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

Yiming SUN\(^{[a]}\), Nonmember and Yoshikazu MIYANAGA\(^{[b]}\), 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.

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.
fected 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:

\[ s_i = [s_{i,1}, s_{i,2}, s_{i,3}, \ldots, s_{i,L}]^T, i = 1, \ldots, M \]  

(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, \( \Delta \) MFCC and \( \Delta \Delta \) MFCC, where \( \Delta \) MFCC is calculated from the differentiation of MFCC and \( \Delta \Delta \) MFCC is calculated from the differentiation of \( \Delta \) MFCC.

In order to obtain the frequency components of MSD, the following equations are applied:

\[ p_k = \sum_{i=1}^{M} s_i e^{-j2\pi ki/N} \]  

(2)

\[ c_\lambda = f_{\text{RSA}}[p_\lambda] \]  

(3)

\[ s_k = \frac{1}{M} \sum_{i=1}^{M} c_\lambda e^{j2\pi ki/N} \]  

(4)

where \( k = 1, 2, \ldots, M \). Equation (3) indicates the function of RSA for \( p_\lambda \). RSA reduces the value of \( p_\lambda \) where \( t = 0 \) and \( t > N \) where \( N \) is decided as the cut-off frequency of higher band in MSD. The vector \( s_k \) 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 \( s_k \) as \( \hat{s}_{k,i} \), DRA calculates the following new value:

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

(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...
phrases and thus its dynamic rage is decided from the max-
imum 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 ob-
served, the selection of each word and the dynamic rage ad-
justment for the selected word are not difficult. However,
under noisy conditions, the selection of words may be dif-
cult issue. In this paper, the following two step processing
is considered.

(1) From an observed noisy continuous speech wave-
form, 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 un-
known continuous speech in recognition.

The definition of the short sentence is given on acousti-
cal 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 followingspeech 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 acous-
tic data, the above simple definition can be used in the pro-
posed 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 sec-
ond step (2), the unbalance of several dynamic ranges ex-
isted 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 nor-
mally 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 ex-
ample. In order to detect a short sentence, 30-frame-width-
window is used and its window is shifted by 15 frames. We
define $E_{jn} (1 \leq n \leq 30)$ as the energy for the $n$-th frame in
the window. In the selected 30-frame-width-window, three
lowest energy values $E_{jn}$, which satisfy the both limitations
$E_{jn} < E_{jn-1}$ and $E_{jn} < E_{jn+1}$ 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
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 includ-
ing non-speech and unvoiced speech are succedingly de-
tected, the zero-crossing point nearest the center point of
the first window among such succeeding windows is esti-
mated 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 be-
tween the zero-crossings of $p_{ji}$ in a short sentence, where
$p_{ji}$ 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 $p_{ji}$. Please note that the differ-
ent location of a block is used for $p_{ji}$ at different $i$. In an esti-
mated block, we use different maximum value to calculate
the $p'_{ji}$ from $p_{ji}$ by DRA.

In Fig. 4, the simple concept of block separation is ex-
plained. 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 posi-
tions 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 be-
tween 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 $L_m$ length from the right edge of the main block.
Between these zero-crossing points, the next block is se-
lected. 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
$L_m$ are given by different lengths and they are decided from
Then, we set \( L_j(P_i) \) from the start-point as the first block. Due to noise, the proposed algorithm searches zero-crossing points which satisfy the limitations. Furthermore, the noise caused some abrupt changes between zero-crossing points. We consider the limitations to select the zero-crossing points of blocks. The limitations focus on preserving the continuity of the zero-crossing points. 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 \( P_{0,j} \) 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 \( L_j(P_{-1}) = L_j(P_0) - L_m \) and \( L_j(P_{+1}) = L_j(P_0) + L_m \). Next the algorithm searches the zero-crossing points nearest to left-hand side of \( L_j(P_{+1}) \) and the right-hand side of \( L_j(P_{-1}) \). 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-point 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 \( L_m \), i.e., \( L_j(P_i) - L_j(P_{i+1}) > L_m \). Nearest to \( L_j(P_{i+1}) \), we use Eq. (6) to search zero-crossing points which satisfy the limitations. Then, we set \( i = \pm 1 \) to search the next block boundary. Symbol \( \pm i \) is the \( \pm i \)-th block whose boundary satisfies the above limitations.

