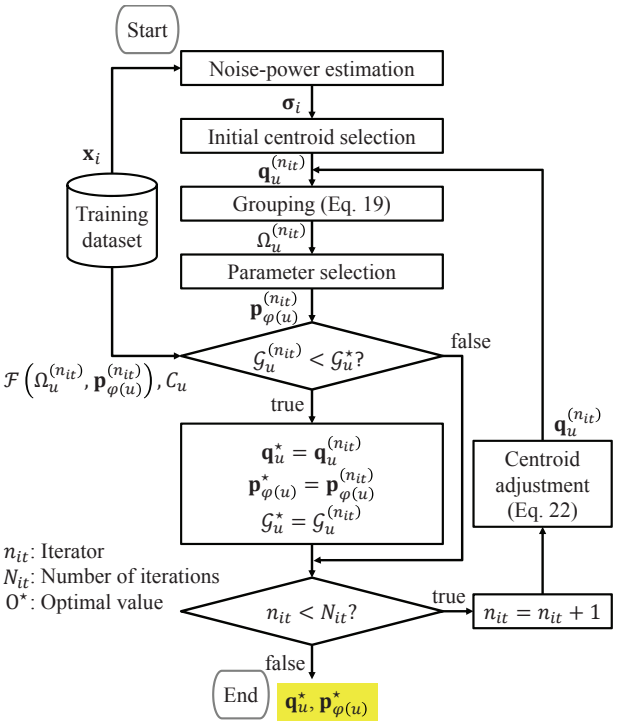


Fig. 3 Flow chart of proposed method

across the whole dataset is minimized. In other words, the i-th dataset is grouped into one of the U groups u that is assigned the best parameter set that maximizes ASR accuracy according to its noise-power vector.

The grouping is made possible by using centroids defined in the space of the noise-power vector given by

$$\mathbf{q}_u = \left[q_{\text{Low},u}, q_{\text{Med},u}, q_{\text{High},u}\right]^{\mathrm{T}} (u = 1, \ldots, U) \tag{18}$$

where $q_{\text{Low},u}$, $q_{\text{Med},u}$, and $q_{\text{High},u}$ are the coordinates of the u-th centroid. Initially, the centroids are set at the same positions as that of randomly selected U noise-power vectors among all noise-power vectors measured from the I speech sentences. Then, all noise-power vectors are grouped to their closest centroid by

$$\Omega_u \ni \forall i \quad \text{s.t.} \quad u = \arg\min_{u'} \|\boldsymbol{\sigma}_i - \mathbf{q}_{u'}\|^2 \tag{19}$$

where Ω_u denotes the set of indices of the speech sentences grouped into the u-th group.

3.3 Optimal grouping for maximizing speech recognition accuracy

The grouping now needs to be optimized in order to minimize the WER. On the basis of the hypothesis that the parameter sets affect ASR accuracy, the WER of speech sentences in the Ω_u group will be represented by a function of Ω_u and $\mathbf{p}_j(\omega)$, i.e., $\mathcal{F}\left(\Omega_u, \mathbf{p}_j(\omega)\right)$. This implies that both Ω_u and $\mathbf{p}_j(\omega)$ have to be adjusted to minimize the WER. The proposed method achieves this by tuning both

Ω_u and $\mathbf{p}_j(\omega)$ alternately as follows. Fig. 3 shows a flow chart of the optimization.

Once Ω_u is determined by the initial grouping given by (19), the best parameter set is selected for each group by

$$\varphi(u) = \arg\min_j \left\{\mathcal{F}\left(\Omega_u, \mathbf{p}_j(\omega)\right)\right\} \tag{20}$$

where $\varphi(u)$ is the index of the chosen noise-reduction parameter set for the u-th group.

