HOKKAIDO UNIVERSITY

Title	A Noise-Robust Continuous Speech Recognition System Using Block-Based Dynamic Range Adjustment
Author(s)	Sun, Yiming; Miyanaga, Yoshikazu
Citation	IEICE Transactions on Information and Systems, E95D(3), 844-852 https://doi.org/10.1587/transinf.E95.D.844
Issue Date	2012-03-01
Doc URL	https://hdl.handle.net/2115/49207
Rights	Copyright © 2012 The Institute of Electronics, Information and Communication Engineers
Type	journal article
File Information	TISe95D-3_844-852.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

IEICE TRANS. INF. & SYST., VOL.E95–D, NO.3 MARCH 2012

PAPER

A Noise-Robust Continuous Speech Recognition System Using Block-Based Dynamic Range Adjustment

Yiming SUN[†a)], *Nonmember and* Yoshikazu MIYANAGA[†], *Fellow*

SUMMARY A new approach to speech feature estimation under noise circumstances is proposed in this paper. It is used in noise-robust continuous speech recognition (CSR). As the noise robust techniques in isolated word speech recognition, the running spectrum analysis (RSA), the running spectrum filtering (RSF) and the dynamic range adjustment (DRA) methods have been developed. Among them, only RSA has been applied to a CSR system. This paper proposes an extended DRA for a noise-robust CSR system. In the stage of speech recognition, a continuous speech waveform is automatically assigned to a block defined by a short time length. The extended DRA is applied to these estimated blocks. The average recognition rate of the proposed method has been improved under several different noise conditions. As a result, the recognition rates are improved up to 15% in various noises with 10 dB SNR.

key words: *CMS, CSR, DRA, noise-robust, RSA*

1. Introduction

Recently, continuous speech recognition (CSR) has made great progress and yielded a high recognition rate [1]. The high recognition rate can be achieved under clean and high SNR, i.e., over 20 dB SNR, environments. However, current CSR technology has not matured to provide high recognition accuracy under severe noisy environments, i.e., the conditions lower than 20 dB SNR [2].

On the other hand, some noise robust speech recognition methods have been developed for isolated speech, i.e., isolated words and phrases. For the improvement of speech recognition performance, the spectrum subtraction (SS), RelAtive SpecTrA (RASTA), Cepstral mean subtraction (CMS), running spectrum filtering and dynamic range adjustment (RSF/DRA) and running spectrum analysis (RSA) have been used [3]–[5]. They can efficiently reduce the noise effects from noisy speech data. Even when an environment noise is lower than 20 dB SNR, the isolated speech recognition system with noise robust techniques can recognize target speech with high recognition rate.

Although RASTA is a well known method focusing on modulation spectrum domain (MSD), a primary RASTA employs IIR filtering and it may cause a problem such as phase distortion [6]. RSF is based on a FIR filter. RSA is directly used in the MSD. Compared with RASTA and RSF, RSA can realize ideal processing [7], [8]. In this paper, we select RSA to reduce any noise effects on the MSD.

Manuscript received September 5, 2011.
Manuscript revised November 9, 2011.

[†]The authors are with the Graduate School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060–0814 Japan.
a) E-mail: sunny@icn.ist.hokudai.ac.jp

DOI: 10.1587/transinf.E95.D.844

Among the noise robust methods used in a CSR system [9]–[11], a method using RSA and CMS has been developed in [11] and it can show a little higher performance than others. The RSA and CMS are used for the reduction of distortion embedded into a training data set and the CMS is also used for the time invariant noise reduction to an observed speech waveform in a recognition stage. By using the above noise robust techniques, the recognition accuracy can be improved. However, compared with the results of isolated speech recognition accuracy, its performance is insufficient for many actual applications.

In this paper, the modified technique of a dynamic range adjustment (DRA) is proposed for a CSR system. The speech waveform is observed within unlimited time length since any continuous speech data are supposed to the CSR system. On the other hand, the dynamic range of speech features disturbed by any noises should be properly adjusted in order to minimize the difference between the dynamic range of clean speech features and that of noisy speech features. The proposed method introduces a short time length block chosen stochastically from the feature sequence of continuous speech. Using these given blocks, the DRA algorithm is properly applied. By using such processing, the proposed CSR system can show higher speech recognition accuracy where 15 different noise types are used with 10, 15 and 20 dB SNR.

Section 2 introduces three methods we used in simulation. Section 3 shows influence of noises in continuous speech data. Section 4 details a block-based DRA algorithm. Section 5 describes the conditions in the procedure of model training and recognition. Section 6 presents all conditions and results in modeling and recognition.

2. Conventional Methods

2.1 CMS

CMS is a channel normalization approach to compensate for the acoustic channel [12]. The time invariant channel parameters in a recording system and convolutional disturbance noise are evaluated by CMS and these noises are reduced from an observed speech waveform. By using CMS, the distortion between training speech data and observed speech data can be improved.

In CMS, the averages of all MFCC components are calculated, and then these averages are subtracted from MFCC [13] components. CMS can remove the channel ef-

Copyright © 2012 The Institute of Electronics, Information and Communication Engineers

fects happened in the convolutional distortion. Since we have no information about the microphone system and any other convolutional effects for recording the speech, we choose CMS as one of preprocessing methods.

2.2 RSA

RSA is applied for both of low and high frequency components in modulation spectrum domain (MSD). The components of low and high frequency in MSD are reduced by using RSA [14]. The reduction of low frequency components has the same effect of CMS technique. In addition, the reduction of high frequency components results in the elimination on time varying noises which cannot be created by a human speech production.

The speech features are calculated from observed speech. In this paper, MFCCs are used in a speech feature vector. Let us assume that we obtain M speech feature vectors defined as:

$$\mathbf{s}_i = [s_{i,1}\ s_{i,2}\ s_{i,3}\ \ldots\ s_{i,L}]^T, i = 1, \ldots, M. \tag{1}$$

where L denotes the number of speech features, i denotes a time index, and T stands for a transpose. The above feature vector consists of MFCC, Δ MFCC and $\Delta\Delta$ MFCC, where Δ MFCC is calculated from the differentiation of MFCC and $\Delta\Delta$ MFCC is calculated from the differentiation of Δ MFCC. In order to obtain the frequency components of MSD, the following equations are applied:

$$\mathbf{p}_k = \sum_{i=1}^{M} \mathbf{s}_i e^{\frac{-j2\pi ki}{M}} \tag{2}$$

$$\mathbf{c}_k = f_{RSA}[\mathbf{p}_k] \tag{3}$$

$$\hat{\mathbf{s}}_k = \frac{1}{M} \sum_{i=1}^{M} \mathbf{c}_k e^{j\frac{2\pi ki}{M}} \tag{4}$$

where $k = 1, 2, \ldots, M$. Equation (3) indicates the function of RSA for $\mathbf{p}_k$. RSA reduces the value of $\mathbf{p}_t$ where $t = 0$ and $t > N$ where N is decided as the cut-off frequency of higher band in MSD. The vector $\hat{\mathbf{s}}_k$ is RSA speech features in which noise components are reduced.

