Noise robust speech recognition for voice driven wheelchair

Akira Sasou, Hiroaki Kojima
National Institute of Advanced Industrial Science and Technology (AIST)
{a-sasou,h.kojima}@aist.go.jp

Abstract
In this paper, we introduce a noise robust speech recognition system for a voice-driven wheelchair. Our system has adopted a microphone array system in order for the user not to need to wear a microphone. By mounting the microphone array system on the wheelchair, our system can easily distinguish the user’s utterances from other voices without using a speaker identification technique. We have also adopted a feature compensation technique. By combining the microphone array system and the feature compensation technique, our system can be applied to various noise environments. This is because the microphone array system can provide reliable information about voice activity detection to the feature compensation method, and the feature compensation method can compensate for the weak point of the microphone array system, which is that the microphone array system tends to be less effective for omni-directional noises.

Index Terms: noise robust speech recognition, wheelchair

1. Introduction
Various voice-driven wheelchairs have already been developed which try to make it possible for disabled people to move according to their own will [1]. However, conventional voice-driven wheelchairs still have some problems. For instance, conventional voice-driven wheelchairs have adopted a headset microphone which can record the user’s command voice in higher SNR even with the presence of surrounding noises. However, disabled people need to put this headset microphone on each time they use the voice-driven wheelchair. When the headset microphone moves away from the position of the mouth, they also need to be able to adjust the position of the headset microphone by themselves. Because these actions are not always easy for disabled people, we think headset microphones are not practical. Our goal is to develop a voice-driven wheelchair that does not need disabled people to wear any microphones on themselves.

2. Microphone array system
When considering surrounding-noise problems, we'd like to put microphones close to the user's mouth. However, such microphones would be dangerous for some disabled people, for instance, those with cerebral palsy with involuntary movements. So the microphone should be placed far enough from the user's mouth so as not to touch the user's head. In this case, however, speech recognition systems suffer from interference from surrounding noise and other voices. In order to overcome such noise problems, we have adopted a microphone array system instead of a headset microphone. Figure 1 shows the wheelchair we have developed. Our microphone array system consists of two circuit boards. Each circuit board has four silicon microphones soldered every 3cm linearly. The size of each circuit board is W130xD10xH5mm. The circuit boards were placed in the diagonal directions of the black square sponges. Because these black sponges are placed on the edges of the arm rests, the user's head never touches the microphone array system even when there are involuntary movements.

Figure 1: The developed wheelchair with microphone array system.

2.1. Detection of user utterance and noises
The speech recognition system should accept only the user's voice and should reject voices coming from outside. If we adopt a speaker identification technique for this purpose, we need to train the system every time a new user uses the wheelchair. This would not be practical. Instead, because our wheelchair has adopted a microphone array system, we can estimate the position or the arrival direction of sounds. That is, the mouth position of a seated user is always in a certain area, which is around the center of the seat at the sitting height. We call this as the user-utterance-area (UUA).

When an observed voice's position is estimated to be in the user-utterance-area, the speech recognition system accepts the voice. On the other hand, when an observed voice is judged to come from the outside, the speech recognition system rejects it. So, by adopting the microphone array system, we can easily distinguish the user's voice from others’ voices without any training procedures.

We have adopted the MUSIC [2] method for estimating the position and arrival direction of noises. We assume that a sound source occurring in the user utterance area is received as a spherical wave by the microphone array system. The steering vectors in the user utterance area are defined as follows.

$$P_q = [P_x, P_y, P_z]^T, q = 1, \ldots, 8$$
The number of sound sources is estimated from the number of eigenvalues larger than the threshold value. Instead of using the estimate of spatial correlation matrix, the estimated arrival directions of noises are given by \( \{ \theta_l, \phi_l \} \, l = 1, \ldots, K \). Instead of using the estimate of spatial correlation matrix \( \mathbf{R}(\omega) \), the modified MVBF uses the following virtual correlation matrix.

\[
\mathbf{V}(\omega) = \mathbf{a}(\omega, \mathbf{P}_\text{v}) \cdot \mathbf{a}^H(\omega, \mathbf{P}_\text{v}) + \mathbf{C}_\text{v}
\]

The third term \( \mathbf{C}_\text{v} \) is the correlation matrix of the virtual background noise, and the power of the virtual noise \( \sigma_c^2 \) can be arbitrarily chosen. By using this virtual correlation matrix, the coefficient vector of the spatial inverse filter becomes

\[
\mathbf{w}(\omega) = \mathbf{V}^{-1}(\omega) \mathbf{a}(\omega, \mathbf{P}_\text{v})
\]

The FFT of the emphasized speech signal is given by

\[
\mathbf{X}(\omega, n) = \mathbf{w}(\omega) \cdot \mathbf{Y}(\omega, n)
\]

### 3. HMM based feature compensation

The microphone array system is very effective at suppressing directional noise sources. However, it tends to be less effective for omni-directional ones. In order to make the speech recognition more robust to various noise environments, we also adopt the HMM based feature compensation method [4] in addition to the microphone array system.
There is another advantage of combining the microphone array system and the feature compensation method. The feature compensation method needs precise voice activity detection, as described below. Generally, however, it is not always easy to detect voice activity from a noise-corrupted speech signal of a single channel. Such poor accuracy of voice activity detection causes degradation of feature compensation accuracy. On the other hand, the microphone array system can detect voice activity even with the presence of surrounding noises. In our system, therefore, the feature compensation method can utilize the reliable voice activity detection from the microphone array system, which guarantees the feature compensation accuracy.