Then, the grouping is further optimized by adjusting the positions of the centroids provided the parameter sets φ_u are fixed. To this end, a cost function that quantifies the overall WER is introduced, that is,

$$\mathcal{G}_u = \frac{\sum_{u=1}^{U} \mathcal{F}\left(\Omega_u, \mathbf{p}_{\varphi(u)}(\omega)\right) C_u}{\sum_{u=1}^{U} C_u} \tag{21}$$

where C_u denotes the number of words in the speech sentences that belong to the u-th group. Since the relationships between $\mathcal{F}\left(\Omega_u, \mathbf{p}_j(\omega)\right)$ and $\mathbf{q}_u$ cannot be described with a simple model, it is difficult to analytically derive $\mathbf{q}_u$ that minimizes $\mathcal{G}_u$. Instead, the hill-climbing algorithm [22] is applied to search the optimal $\mathbf{q}_u$. This method finds an optimal value by updating parameters while evaluating the change in the cost function. In this study, $\mathcal{G}_u$ in (21) was evaluated while $\mathbf{q}_u$ was perturbed by $\boldsymbol{\epsilon}$, given by

$$\mathbf{q}_u \leftarrow \mathbf{q}_u \pm \boldsymbol{\epsilon}, \tag{22}$$

$$\boldsymbol{\epsilon} = \left[\epsilon_{\mathrm{Low},u}, \epsilon_{\mathrm{Med},u}, \epsilon_{\mathrm{High},u}\right]^{\mathrm{T}} \tag{23}$$

After updating $\mathbf{q}_u$, the parameter sets assigned to groups will be reviewed and updated by (20). This two-step process will carry on until no further reduction in $\mathcal{G}_u$ is observed by adjusting $\mathbf{q}_u$. Because $\mathcal{G}_u$ may not be minimized when the initial value of $\mathbf{q}_u$ is not arranged properly, the optimal search needs to be run multiple times with randomly selected initial centroids; then, $\mathbf{p}_{\varphi(u)}(\omega)$ and $\mathbf{q}_u$ that provide the lowest $\mathcal{G}_u$ after reaching its minimum value need to be selected.

4. Experiments

We compared ASR accuracy obtained using the proposed method with that using the conventional method, in which a single noise-reduction parameter set was applied.

4.1 Experimental conditions

The microphone array used in the experiment consisted of three cardioid microphones. Each microphone was oriented 120° from the others. Beamformings were designed by using the MVDR method to point their directivity towards the three directions to which the microphones were facing.

Training and evaluation data were manually prepared as follows. We measured the impulse responses from the positions of the target and noise sources to the microphone array as shown in Fig. 4. Impulse responses from eight different directions were measured to simulate incoherent background noise. The measurements were carried out in a reverberant chamber of which two different amounts of sound-absorbing material was put up on walls and the ceiling and in a meeting room to confirm that the proposed method is effective in various reverberant environments. Clean speech signals were convolved with the recorded impulse responses of the target to make the target sound signals. In the same way, one of four different types of background noise recorded in cars, offices, shopping centers, or exhibition halls was convolved with the impulse responses of the noise from the eight directions, which were summed together. Finally, the target sound and background noise were added up with different SNRs, which were varied from −10 to 10 dB.

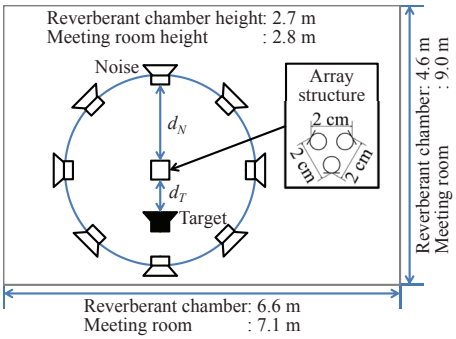


Fig. 4 Noise and impulse response measurement setup to create evaluation data simulating microphone array observation

The Complete Continuous Speech Recognition-I (CSR-I) corpus [23] was used for the clean speech signals. The corpus was divided into two subsets, that is, training and evaluation datasets, which were composed of 323 and 328 English speech sentences spoken by four individuals, respectively. An acoustic model and language model were trained using the Kaldi [24] baseline tool for the CHiME Challenge [25]. The acoustic model was constructed with a GMM-HMM. For the proposed method, 28 sets of parameters were manually prepared.

It is expected that the number of the groups U would also affect the ASR accuracy apart from the position of centroids and selected parameter sets. Since the relationship between U and ASR accuracy is not easily modeled, we experimentally investigated the relationship by varying U from 2 to 6.

The other experimental conditions are listed in Table 1.