2.3 DRA

When any noises are added to speech data, the estimated speech features are affected and distorted by these noises. The dynamic range of each MFCC component in MSD is normally affected. In addition, the RSA which reduces the influences of noises changes the dynamic range of MFCC component in MSD. From these reasons, the adjustment of the dynamic range on MFCC trajectory in MSD has been developed [15].

The dynamic range adjustment (DRA) adjusts the dynamic range of MFCC on MSD by normalizing the amplitude of each component.

If we define the i-th component of $\hat{\mathbf{s}}_k$ as $\hat{s}_{k,i}$, DRA calculates the following new value:

$$s'_{k,i} = \frac{\hat{s}_{k,i}}{\max_{j=1,\ldots,M}[\hat{s}_{j,i}]} \tag{5}$$

where $s'_{k,i}$ denotes the i-th element of the MFCC feature vector after DRA.

3. Influence of Noises in Continuous Speech Data

3.1 Noise Disturbance

Figure 1 shows an example of noise influence in an isolated word. In Fig. 1, there are two different trajectories, i.e., the trajectories of the 2nd MFCC calculated from a clean speech and a noisy speech with 10 dB SNR white noise. Note that both MFCC are estimated from the same speech sound. However, due to the serious influence of noise, the dynamic range of MFCC from the noisy speech is much smaller than the others. If a clean speech is used in the HMM straining stage, the automatic speech recognition (ASR) system cannot correctly recognize any noisy speech because of such difference. The DRA method has been developed as the compensation method for such difference in an isolated word and phrase.

Figure 2 shows an example of a clean continuous speech waveform and its instant power trajectory. Normally an observed continuous speech consists of many words and

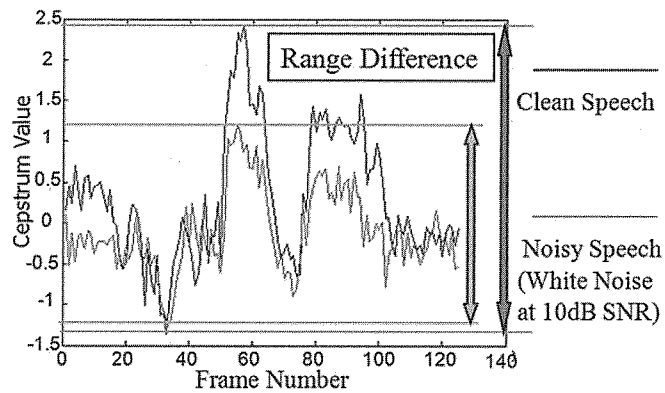


Fig. 1 Noise influences in word feature vectors.

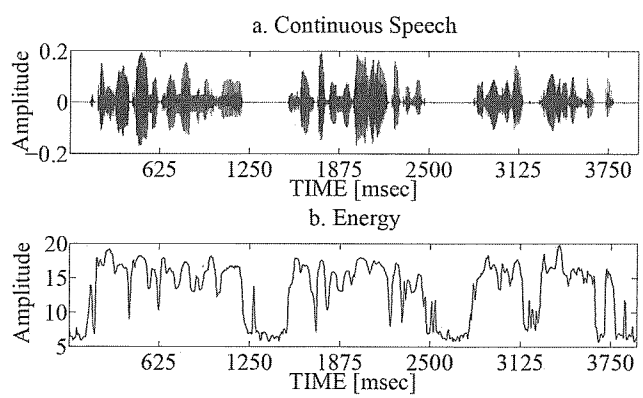


Fig. 2 Continuous speech in clean condition.

phrases and thus its dynamic rage is decided from the maximum energy selected among the continuous speech. The conventional DRA may employ its maximum value and then apply its value to all MFCC components. However, as we can easily recognize the difference among the dynamic ranges of all words in the example of Fig. 2a, the dynamic rage should be carefully adjusted in each word.

Although only a clean continuous speech can be observed, the selection of each word and the dynamic rage adjustment for the selected word are not difficult. However, under noisy conditions, the selection of words may be difficult issue. In this paper, the following two step processing is considered.

(1) From an observed noisy continuous speech waveform, all short sentences are selected.

(2) A short sentence is divided into several blocks and then each block is independently applied by DRA.

The above processing is applied to an observed unknown continuous speech in recognition.

The definition of the short sentence is given on acoustical conditions. If a speech waveform has a certain length of silent, e.g., 200 *msec*, the location of this silent part is called "speech pause". The short sentence consists of "speech pause", its following speech and again "speech pause" after the speech. The defined short sentence may represent a word, several words, a phrase and some phrases. Although the exact meaning of a sentence cannot be defined on acoustic data, the above simple definition can be used in the proposed method.

In the first step (1), non-speech parts are eliminated. A continuous speech has many non-speech parts and only noises. These parts effects DRA inappropriately. In the second step (2), the unbalance of several dynamic ranges existed in a continuous speech can be compensated.

3.2 Sentence Selection

In the training stage, the set of continuous speech data is given as a prior information. From these given speech data, a short sentence is manually selected. There are many speech waveforms with high SNR environment. Figure 2 shows one of example. In other words, the selection of a short sentence can be easily executed. The length of speech pause is defined as 12 window frames. In Fig. 2, we can select three short sentences.