From the above selection, we can get all zero-crossing points which give the block boundaries. They are given as \( L_j(P_{-N}), L_j(P_{-N+1}), \ldots, L_j(P_{-M}), L_j(P_{-M+1}), \ldots, L_j(P_{-1}), L_j(P_1), \ldots, L_j(P_{M}) \). The main block is given from \( L_j(P_{-i}) \) to \( L_j(P_i) \). In the left-hand side, the \( i \)-th block is given from \( L_j(P_{-i}) \) to \( L_j(P_{-i+1}) \). 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 \( \delta_p \), then the adjustment value in the right-hand side is calculated as follows:

1. Determine the maximum value among \( P_{i,j} \) (i.e., \( 1, 2, \ldots, M \)) as \( T_{1,i} \).
2. If \( P_{0,j} - T_{1,i} < \delta_p \) and \( T_{1,i} - P_{i,j} < \delta_p \), then \( P_{i,j} \) is selected as the adjustment value in the \( \pm i \)-th block.
3. If \( P_{0,j} - T_{1,i} < \delta_p \) or \( T_{1,i} - P_{i,j} < \delta_p \), then the adjustment value is given as \( T_{1,i} \). In other words, \( P_{i,j} = T_{1,i} \).
4. If \( P_{0,j} - T_{1,i} > \delta_p \) or \( T_{1,i} - P_{i,j} > \delta_p \), then the adjustment value is given 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. 4** An example for a short sentence and blocks.

**Fig. 5** An example for a short sentence and blocks.
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:

\[ p'_{k,j} = \frac{p_{k,j}}{p_{sl,j}} \]  

(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 \( |p_{k,j}| \) 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 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 range normalization of their estimated and processed speech features.

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

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Averages</th>
<th>Variance</th>
<th>Appear Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>13.35</td>
<td>13.50</td>
<td>2054</td>
</tr>
<tr>
<td>e</td>
<td>14.46</td>
<td>15.59</td>
<td>12688</td>
</tr>
<tr>
<td>i</td>
<td>14.93</td>
<td>20.97</td>
<td>1724</td>
</tr>
<tr>
<td>o</td>
<td>13.83</td>
<td>19.01</td>
<td>37657</td>
</tr>
<tr>
<td>u</td>
<td>10.64</td>
<td>17.50</td>
<td>4831</td>
</tr>
</tbody>
</table>

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 range 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 \( \frac{P_{0,j}}{P_{0,j}} \) for normalization. From the Sect. 4.3 (2) to (4), we can compute the inequality \( P_{0,j} > P_{0,j} > P_{0,j} - 2\delta_p \) 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 \( P_{0,j} = 13, \ T_{i,j} = 11 \), 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 [%].

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Averages</th>
<th>Variance</th>
<th>Appear Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>13.35</td>
<td>13.50</td>
<td>2054</td>
</tr>
<tr>
<td>e</td>
<td>14.46</td>
<td>15.59</td>
<td>12688</td>
</tr>
<tr>
<td>i</td>
<td>14.93</td>
<td>20.97</td>
<td>1724</td>
</tr>
<tr>
<td>o</td>
<td>13.83</td>
<td>19.01</td>
<td>37657</td>
</tr>
<tr>
<td>u</td>
<td>10.64</td>
<td>17.50</td>
<td>4831</td>
</tr>
</tbody>
</table>
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 criteria for the evaluation of speech recognition:

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

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

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 5 Noise definition.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Noise Name</th>
<th>Symbol</th>
<th>Noise Name</th>
<th>Symbol</th>
<th>Noise Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>babble</td>
<td>N2</td>
<td>buccaneer1</td>
<td>N3</td>
<td>buccaneer2</td>
</tr>
<tr>
<td>N4</td>
<td>destroyerenger</td>
<td>N5</td>
<td>destroyerops</td>
<td>N6</td>
<td>H16</td>
</tr>
<tr>
<td>N7</td>
<td>factory1</td>
<td>N8</td>
<td>factory2</td>
<td>N9</td>
<td>lifchanel</td>
</tr>
<tr>
<td>N10</td>
<td>leopard</td>
<td>N11</td>
<td>m109</td>
<td>N12</td>
<td>machingun</td>
</tr>
<tr>
<td>N13</td>
<td>pink</td>
<td>N14</td>
<td>volvo</td>
<td>N15</td>
<td>white</td>
</tr>
</tbody>
</table>
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


<table>
<thead>
<tr>
<th>Proposed CSR system</th>
<th>Unknown data recognition rates at 10 dB SNR [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nr1</td>
<td>32.12</td>
</tr>
<tr>
<td>Nr2</td>
<td>31.85</td>
</tr>
<tr>
<td>Nr3</td>
<td>31.51</td>
</tr>
<tr>
<td>Nr4</td>
<td>31.16</td>
</tr>
<tr>
<td>Nr7</td>
<td>29.75</td>
</tr>
<tr>
<td>Nr8</td>
<td>29.26</td>
</tr>
<tr>
<td>Nr9</td>
<td>28.79</td>
</tr>
<tr>
<td>Nr10</td>
<td>28.30</td>
</tr>
<tr>
<td>Nr12</td>
<td>27.32</td>
</tr>
<tr>
<td>Ave</td>
<td>27.29</td>
</tr>
</tbody>
</table>

10 dB SNR, all results have been improved, and the average recognition rate has been increased 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. The average recognition rate under lower SNR noise environments. The DRA normalization for robust speech feature extraction and recognition, International Congress on Image and Signal Processing, pp.1–4, Oct. 2009.


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 Hokkaido 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.