4.2 Experimental results

In the experiments, the centroids were derived by using the training process according to the flow chart in Fig. 3. The centroids derived when $U = 4$ are listed in Table 2, where the noise power is described in dBov, which denotes the level relative to the maximum value that can be stored

Table 1 Experimental conditions

Sampling rate (kHz)	16
Quantization bit rate (bit)	16
# of microphones, M	3
# of beamformers, L	3
Target distance, d_T (m)	0.5, 1.0, 1.5, 2.0
Noise distance, d_N (m)	2.0, 4.0
Frame length (ms)	32
Frame shift (ms)	16
# of sentences for training, I	323
# of sentences for evaluation	328
# of groups, U	2, 3, 4, 5, 6
# of parameter sets, J	28
Perturbation amplitude, $\boldsymbol{\epsilon}$	$[1, 1, 1]^{\mathrm{T}}$

Table 2 Centroids obtained by training when $U = 4$

	σ_{Low} (dBov)	σ_{Med} (dBov)	σ_{High} (dBov)
Ω_1	−66	−61	−72
Ω_2	−56	−64	−98
Ω_3	−55	−42	−42
Ω_4	−27	−30	−61

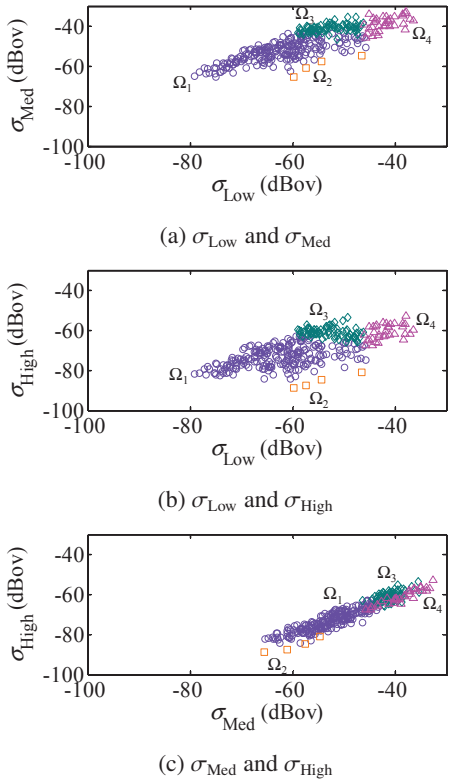
in an integer format on a computer [26]. The centroids specified the groups of noise power observed in the speech sentences. Fig. 5 shows the noise-power vectors, which were divided into four groups using the centroids, in different colors/markers.

A parameter set was assigned for each of the groups. The component values of the parameter sets for different U are listed in Table 3. Note that we averaged the values with respect to the frequency bins for simplicity. The experimental results obtained when $U = 1$ are equivalent to those with the conventional method. The same parameter set used for $U = 1$ was also used to evaluate the conventional method when $U \geq 2$. The ASR results obtained using the assigned parameter sets are described by WER in Table 4, in which WER values reduced by using the proposed method are highlighted in boldface type.

The results clearly show that the WER was reduced in most groups with the proposed method than with the conventional method ($U = 1$).

Regardless of the value of U, there was a majority group, or a group that included a relatively large number of speech sentences. Although the WER was not reduced in the majority groups, it was reduced in other groups, or minority groups. With the conventional method, in which a single parameter set was assigned to all the speech sentences, the parameter set suitable for the majority group was selected as the optimal one. However, such a parameter set was not suitable for the minority groups. With the proposed method, another parameter set, which was especially suitable for each of the minority groups was selected to reduce the WER.

Referring to the experiment with $U = 6$, one of the groups included no speech sentences. The centroid was located too far from the whole dataset to form its group in



(a) σ_{Low} and σ_{Med}

(b) σ_{Low} and σ_{High}

(c) σ_{Med} and σ_{High}

Fig. 5 Grouping results when $U = 4$

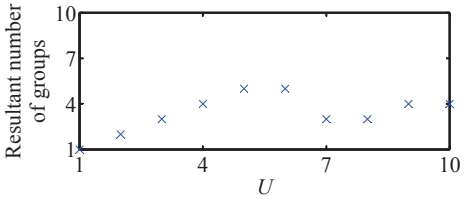


Fig. 6 Relationship between given U and resultant number of groups

the noise-power vector space during the training process. This shows the proposed method allows redundant groups to disappear in order to minimize the WER. As a result, although an ad hoc number of groups must be given at the beginning of the training process, the number of groups tends to be reduced to an appropriate number, not necessarily the optimal number, automatically. Fig. 6 shows the resultant numbers of groups when U was increased. The numbers of groups saturated at five even if U was increased beyond five. In these experiments, five groups were enough for the data to reduce the WER by the parameter switching.