The feature compensation method assumes that a distortion of the noisy speech feature in the cepstral domain can be divided into a stationary and a non-stationary distortion component. The temporal trajectory of the non-stationary distortion component is assumed to be zero almost everywhere but temporarily changes. The stationary distortion component is absorbed by adding the estimated stationary distortion component to the expectation value of each Gaussian distribution in the output probability density functions (pdfs) of the clean speech’s HMMs. The degradation of feature-compensation accuracy caused by the non-stationary distortion component is compensated by evaluating each noise-adapted Gaussian distribution’s posterior probability multiplied by the forward path probability.

The noisy speech feature $X$ in the cepstral domain can be represented by $x = s + g(s, n)$ where $s$ represents the clean speech feature, $n$ represents the noise feature and $g(s, n)$ represents the distortion feature given by

$$g(s, n) = C \cdot \log \left[ 1 + \exp \left( C^{-1} \cdot (n - s) \right) \right]$$

(13)

where $C$ and $C^{-1}$ denote the DCT matrix and its inverse transform matrix, respectively. The first step of the feature-compensation process is to generate the copied Gaussian distributions of clean speech in an output pdf of each state. The output pdf of the $j$th state is represented by

$$b_j(s) = \sum_{n_j \in \text{States}} w_{n_j} N(s; \mu_{n_j}, V_{n_j}).$$

(14)

The second step is to evaluate the stationary distortion component $d_{n_j}$ for each copied Gaussian distribution, which is evaluated using the expectation value of each Gaussian distribution and the several noise-only frames prior to each utterance.

$$d_{n_j} = (1/N_{n_j}) \sum_{n \in \text{Noise-Only}} \log \left[ 1 + \exp \left( C^{-1} \cdot (n - \mu_{n_j}) \right) \right]$$

(15)

The third step is to adapt each copied Gaussian distribution of clean speech to the noisy speech by adding each evaluated stationary distortion component to the expectation value of each copied Gaussian distribution. In this noise adaptation process, we take into account only the expectation value of each Gaussian distribution. The diagonal covariance matrix in the noise-adapted Gaussian distribution is assumed to be the same as the covariance matrix of the clean speech. The noise-adapted output pdf is given by

$$b_j(x) = \sum_{n \in \text{Noise-Only}} w_{n_j} N(x; \mu_{n_j} + d_{n_j}, V_{n_j}).$$

(16)

The fourth step is to evaluate each noise-adapted Gaussian distribution’s posterior probability multiplied by the normalized forward path probability,

$$p_{n_j} = \frac{\alpha'(j, n-1) w_{n_j} N(x; \mu_{n_j} + d_{n_j}, V_{n_j})}{\sum_{j \in \text{States}} \sum_{n \in \text{Noise-Only}} \alpha'(j, n-1) w_{n_j} N(x; \mu_{n_j} + d_{n_j}, V_{n_j})}$$

(17)

where $\alpha'(j, n)$ denotes the normalized forward path probability which is given by

$$\alpha'(j, n-1) = \exp \left\{ \frac{\alpha(j, n-1)}{n} \right\}.$$  

(18)

The $\alpha(j, n)$ denotes the forward path probability that is obtained from the Viterbi decoding process. The fifth step is to evaluate the expectation value of the stationary distortion component weighted by the posterior probability of each noise-adapted Gaussian distribution.

$$\bar{d} = \sum_{j \in \text{States}} \sum_{n \in \text{Noise-Only}} p_{n_j} d_{n_j}$$

(19)

Finally, the compensated speech feature $\tilde{s}$ is obtained by subtracting the expectation value of the stationary distortion component from the noise-corrupted speech feature.

$$\tilde{s} = x - \bar{d}$$

(20)

The original Gaussian distributions of clean speech are used to evaluate the output probability of the compensated speech feature $b_j(\tilde{s})$ in the Viterbi decoding process.

### 4. System overview

Our system consists of one CPU board of a Pentium-M 2.0GHz, 8ch A/D converter, and DC-DC converter. These devices can be put in an aluminum case of size W30xH7xD18cm, which can be hidden in a space under the seat. The system devices that can be seen from the outside are just the microphone array system and the LCD showing the results of recognition.

The system embedded in the wheelchair must execute the following five functions.

1. Detection of user utterance and noises
2. Enhancement of user utterance
3. Speech feature compensation
4. Speech recognition
5. Wheelchair control

We have developed software that can execute these functions in a real time with the CPU board. The sampling rate of A/D converter was set to 8 kHz due to the limitation of processing capacity of the CPU board.