However, in the recognition stage, a speech was normally observed with several noises where SNR was low. In addition, a short sentence should be selected automatically. Figure 3 represents an observed speech waveform as an example. In order to detect a short sentence, 30-frame-width-window is used and its window is shifted by 15 frames. We define $E_{j,n}$ ($1 \leq n \leq 30$) as the energy for the n-th frame in the window. In the selected 30-frame-width-window, three lowest energy values $E_{j,n}$, which satisfy the both limitations $E_{j,n} < E_{j,n-1}$ and $E_{j,n} < E_{j,n+1}$ at the same time among 30 different energy values, are selected. The average of three low energy values is assumed to be noise energy and thus

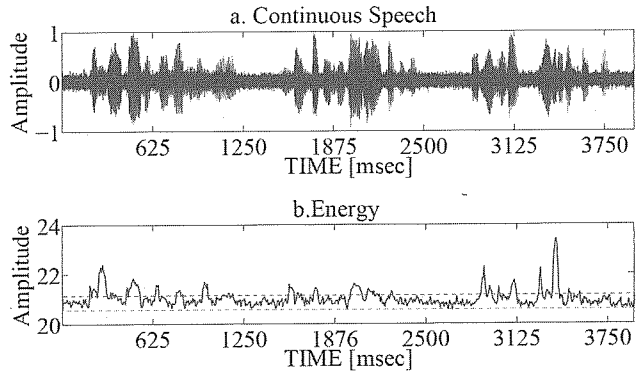


Fig. 3 Continuous speech in 10 dB SNR white noise condition.

the 1.5 times as much as the average value is defined as a threshold. When all 30 energy values in a 30-frame-width-window are lower than the threshold, this window includes non-speech and unvoiced speech. If such windows including non-speech and unvoiced speech are succeedingly detected, the zero-crossing point nearest the center point of the first window among such succeeding windows is estimated as the end point of a short sentence and that of the last window is estimated as the start point of the next short sentence.

4. Block Based DRA

4.1 A Short Sentence and Blocks

In this paper, the proposed algorithm identifies a block between the zero-crossings of $p_{j,i}$ in a short sentence, where $p_{j,i}$ denotes the i-th feature vector in j-th dimension. The definition of the block is a part between the zero-crossing points in the trajectory of $p_{j,i}$. Please note that the different location of a block is used for $p_{j,i}$ at different i. In an estimated block, we use different maximum value to calculate the $p'_{j,i}$ from $p_{j,i}$ by DRA.

In Fig. 4, the simple concept of block separation is explained. The algorithm finds out the maximum value in a given short sentence, i.e., "Peak Point" in Fig. 4. From the peak point, the L_m length of the forward and backward positions is decided. In the forward short sentence from the peak point, the algorithm finds out the first zero-crossing point over the L_m length. In the backward short sentence from the peak point, the algorithm finds out the first zero-crossing point over the L_m length. The main block is selected between the above two zero-crossing points.

In the right-hand side of the short sentence from the main block, the algorithm finds out the first zero-crossing point over L_w length from the right edge of the main block. Between these zero-crossing points, the next block is selected. In the left hand side of the short sentence from the main block, the same procedure is applied.

In a short sentence, the main block has a larger peak value than others. In other words, the lengths of $2L_m$ and L_w are given by different lengths and they are decided from

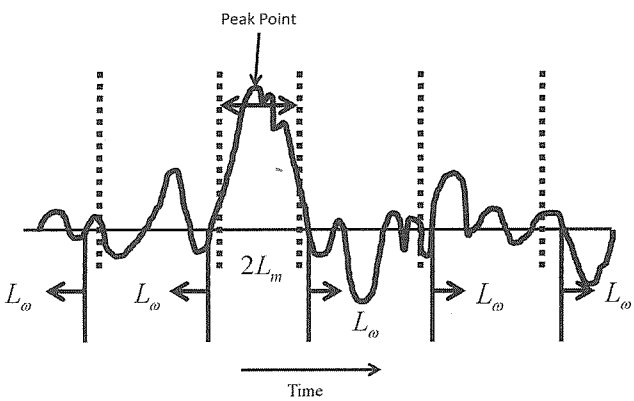


Fig. 4 An example for a short sentence and blocks.

prior experiments.

4.2 First Step: Block Separation

As mentioned in Sect. 4.1, the trajectory of $p_{j,i}$ given in a sentence is divided into some blocks with the zero-crossing points. The proposed algorithm searches the zero-crossing points in $p_{j,i}$ by the equation:

$$f_{j,i} = p_{j,i-1}/p_{j,i}. \tag{6}$$

If $f_{j,i} < 0$, there is a zero-crossing point between $p_{j,i-1}$ and $p_{j,i}$.

The value of $P_{0,j}$ is defined as the maximum value of the peak point. Then $L_j(P_0)$ is recorded as the location of $P_{0,j}$. After that, the length of L_m frames is subtracted and added from $L_j(P_0)$, and they are recorded as $\hat{L}_j(P_{-1}) = L_j(P_0) - L_m$ and $\hat{L}_j(P_1) = L_j(P_0) + L_m$. Next the algorithm searches the zero-crossing points nearest to left-hand side of $\hat{L}_j(P_{-1})$ and the right-hand side of $\hat{L}_j(P_1)$. Once the algorithm finds the zero-crossing points, we define them as the $L_j(P_{-1})$ and $L_j(P_1)$ as the locations of the zero-crossing points in this block. From $L_j(P_{-1})$ to $L_j(P_1)$, we can get the main block. The range of the main blocks is from the start-point as $L_j(P_{-1})$ to the end-pint as $L_j(P_1)$.

There are numerous zero-crossing points in a short sentence due to noise. Furthermore, the noise caused some abrupt changes between zero-crossing points. We consider limitations to select the zero-crossing points of blocks. The limitations focus on preserving the continuity of the $p_{j,i}$ in zero-crossing points. If $p_{j,i}$ is zero-crossing point, $|p_{j,i+1}| < 2$ and $|p_{j,i-1}| < 2$, it means a smooth variation near this zero-crossing point. Otherwise, there is a discontinuity between $p_{j,i+1}$ and $p_{j,i-1}$. In other words, the zero-crossing points used in a short sentence are selected under the above limitations.

We continue to divide the other two segments into blocks. The shortest length of a block is defined as L_w, i.e., $L_j(P_i) - L_j(P_{i+1}) > L_w$. Nearest to $L_j(P_{i+1})$, we use Eq. (6) to search zero-crossing points which satisfy the limitations. Then, we set $i = \pm i \pm 1$ to search the next block boundary. Symbol $\pm i$ is the $\pm i$-th block whose boundary satisfies the above limitations.