Finally, we aggregated the WER for the whole dataset, as shown in Table 5. Note that the results listed in Tables 4 and 5 were obtained from the same experiments, although Table 5 includes results for the training dataset as well as the evaluation dataset. The WER reduction for the minority groups resulted in eight to ten-point reductions for the whole evaluation dataset. The experiment using the dataset described above showed that the WER with $U \geq 3$ was roughly the same as that with $U = 2$. Because there

Table 3 Frequency-averaged noise-reduction parameter sets

		p_1	$p_2 \times 10^{-3}$	p_3
$U = 1$	$\mathbf{p}_{\varphi(1)}$	0.8	9.3	0.20
$U = 2$	$\mathbf{p}_{\varphi(1)}$	0.8	9.3	0.20
	$\mathbf{p}_{\varphi(2)}$	2.0	9.3	0.20
$U = 3$	$\mathbf{p}_{\varphi(1)}$	0.8	9.3	0.20
	$\mathbf{p}_{\varphi(2)}$	2.0	11.6	0.20
	$\mathbf{p}_{\varphi(3)}$	0.8	9.9	0.20
$U = 4$	$\mathbf{p}_{\varphi(1)}$	0.8	9.3	0.20
	$\mathbf{p}_{\varphi(2)}$	2.0	9.9	0.20
	$\mathbf{p}_{\varphi(3)}$	2.0	9.9	0.16
	$\mathbf{p}_{\varphi(4)}$	0.8	9.9	0.20
$U = 5$	$\mathbf{p}_{\varphi(1)}$	0.8	9.3	0.20
	$\mathbf{p}_{\varphi(2)}$	2.0	11.6	0.20
	$\mathbf{p}_{\varphi(3)}$	0.8	9.3	0.20
	$\mathbf{p}_{\varphi(4)}$	1.0	0.0	1.00
	$\mathbf{p}_{\varphi(5)}$	1.0	0.0	1.00
$U = 6$	$\mathbf{p}_{\varphi(1)}$	0.8	9.3	0.20
	$\mathbf{p}_{\varphi(2)}$	2.0	11.6	0.20
	$\mathbf{p}_{\varphi(3)}$	0.8	9.3	0.20
	$\mathbf{p}_{\varphi(4)}$	1.0	0.0	1.00
	$\mathbf{p}_{\varphi(5)}$	2.0	9.9	0.20
	$\mathbf{p}_{\varphi(6)}$	1.0	0.0	1.00

is no method of deriving the optimal U deterministically, we suggest using the proposed method with varying U and selecting the U resulting in the lowest WER.

Overall, the experimental results verified that the proposed method was effective for improving ASR accuracy in noisy environments by switching the noise reduction parameter sets according to the noise level measured in different frequency bands. It was also found that the proposed method was able to adjust the number of groups automatically.

5. Conclusion

We proposed and evaluated a method for automatically switching parameters for noise reduction to improve the accuracy of ASR systems. Noise characteristics are quantified by the noise-power vector measured from noisy speech sentences, which is used to group the sentences and assign the best parameter set to achieve the highest recognition accuracy. Experiments using datasets in various noisy environments revealed that the WER could be reduced to 31.0% with the proposed method compared to 41.7% with the conventional method.

Pursuing efficient features other than the noise-power vector in order to group noisy speech sentences more accurately would be a further research subject.

Acknowledgments

We would like to express our sincere appreciation to Dr. Fujimoto at the National Institute of Information and Communications Technology, who was earlier at NTT Communication Science Laboratories, Dr. Kamado at NTT DOCOMO, Inc., who was earlier at NTT Media Intelligence Laboratories, and Dr. Masumura and Mr. Asami at NTT Media Intelligence Laboratories for their valuable comments.