The wheelchair of Fig.1 has two motors, which drive the left and the right wheels independently. The motor controller is connected with the CPU board by an RS232C serial cable. The CPU board can indicate the rotation speeds of motors independently by way of the controller. So not only the speed the wheelchair moves but also the radius the wheelchair rotates can be easily controlled from the CPU board.

### 5. Experiments

In order to assess the noise robustness of the proposed method, we conducted experiments as follows. We recorded clean speech signals and environmental noises separately, and then mixed the digital signals of them together at six different kinds of SNR levels (20dB, 15dB, 10dB, 5dB, 0dB, -5dB), using a computer in order to generate noise-corrupted speech signals. The clean speech signals were recorded with the speakers sitting on the wheelchair, and uttering a command
voice in a silent room. The speakers are able-bodied, because the purpose of the experiments is assessment of the noise robustness, and consist of 29 females and 19 males. The speakers uttered 13 kinds of command-voices while looking forward, right and left in the user utterance area. In this experiment, we used five kinds of Japanese command voices out of the recorded ones, which are mae, migi, hidari, ushiro and teish. The environmental noises were recorded by actually moving the wheelchair around 13 kinds of places, which are 1. near a kindergarten, 2. a construction site near a train, 3. under train rails, 4. in front of an amusement arcade, 5. a building under construction, 6. a public office, 7. wind noise, 8. along a big street, 9. a road crossing, 10. a construction site, 11. a shop, 12. in front of a station and 13. in front of a ticket gate. The sound source localization and beam forming of the microphone array system are executed in every 125ms. The triphone acoustic models were trained from clean speech data that was obtained by downsampling the JNAS [5] data to 8 kHz. For comparison, we also evaluated baseline performance, in which we used speech data that were recorded by one microphone placed on the right hand microphone array, and closest to the user. The baseline performance was evaluated by recognizing each utterance that was segmented manually to include silence durations. On the other hand, the performance of the proposed method was evaluated without any segmentation information except the detection of voice activity of the microphone array system. From these results, we could confirm the effectiveness of the proposed method.

Table 3: Evaluated performances of the baseline and the proposed method.

<table>
<thead>
<tr>
<th></th>
<th>Baseline performance</th>
<th>Proposed system's performance</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Word Accuracy(%)</td>
<td>Word Accuracy(%)</td>
<td></td>
</tr>
<tr>
<td>near a kindergarten</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a construction site</td>
<td>59.63</td>
<td>61.76</td>
<td>60.93</td>
</tr>
<tr>
<td>Site near a train</td>
<td>60.07</td>
<td>61.64</td>
<td>60.37</td>
</tr>
<tr>
<td>under train rails</td>
<td>61.19</td>
<td>63.56</td>
<td>62.16</td>
</tr>
<tr>
<td>in front of an</td>
<td>63.94</td>
<td>66.36</td>
<td>65.61</td>
</tr>
<tr>
<td>amusement arcade</td>
<td>65.72</td>
<td>68.86</td>
<td>67.51</td>
</tr>
<tr>
<td>a building under</td>
<td>69.76</td>
<td>73.14</td>
<td>71.38</td>
</tr>
<tr>
<td>construction</td>
<td>72.67</td>
<td>76.10</td>
<td>74.34</td>
</tr>
<tr>
<td>a public office</td>
<td>76.56</td>
<td>79.93</td>
<td>78.72</td>
</tr>
<tr>
<td>wind noise</td>
<td>79.84</td>
<td>83.89</td>
<td>81.82</td>
</tr>
<tr>
<td>along a big street</td>
<td>83.15</td>
<td>86.96</td>
<td>85.00</td>
</tr>
<tr>
<td>a road crossing</td>
<td>86.32</td>
<td>90.44</td>
<td>88.88</td>
</tr>
<tr>
<td>a construction site</td>
<td>90.11</td>
<td>94.18</td>
<td>92.14</td>
</tr>
<tr>
<td>a shop</td>
<td>94.47</td>
<td>98.55</td>
<td>96.30</td>
</tr>
<tr>
<td>in front of a</td>
<td>98.28</td>
<td>99.09</td>
<td>98.68</td>
</tr>
<tr>
<td>station</td>
<td>99.00</td>
<td>99.75</td>
<td>99.75</td>
</tr>
<tr>
<td>in front of a ticket</td>
<td>99.00</td>
<td>99.44</td>
<td>99.00</td>
</tr>
<tr>
<td>gate</td>
<td>99.00</td>
<td>99.88</td>
<td>99.44</td>
</tr>
<tr>
<td>Avg</td>
<td>99.00</td>
<td>99.88</td>
<td>99.44</td>
</tr>
</tbody>
</table>

6. Conclusions

This research has been conducted as part of “Development of technologies for supporting safe and comfortable lifestyles for persons with disabilities,” funded by the solution oriented research for science and technology (SORST) program of the Japan Science and Technology Agency (JST), Ministry of Education, Culture, Sports, Science and Technology (MEXT) of the Japanese Government.

7. Acknowledgements

8. References