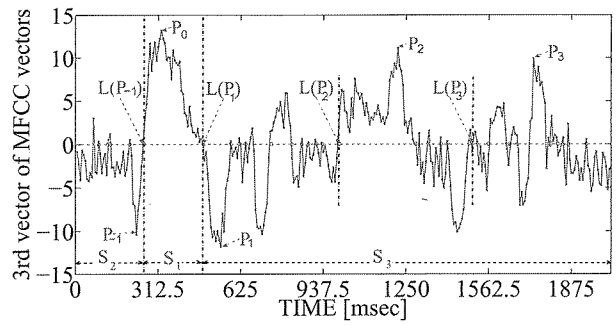


Fig. 5 An example of $p_{j,i}(j = 3)$ for block separation and determination of maximum value.

From the above selection, we can get all zero-crossing points which give the block boundaries. They are given as $L_j(P_{-N})$, $L_j(P_{1-N})$, $L_j(P_{2-N})$, ..., $L_j(P_{-1})$, $L_j(P_1$, ..., $L_j(P_M)$. The main block is given from $L_j(P_{-1})$ to $L_j(P_1)$. In the left-hand side, the $-i$th block is given from $L_j(P_{-i-1})$ to $L_j(P_{-i})$. In the right-hand side, the i-th block is given from $L_j(P_i)$ to $L_j(P_{i+1})$.

Figure 5 shows an example of blocks, where $j = 3$. In Fig. 5, two longer vertical dot-lines show the boundary of the main block. S_1, S_2 and S_3 indicate the main block, the left-hand side block and the right-hand side blocks, respectively. The shorter vertical dot-lines show the boundary of a block in S_3.

4.3 Second Step: Determination of the Maximum Value

From the above step, several blocks are selected and they have many peaks. In this step, the algorithm finds out the adjustment value used in the block-based DRA. Although the conventional DRA employs the maximum value in an observed MFCC trajectory as the adjustment value, the proposed block-based DRA has an additional restriction for this determination.

The value of $P_{\pm i,j}$ is defined as the maximum value within the $\pm i$-th block in a right-hand side block and a left-hand side block.

The proposed algorithm uses the assumption in which there is not large difference between the adjustment values of neighborhood blocks. If we assume such difference value is δ_p, then the adjustment value in the right-hand side is calculated as follows:

(1) Determine the maximum value among $P_{i,j}$ (i = 1, 2, ..., M) as $T_{1,j}$.

(2) If $P_{0,j} - T_{1,j} < \delta_p$ and $T_{1,j} - P_{i,j} < \delta_p$, then $P_{i,j}$ is selected the adjustment value in the $\pm i$-th block.

(3) If $P_{0,j} - T_{1,j} < \delta_p$ or $T_{1,j} - P_{i,j} < \delta_p$, then the adjustment value is given as $T_{1,j}$. In other words, $P_{i,j} = T_{1,j}$.

(4) If $P_{0,j} - T_{1,j} > \delta_p$ and $T_{1,j} - P_{i,j} > \delta_p$, then the adjustment value is give as $P_{0,j} - \delta_p$. In other words, $P_{i,j} = P_{0,j} - \delta_p$.

In the left-hand side, we apply the same calculation.

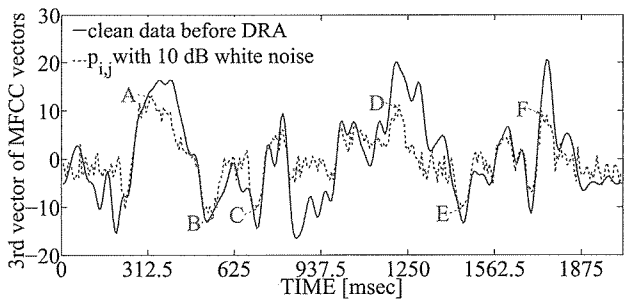


Fig. 6 Before DRA in CSR.

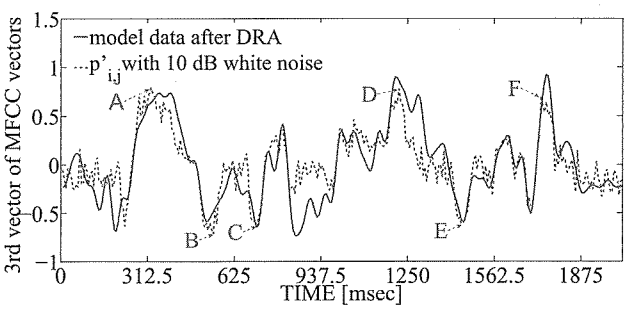


Fig. 7 Results for block-based DRA algorithm in CSR.

4.4 Third Step: Using Block-Based DRA

We have obtained all blocks from a short sentence and determined adjustment values. In each block, the following block-based DRA is applied:

$$p'_{k,i} = \frac{p_{k,i}}{P_{\pm i,j}}, \tag{7}$$

Figure 6 shows the same MFCC feature vectors between the clean and 10 dB SNR white noise conditions. Figure 7 shows the results for the block-based DRA algorithm. In Fig. 6, almost all of $|p_{j,i}|$ in noisy speech are smaller than that of clean speech feature at the same time, especially in marked position from A to F. If we use the proposed algorithm, we can adjust the features of noisy speech. Figure 7 shows the adjustment happened at the mark positions from A to F. It means the proposed algorithm effectively increases the similarity between clean and noisy speech features.

5. Discussion

In the training of HMM [16], [17], all sentences are assumed to be recorded under clean or low noise situation. In other words, any time varying noises and high level noises are not considered in this training stage. From these reasons, conventional CMS, RSA and DRA are applied to all given training speech data set.

The cepstral variance normalization (CVN) technique normalizes the feature variance to the same scale. In particular, CVN has been developed in [18] which is applied to the recognition of Japanese digit strings. The cepstral mean normalization (CMN) and CVN are used in cascade to execute the mean and variance normalization (MVN). The concept of our proposed method is similar to the above method. Our proposed method focuses on any Japanese character strings. In our method, the segmentation, i.e., 4.2, is designed for any character strings against high noisy circumstances. The result comparisons are given in Table 4 and from Table 6 to Table 11.

Furthermore, in the recognition, many various and different noises should be considered during the recording to speech waveform. Accordingly, the proposed block-based DRA is applied. Numerical comparison results for MVN and our proposed method will be shown in Sect. 6.