Table 4 ASR results with WER $\mathcal{F}\left(\Omega_u, \mathbf{p}_{\varphi(u)}\right)$

		# of speech sentences	Conventional method	Proposed method
$U = 1$	Ω_1	328	41.7%	41.7%
$U = 2$	Ω_1	219	26.4%	26.4%
	Ω_2	109	73.5%	**41.7**%
$U = 3$	Ω_1	186	25.9%	25.9%
	Ω_2	135	65.9%	**39.4**%
	Ω_3	7	11.6%	12.3%
$U = 4$	Ω_1	211	25.9%	25.9%
	Ω_2	79	72.0%	**38.8**%
	Ω_3	34	76.4%	**48.8**%
	Ω_4	4	11.9%	11.9%
$U = 5$	Ω_1	153	24.2%	24.2%
	Ω_2	98	74.2%	**39.1**%
	Ω_3	61	35.9%	35.9%
	Ω_4	15	35.9%	**31.4**%
	Ω_5	1	100.0%	100.0%
$U = 6$	Ω_1	195	26.1%	26.1%
	Ω_2	87	74.8%	**38.3**%
	Ω_3	40	49.4%	49.4%
	Ω_4	3	34.3%	**19.4**%
	Ω_5	3	21.4%	**10.7**%
	Ω_6	0	-	-

Table 5 WER for whole dataset

	U	Training dataset	Evaluation dataset
Conventional method	1	28.9%	41.7%
Proposed method	2	22.9%	31.4%
	3	23.5%	31.0%
	4	22.5%	31.1%
	5	22.4%	31.2%
	6	22.5%	32.1%

References

[1] L. Rabiner and B.H. Juang: An introduction to hidden Markov models, IEEE ASSP Magazine, Vol.3, No.1, pp.4–16, 1986.

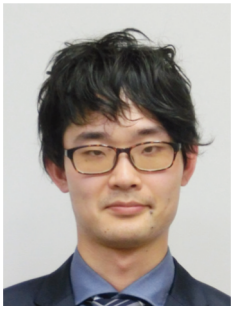
[2] K.F. Lee: Automatic Speech Recognition: The Development of the SPHINX System, Kluwer Academic, 1989.

[3] T. Hori, C. Hori, Y. Minami and A. Nakamura: Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition, IEEE Trans. on Audio, Speech, and Language Processing, Vol.15, No.4, pp.1352–1365, 2007.

[4] G. Hinton, S. Osindero and Y. Teh: A fast learning algorithm for deep belief nets, Neural Comput., Vol.18, No.7, pp.1527–1554, 2006.
[5] M.L. Seltzer, D. Yu and Y. Wang: An investigation of deep neural networks for noise robust speech recognition, Proc. International Conference on Acoustics, Speech, and Signal Processing, pp.7398–7402, 2013.
[6] J. Li and L. Deng: Robust Automatic Speech Recognition: A Bridge to Practical Applications, Academic Press, 2015.
[7] P.C. Loizou: Speech Enhancement: Theory and Practice, 2nd ed., CRC Press, Inc., 2013.
[8] D.H. Johnson and D.E. Dudgeon: Array Signal Processing: Concepts and Techniques, Simon & Schuster, 1993.
[9] H. Cox, R.M. Zeskind and M.M. Owen: Robust adaptive beamforming, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol.35, No.10, pp.1365–1376, 1987.
[10] C. Marro, Y. Mahieux and K.U. Simmer: Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering, IEEE Trans. on Speech and Audio Processing, Vol.6, No.3, pp.240–259, 1998.
[11] R. Zelinski: A microphone array with adaptive post-filtering for noise reduction in reverberant rooms, Proc. International Conference on Acoustics, Speech, and Signal Processing, pp.2578–2581, Apr 1988.
[12] I.A. McCowan and H. Bourlard: Microphone array post-filter based on noise field coherence, IEEE Trans. on Speech and Audio Processing, Vol.11, No.6, pp.709–716, 2003.
[13] M. Brandstein and D. Ward, eds.: Microphone Arrays: Signal Processing Techniques and Applications, Springer, 2001.
[14] K. Kumatani, B. Raj, R. Singh and J.W. McDonough: Microphone array post-filter based on spatially-correlated noise measurements for distant speech recognition, Proc. Annual Conference of the International Speech Communication Association, pp.298–301, 2012.
[15] Y. Hioka, K. Furuya, K. Kobayashi, S. Sakauchi and Y. Haneda: Angular region-wise speech enhancement for hands-free speakerphone, IEEE Trans. on Consumer Electronics, Vol.58, No.4, pp.1403–1410, 2012.
[16] K. Niwa, Y. Hioka and K. Kobayashi: Post-filter design for speech enhancement in various noisy environments, Proc. International Workshop on Acoustic Signal Enhancement, pp.35–39, Sept 2014.
[17] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa and Y. Haneda: Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain, IEEE Trans. on Audio, Speech, and Language Processing, Vol.21, No.6, pp.1240–1250, 2013.
[18] Y. Hioka and K. Niwa: PSD dstimation in beamspace for source separation in a diffuse noise field, Proc. International Workshop on Acoustic Signal Enhancement, pp.85–88, Sept 2014.
[19] T. Arakawa, H. Al-Hassanieh, M. Tsujikawa and R. Isotani: Extended minimum classification error training in voice activity detection, Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp.232–236, Nov 2009.
[20] K. Horii, T. Fukumori, M. Morise, T. Nishiura and Y. Yamashita: The determination of dynamic subtraction for spectral subtraction towards musical tone reduction, The Journal of the Acoustical Society of America, Vol.131, No.4, pp.3446–3446, 2012.
[21] S. Sakauchi, A. Nakagawa, Y. Haneda and A. Kataoka: Implementing and evaluating an audio teleconferencing terminal with noise and echo reduction, Proc. International Workshop on Acoustic Signal Enhancement, pp.191–194, 2003.
[22] S.J. Russell and P. Norvig: Artificial Intelligence: A Modern Approach, Prentice Hall, 2010.
[23] J. Garofolo, D. Graff, D. Paul and D. Pallett: CSR-I (WSJ0) complete. https://catalog.ldc.upenn.edu/LDC93S6A.
[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer and K. Vesely: The kaldi speech recognition toolkit, Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, 2011.
[25] J. Barker, R. Marxer, E. Vincent and S. Watanabe: The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines, Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp.504–511, Dec 2015.
[26] ITU-T: G.191 : Software tools for speech and audio coding standardization. https://www.itu.int/rec/T-REC-G.191.



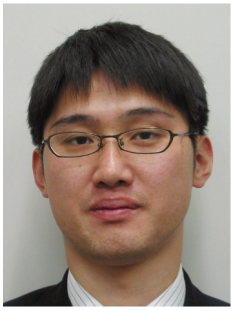
Tomoko Kawase received her B.E. and M.E. degrees in system design engineering from Keio University in 2011 and 2013, respectively. Since joining the Nippon Telegraph and Telephone Corporation (NTT) in 2013, she has been engaged in research on microphone array signal processing. She is now a research engineer at NTT Media Intelligence Laboratories. She is a member of the Acoustical Society of Japan (ASJ).



Kenta Niwa received his B.E., M.E., and Ph.D. degrees from Nagoya University in 2006, 2008, and 2014, respectively. Since joining the Nippon Telegraph and Telephone Corporation (NTT) in 2008, he has been engaged in research on microphone array signal processing. He is now a research engineer at NTT Media Intelligence Laboratories. He was awarded the Awaya Prize by the Acoustical Society of Japan (ASJ) in 2010. He is a member of IEEE, ASJ, and the Institute of Electronics, Information and Communication Engineers (IEICE).



Yusuke Hioka received his B.E., M.E., and Ph.D. degrees in engineering from Keio University in 2000, 2002, and 2005, respectively. From 2005 to 2012, he was with NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories), Nippon Telegraph and Telephone Corporation (NTT). From 2010 to 2011, he was also a visiting researcher at Victoria University of Wellington, New Zealand. In 2013, he moved to New Zealand and was appointed as a Lecturer at the University of Canterbury, Christchurch. Then, in 2014, he joined the Department of Mechanical Engineering, the University of Auckland, Auckland, where he is currently a Senior Lecturer. He has wide interests in audio and acoustics, especially in microphone array signal processing and room acoustics. He is a senior member of IEEE and a member of ASJ and IEICE.



Kazunori Kobayashi received his B.E., M.E., and Ph.D. degrees in electrical and electronic system engineering from Nagaoka University of Technology in 1997, 1999, and 2003, respectively. Since joining the Nippon Telegraph and Telephone Corporation (NTT) in 1999, he has been engaged in research on microphone arrays, acoustic echo cancellers, and hands-free systems. He is now a senior research engineer at NTT Media Intelligence Laboratories. He is a member of the Acoustic Society of Japan (ASJ) and the Institute of Electronics, Information and Communication Engineers (IEICE).

(Received November 4, 2016; revised December 8, 2016)