Table 1 Long vowel frame average length [%].

Phoneme	Averages	Variance	Appear Times
a:	13.35	13.50	2054
e:	14.46	15.59	12688
i:	14.93	20.97	1724
o:	13.83	19.01	37657
u:	10.64	17.50	4831

5.1 Model Training Stage

Even when the speech data sets for the training are recorded under low noise circumstances, the effect of convolutional disturbance, i.e., microphone, may influence speech features. During the training stage for HMMs, CMS, RSA and DRA should be used where conventional systems have employed only CMS and CMS/RSA.

As the merit of RSA, the un-speech feature over 15 Hz on MSD can be accurately reduced than RSF. In addition, using RSA with CMS, the noise and disturbance components can be eliminated effectively.

The effects of CMS and RSF are not small for the dynamic range of speech feature trajectory mentioned in the previous section. The conventional DRA is applied for the dynamic rage normalization of their estimated and processed speech features.

Table 1 shows the averages and variance of five long vowels in all training data.

5.2 Speech Recognition Stage

CMS and RSA can reduce any impulsive noise before block-based DRA. In the speech recognition stage, the block-based DRA is applied. In the speech recognition, it is impossible to know the length of speech waveform as prior information. In addition, during recording, some different noises and disturbances may happen. For the reduction of noise and disturbance, the proposed block-based DRA is applied.

For conventional DRA, we use $\frac{p_{j,i}}{P_{0,j}}$ for normalization. From the Sect. 4.3 (2) to (4), we can compute the inequality $P_{0,j} > P_{i,j} > P_{0,j} - 2\delta_p$ and $P_{0,j} > P_{i,j} > P_{0,j} - \delta_p$. In Sect. 4.3, we have set $P_{i,j} = P_{0,j} - 2\delta_p$ or $P_{i,j} = P_{0,j} - \delta_p$ as the adjustment value. If we suppose $P_{0,j} = 13$, $T_{1,j} = 11$,

$P_{i,j} = 9$ and $\delta_p = 2$, it satisfies the Sect. 4.3 (4). Then, we substitute $P_{i,j} = P_{0,j} - \delta_p$ into Eq. (7). For the point $|p_{j,i}| = 9$ in Fig. 8, the $|p'_{j,i}|$ improves 0.1 compared with conventional DRA. If the values of $P_{0,j}$, $T_{1,j}$ and $P_{i,j}$ satisfy Sect. 4.3 (2) or (3), we substitute $P_{i,j} = P_{0,j} - 2\delta_p$ into Eq. (7). The $|p'_{j,i}|$ improves more compared with conventional DRA. Therefor, we set the δ_p to 2 in Sect. 4.3.

In Fig. 8, the horizontal axis denotes the MFCC value before DRA and the vertical axis denotes the MFCC value after DRA. We have known as the dynamic range of MFCC from the noisy speech is much smaller than the others from Fig. 1. In horizontal axis, the large value means the MFCC value under clean condition. Otherwise, the small value means the MFCC value under noise conditions.

As well as for the same $|p_{j,i}|$ in Fig. 8, if $|p_{j,i}| < 2$, the deviation is less than 0.03 between the neighbor maxima. This deviation between the neighbor maxima is acceptable in the range from -1 to 1 and thus we set $|p_{j,i-1}| < 2$ or $|p_{j,i+1}| < 2$ in Sect. 4.2.

From Table 1, the averages of all vowels are less than 15. The main block width is longer than $2L_m$ from Sect. 4.2. If we set $L_m = 15$, the main block width include at least a vowel.

The value of L_w determines the block width. If the value of L_w is large, it causes large changes into a block and leads to recognition rate abrupt decrease. On the other hand, if the value of L_w is small, it causes small changes into a block and leads to the recognition rate close to the results by using conventional DRA. As Fig. 9 shown, the recognition result becomes high when we set $L_w = 80$.

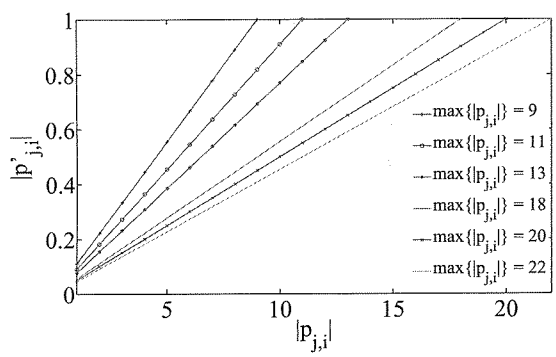


Fig. 8 The normalization effect by using different maxima for same $|p_{j,i}|$.

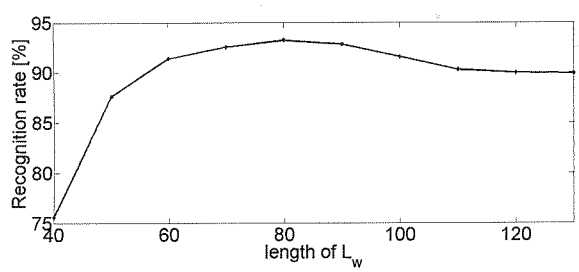


Fig. 9 Recognition rate with different L_w.

6. Results

In our experiments, all HMMs have been trained by using JNAS database [19]. It is produced by 153 males' native Japanese speakers. The conditions on speech analysis are given in Table 2.

We use two criterions for the evaluation of speech recognition:

$$R_C = \frac{N - S - D}{N} \times 100 \quad [\%], \tag{8}$$

$$R_A = \frac{N - S - D - I}{N} \times 100 \quad [\%], \tag{9}$$

where N is the total number of words in the set of speech sentences, S is the number of misrecognized words, D denotes the number of words which are not selected as words, I denotes the number of words which are misrecognized as words, i.e., noise components and non-speech. Above, R_C shows the correct word recognition rate for the entire set of speech words, and R_A shows the accuracy of the total CSR performance.

In recognition, we use the sets of known and unknown data for our recognition tests. Known data denotes the test data which comes from the training database. Unknown data denotes the test data are collected from by Hokkaido university students, where the sentences are different from the training database. The conditions in this experiment are shown as Table 3. Additionally, we use Julius as an evaluation tool in the recognition.

We have simulated all data under both of clean condition and noise conditions with different SNRs. We define 15 kinds noise as Table 5.

In all tables, the 'Proposed' column denotes the method

Table 2 Acoustic analysis conditions.

Sampling frequency	16 kHz
Frame shift	10.0 ms
Frame length	25.0 ms
Window type	Hanning
Training data	23651 sentences from 153 people
Emphasizing of High Frequency	$1 - 0.97z^{-1}$
HMM state number	5 states (include start and end states)
Number of Gaussian Mixtures	16
Clustering	about 2000 states

Table 3 Recognition conditions.

Known data for testing	50 sentences from 12 people
Unknown data for testing	180 sentences from 6 people
Sampling and frame conditions	the same with Table 2

Table 4 Recognition rates for clean condition [%].

	Proposed		RSA		MVN		Con DRA	
	Corr	Acc	Corr	Acc	Corr	Acc	Corr	Acc
known data	93.22	92.29	92.55	91.22	91.29	90.09	89.63	88.03
unknown data	83.90	82.43	82.69	81.00	82.87	81.43	82.69	81.33

Table 5 Noise definition.

Symbol	Noise Name	Symbol	Noise Name	Symbol	Noise Name
N1	babble	N2	buccaneer1	N3	buccaneer2
N4	destroyerenginer	N5	destroyerops	N6	f16
N7	factory1	N8	factory2	N9	hfchannel
N10	leopard	N11	m109	N12	machinegun
N13	pink	N14	volvo	N15	white

Table 6 Known data recognition rates at 20 dB SNR [%].

	Proposed		RSA		MVN		Con DRA	
	Corr	Acc	Corr	Acc	Corr	Acc	Corr	Acc
N1	76.06	70.84	76.46	73.01	76.26	68.48	75.12	72.76
N2	76.46	74.87	73.40	71.68	72.33	71.00	70.69	69.59
N3	76.86	74.34	73.94	72.74	74.52	73.59	69.22	67.84
N4	75.66	73.67	77.39	75.93	75.20	74.00	73.80	71.94
N5	81.65	80.32	82.85	81.78	80.04	79.11	79.79	78.86
N6	78.32	76.86	74.60	72.74	75.85	74.79	74.34	72.87
N7	81.65	80.32	77.79	76.60	77.79	76.46	75.80	74.87
N8	87.90	86.97	77.79	76.60	80.90	79.84	75.80	74.87
N9	67.42	64.49	57.31	55.05	64.36	62.33	63.56	60.77
N10	85.64	81.65	86.57	83.11	85.64	81.52	82.63	81.24
N11	88.56	87.50	89.89	88.96	88.09	87.89	87.23	86.30
N12	84.04	78.19	80.98	74.47	82.91	75.86	81.48	75.92
N13	80.32	79.26	76.99	76.06	79.78	78.45	79.41	78.09
N14	89.89	88.83	91.09	89.89	90.69	89.36	90.43	88.83
N15	68.22	65.65	69.95	68.62	69.40	68.07	63.56	61.84
Ave	80.08	77.72	77.80	75.82	78.25	76.05	76.19	74.44

Table 7 Known data recognition rates at 15 dB SNR [%].

	Proposed		RSA		MVN		Con DRA	
	Corr	Acc	Corr	Acc	Corr	Acc	Corr	Acc
N1	62.90	55.32	62.63	59.31	59.84	48.54	55.23	50.33
N2	54.79	52.93	41.49	40.65	50.24	47.71	54.52	53.17
N3	60.11	57.31	44.02	41.89	52.63	50.9	52.53	50.04
N4	63.83	60.24	56.78	54.79	52.63	49.44	58.27	55.80
N5	74.07	71.81	66.36	64.63	65.80	63.40	68.44	65.67
N6	64.23	62.23	54.39	52.26	57.95	55.69	60.18	58.41
N7	65.03	63.30	55.72	54.26	55.69	52.77	61.16	59.16
N8	77.93	76.46	77.26	75.53	76.99	75.26	73.49	71.42
N9	47.61	43.35	35.77	32.85	46.28	43.25	45.55	42.27
N10	81.91	76.06	82.18	77.53	80.38	75.00	75.57	72.40
N11	85.11	84.04	81.38	80.05	80.30	80.11	77.11	75.11
N12	80.98	72.47	78.86	69.81	80.31	71.27	74.63	66.01
N13	65.82	64.76	49.07	47.07	58.06	55.69	57.81	55.73
N14	89.89	88.83	91.36	90.16	89.69	88.49	79.71	77.45
N15	46.68	43.09	39.23	36.04	42.98	39.79	39.56	37.14
Ave	68.06	64.81	61.10	58.40	63.32	59.82	62.25	59.34

using CMS, RSA and conventional DRA for HMM training, and using CMS and block-based DRA for recognition. The 'RSA' column denotes the method using CMS and RSA for HMM training, and using CMS for recognition. The 'MVN' column denotes CMS, RSA and MVN for HMM training and using CMS and MVN for recognition. The 'Con DRA' column denotes the method using CMS, RSA and conventional DRA for HMM training, and using CMS and conventional DRA for recognition. 'Ave' denotes the average recognition rate in all Tables. Table 4 shows the results in the clean conditions, Table 6, 7 and 8 show the recognition results on training data with 20, 15 and 10 dB SNR conditions. The Table 9, 10 and 11 show the recognition results

Table 8 Known data recognition rates at 10 dB SNR [%].

	Proposed		RSA		MVN		Con DRA	
	Corr	Acc	Corr	Acc	Corr	Acc	Corr	Acc
N1	36.84	25.80	36.04	32.05	36.30	22.61	34.49	25.90
N2	30.19	28.19	15.56	14.63	25.90	22.45	30.35	28.62
N3	34.31	32.31	18.08	16.87	31.62	29.23	30.51	28.88
N4	40.96	38.03	24.07	23.14	34.22	29.70	38.84	37.07
N5	51.33	49.07	34.18	33.24	47.58	43.59	47.41	43.57
N6	39.10	36.17	24.07	23.14	36.82	32.29	31.86	30.61
N7	42.15	39.23	25.93	25.00	36.28	31.49	39.63	38.46
N8	61.70	58.64	52.66	51.06	52.64	48.91	53.12	51.24
N9	25.54	23.24	15.29	13.96	18.80	16.10	18.00	16.08
N10	79.26	73.81	77.93	72.87	77.39	70.61	72.21	68.55
N11	74.07	72.07	61.30	59.97	67.39	65.00	67.04	64.78
N12	79.46	69.88	77.53	66.89	79.19	69.35	72.13	59.54
N13	36.64	33.91	19.28	18.09	35.88	33.75	34.62	33.14
N14	90.03	88.38	88.96	87.50	88.23	87.50	78.21	76.90
N15	28.32	25.00	17.55	16.22	24.04	21.25	23.30	21.49
Ave	49.93	46.25	39.23	36.98	46.15	41.86	44.78	41.66

Table 9 Unknown data recognition rates at 20 dB SNR [%].

	Proposed		RSA		MVN		Con DRA	
	Corr	Acc	Corr	Acc	Corr	Acc	Corr	Acc
N1	73.72	70.14	73.38	69.42	72.96	69.15	73.38	69.42
N2	70.66	68.51	67.46	65.69	70.11	67.85	70.14	68.02
N3	69.61	67.23	66.63	64.74	67.98	65.83	67.36	64.80
N4	77.91	75.17	69.65	67.84	72.66	71.00	72.51	70.89
N5	77.64	75.49	77.11	75.45	77.56	75.08	77.15	75.04
N6	73.00	71.68	70.81	69.16	71.23	69.42	72.93	71.53
N7	73.04	71.34	72.81	71.53	72.40	70.21	73.27	71.53
N8	78.39	76.92	80.09	78.92	79.34	77.19	77.87	76.21
N9	62.10	59.58	60.26	58.63	61.97	60.20	60.07	59.58
N10	76.21	73.30	77.39	74.81	76.13	73.00	75.98	72.85
N11	80.77	79.37	82.65	81.45	80.96	79.03	80.69	79.37
N12	77.92	71.70	76.06	69.57	76.85	71.23	76.06	69.57
N13	74.13	72.44	70.32	68.51	73.77	72.19	73.94	72.13
N14	80.05	77.87	81.83	79.71	81.83	80.13	80.13	77.90
N15	61.16	58.90	60.41	58.07	60.57	58.08	60.63	58.79
Ave	73.76	71.31	72.46	70.23	73.08	70.63	72.81	70.51

Table 10 Unknown data recognition rates at 15 dB SNR [%].

	Proposed		RSA		MVN		Con DRA	
	Corr	Acc	Corr	Acc	Corr	Acc	Corr	Acc
N1	65.20	60.41	65.27	63.08	55.20	48.26	53.28	51.04
N2	54.64	53.39	41.25	40.08	48.14	46.03	48.45	47.64
N3	52.71	50.15	40.05	38.54	48.11	45.92	47.78	45.92
N4	58.61	56.03	53.09	51.36	52.22	51.22	51.85	50.39
N5	68.44	65.80	63.16	61.49	65.59	62.38	61.12	59.53
N6	60.78	58.79	51.58	49.96	53.91	51.80	52.68	51.35
N7	60.97	59.01	54.22	52.83	55.14	52.58	53.67	52.61
N8	73.27	71.42	71.42	70.02	74.19	72.63	73.47	71.18
N9	45.85	42.72	39.22	37.37	42.74	40.63	42.69	39.36
N10	75.64	72.55	76.24	72.51	74.68	71.21	72.68	70.19
N11	77.26	75.49	78.17	76.89	76.49	74.38	74.67	73.60
N12	74.52	65.82	74.96	66.10	74.24	65.57	73.25	68.41
N13	58.37	56.26	45.29	43.78	52.22	49.99	46.20	44.65
N14	79.64	77.45	81.98	79.90	80.67	79.02	78.23	77.36
N15	39.22	36.88	36.73	35.33	40.45	38.04	37.67	35.41
Ave	63.01	60.14	58.18	55.95	59.60	56.64	57.85	55.91

on unspecific speakers with 20, 15 and 10 dB SNR conditions. All averaged results have shown highest accuracy on the proposed method under noisy conditions. Especially, at

Table 11 Unknown data recognition rates at 10 dB SNR [%].

	Proposed		RSA		MVN		Con DRA	
	Corr	Acc	Corr	Acc	Corr	Acc	Corr	Acc
N1	44.72	36.05	38.80	36.69	33.51	32.65	32.80	30.69
N2	30.24	28.43	14.48	13.88	24.01	22.09	14.48	13.88
N3	31.11	29.52	18.93	18.29	26.54	25.35	18.93	18.29
N4	38.35	36.20	21.08	20.51	34.12	31.67	21.08	20.51
N5	49.98	47.18	34.92	34.28	42.64	39.13	34.92	34.28
N6	41.70	40.42	19.42	18.55	36.08	34.08	19.42	18.55
N7	39.44	38.08	23.23	22.89	31.03	28.73	23.23	22.89
N8	62.97	60.90	54.86	53.51	52.08	51.12	51.86	50.51
N9	22.81	20.59	14.29	13.57	16.06	14.10	14.29	13.57
N10	74.21	70.14	72.21	68.55	72.25	68.55	74.74	71.98
N11	69.85	67.55	59.31	57.88	68.78	66.02	59.31	57.88
N12	74.21	61.46	72.13	59.54	74.06	62.86	73.19	60.48
N13	36.29	34.90	17.87	17.16	33.44	31.63	17.87	17.16
N14	79.32	77.24	78.21	76.90	79.05	77.05	78.09	76.12
N15	24.04	22.62	16.25	15.61	19.41	17.68	16.25	15.61
Ave	47.91	44.75	37.06	35.19	42.87	40.18	36.70	34.95

10 dB SNR, all results have been improved, and the average recognition rate has been improved by more than 10%.

7. Conclusions

In this paper, a new noise robust continuous speech recognition system has been proposed. In this system, a new block-based dynamic range adjustment (DRA) algorithm has been implemented into the module of unspecific speaker recognition. The proposed method has enhanced the recognition rate under lower SNR noise environments. The DRA normalizes the maximum amplitudes of MFCC in each selected block. The proposed CSR system can show higher accuracy than conventional systems under 20, 15 and 10 dB SNR noise environments. Especially in *destroyerenginer*, *destroyerops*, *factory*1 and *pink* noise environment, more than 15% improvement can be obtained for known and unknown data. Our target is noise-robust Japanese character string recognition. Compared with [20] and [21], the both paper are aiming to Japanese digital strings recognition. The training database and language model are different as well.

Acknowledgement

The authors would like to thank the global COE program, Graduate School of Information Science and Technology, Hokkaido University for fruitful discussions. This study is supported in parts by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B2) (20300014).

References

[1] K. Iwano, T. Seki, and S. Furui, "Noise robust speech recognition using F_0 contour information," IEICE Trans. Inf. & Syst., vol.E87-D, no.5, pp.1102–1109, May 2004.

[2] A. Acero, Acoustical and Environmental Robustness in Automatic Speech Recognition, Ph.D. Thesis, Carnegie Mellon University, USA, 1990.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol.27, no.2, pp.113–120, April 1979.

[4] N. Wada, N. Hayasaka, S. Yoshizawa, and Y. Miyanaga, "Direct control on modulation spectrum for noise-robust speech recognition and spectral subtraction," International Symposium on Circuits and Systems, pp.2532–2536, 2006.

[5] K. Ohnuki, The Research about Robust Acoustic Modeling and Continuous Speech Recognition System, Ph.D. Thesis, Hokkaido University, Japan, 2009.

[6] S. Yoshizawa and Y. Miyanaga, "Robust recognition of noisy speech and its hardware design for real time processing," ECTI Trans. Elect. Eng., Electron., and Commun., vol.3, no.1, pp.36–43, Feb. 2005.

[7] N. Ohtsuki, Y. Uchikawa, and Y. Miyanaga, "Speech noise reduction using high-precision RSA," IEICE Technical Report, SIS2008-15, June 2008 (in Japanese).

[8] K. Ohnuki, W. Takahashi, and Y. Miyanaga, "Acoustic modeling for robust recognition using running spectrum analysis," IEICE Technical Report, SIS2008-16, June 2008 (in Japanese).

[9] X. Lu, S. Matsuda, M. Unoki, and S. Nakamura, "Temporal modulation normalization for robust speech feature extraction and recognition," International Congress on Image and Signal Processing, pp.1–4, Oct. 2009.

[10] H. Yasui, K. Shinoda, S. Furui, and K. Iwano, "Noise robust speech recognition using spectral subtraction and F_0 information extracted by hough transform," Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference., pp.631–634, Oct. 2009.

[11] K. Ohnuki, W. Takahashi, S. Yoshizawa, and Y. Miyanaga, "New acoustic modeling for robust recognition and its speech recognition system," International Conference on Embedded Systems and Intelligent Technology, 2009.

[12] S.V. Vaseghi, ed., Advanced Digital Signal Processing and Noise Reduction, Second ed., John Wiley & Sons, 2000.

[13] S. Dharanipragada, U.H. Yapanel, and B.D. Rao, "Robust feature extraction for continuous speech recognition using the MVDR spectrum estimation method," IEEE Trans. Audio Speech Language Process., vol.15, no.1, pp.224–234, Jan. 2007.

[14] K. Ohnuki, W. Takahashi, and Y. Miyanaga, "Noise robust speech features for automatic continuous speech recognition using running spectrum analysis," International Symposium on Communications and Information Technologies, pp.150–153, 2008.

[15] W. Naoya and Y. Miyanaga, "Robust speech recognition with MSC/DRA feature extraction on modulation spectrum domain," International Symposium on Communications, Control and Signal Processing, pp.11–14, March 2006.

[16] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol.77, no.2, pp.257–286, Feb. 1989.

[17] P. Banerjee, G. Garg, P. Mitra, and A. Basu, "Application of triphone clustering in acoustic modeling for continuous speech recognition in Bengali," International Conference on Pattern Recognition, pp.1–4, Dec. 2008.

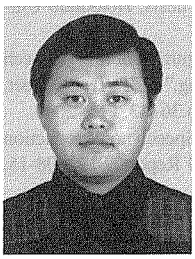
[18] T. Endo and H. Kawai, "Utterance-based mean and segmental variance normalization for robust speech recognition in noisy environments," IEICE Technical Report., SP2007-90, Nov. 2007 (in Japanese).

[19] I. Katunobu, Y. Mikio, T. Kazuya, M. Tatsuo, K. Tetsunori, S. Kiyohiro, and I. Shuichi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," J. Acoust. Soc. Jpn (E), vol.120, no.3, pp.119–206, 1999.

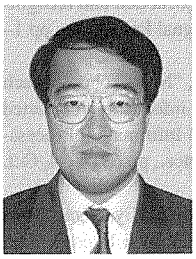
[20] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, K. Itoh, M. Yamamoto, A. Tamada, T. Utsuro, and K. Shikano, "Japanese dictation toolkit: 1998 version," J. Acoust. Soc. Jpn., vol.56, no.4, pp.255–259, April 2000 (in Japanese).

[21] N. Kanedera, T. Arai, K. Okada, and K. Asai, "Continuous speech recognition based on the contribution of modulation spectrum,"

IEICE Technical Report, SP2003-54, June 2003 (in Japanese).



Yiming Sun was born in Jilin, China in 1978. He received the M.S. degree in computer engineering from Northeast Normal University of China in 2007. Currently, he is working toward the Ph.D. degree at the Hokkudai University of Japan. His main research interest is robust continuous speech recognition in clean and noisy environments.



Yoshikazu Miyanaga received the B.S., M.S., and Dr. Eng. degrees from Hokkaido University, Sapporo, Japan, in 1979, 1981, and 1986, respectively. Since 1983 he has been with Hokkaido University. He is now Professor at Division of Information Communication Systems in Graduate School of Information Science and Technology, Hokkaido University. From 1984 to 1985, he was a visiting researcher at Department of Computer Science, University of Illinois, USA. He served as an associate editor of IEICE Transactions on Fundamentals of Electronics, Communications and Computer Science from 1996 to 1999, editors of IEICE Transactions on Fundamentals, Special Issues. He is also an associate editor of Journal of Signal Processing, RISP Japan (2005-present). He was a delegate of IEICE, Engineering Sciences Society Steering Committee, i.e., IEICE ESS Officers from 2004 to 2006. He was a chair of Technical Group on Smart Info-Media System, IEICE (IEICE TG-SIS) during the same period and now a member of the advisory committee, IEICE TG-SIS. He served as a member in the board of directors, IEEE Japan Council as a chair of student activity committee from 2002 to 2004. He is a chair of student activity committee in IEEE Sapporo Section (1998-present). His research interests are in the areas of speech signal processing, wireless communication signal processing and low-power VLSI system design.