Internal Noise Suppression for Speech Recognition by Small Robots

Akinori Ito, Takashi Kanayama, Motoyuki Suzuki and Shozo Makino

Graduate School of Engineering
Tohoku University, Sendai, Japan
{aito,gmount,moto,makino}@makino.ecei.tohoku.ac.jp

Abstract

Speech recognition by a small robot is difficult because the robot makes noise itself. In this paper, two new methods are proposed that suppresses internal noise of the small robots. These methods are based on spectral subtraction (SS). The difference of the proposed methods from the original SS is that the proposed methods use the estimated noise spectrum dependent on the motion of the robot. One method, called MDSS, prepares the noise spectrums for all motions. Another method, called NPSS, predicts the noise spectrum from angular velocities of all joints of the robot using a neural network. From the results of the comparison between the original SS and the proposed methods, the proposed methods outperformed the conventional SS. The NPSS worked well even when the noise of the motion was unstable, while the MDSS method gave good result when the noise in one motion was stable.

1. Introduction

Small robots for entertainment use have been more and more popular nowadays[1, 2]. Spoken dialog is the most natural way for interacting with them. However, speech recognition using a small robot is very difficult. One big reason is that a robot makes noise by moving its motors. If the size of the robot is sufficiently large, the problem can be avoided by placing microphones away from the noise sources or by using the array microphone technology[3]. However, this solution cannot be applied to small-sized robots like the Aibo[4]. Conventional noise suppression techniques such as spectral subtraction (SS)[5] can be used to improve the quality of the input speech, but SS cannot suppress the internal noise completely because the internal noise is unstable.

In this paper, we propose a new method to suppress the internal noise made by the robot. The essential point is to predict the spectrum of the internal noise from the status of the motors in the robot. We used a neural network to predict the noise spectrum from angular velocities of all joints in the robot. We carried out a speech recognition experiment to prove the effectiveness of the proposed method.

2. The robot and internal noise

We used the AIBO ERS-210 as a platform of the research. This robot has two microphones at the both side of the head. These microphones can be used as omni-directional or unidirectional microphones. We used only one microphone at the right side of the head as an omni-directional microphone. The ERS-210 has 15 motors as well as sensors to measure the angles of the motors.

As an ERS-210 can communicate with a PC via wireless network (IEEE802.11b), we recorded the speech using an external PC through the wireless network.

The ERS-210 makes two kinds of noises: electrical noise and mechanical noise. The electrical noise is derived from a power supply or other electrical circuits and is observed as a static noise. This kind of noise is observed even when no motor is loaded. Another one, the mechanical noise, is mainly generated by rotation of the motors, as well as creaking of the joints and contact of body parts. The mechanical noise is unstable and depends on the motion of the robot. Figure 1 shows an example of speech recorded by the ERS-210 while moving its head and legs. This example shows the dependency of the noise on the motion. As the motion of the robot is independent of the input speech, the internal noise may change during an utterance.

The signal-to-noise ratio (SN ratio) of the input speech depends on the distance between a speaker and the robot. To obtain rough estimation of the SN ratio, we carried out an experiment to investigate the relationship between the distance and SN ratio. The conditions of the measurement were shown in Table 1. The measurement is done in a silent room and the speakers were asked to talk to the robot from each distance. Table 2 shows the result. The speakers tended to utter loudly when the distance between the speaker and the robot was great. However, as Table 2 shows, the SN ratio decreased as the distance becomes greater.
3. Motion-dependent spectral subtraction

The spectral subtraction[5] is a technique to suppress noise in the input signal by subtracting the spectrum of the noise from the spectrum of the input signal. This method assumes that the spectrum of the noise does not change during speech input. The spectrum of the noise is often estimated from the beginning part of the input signal where the speech signal does not exist.

Let \( \tilde{S}(f,t) \) and \( N(f,t) \) be the spectrum of speech and noise at time \( t \), respectively. Then the spectrum of the observed signal \( Y(f,t) \) is

\[
|Y(f,t)| = |\tilde{S}(f,t)| + |N(f,t)|.
\]

Assuming that the noise \( N(f,t) \) is independent of time \( t \), the spectrum \( \bar{N}(f) \) can be estimated from the silent part of the input signal. Then the spectrum of the speech can be estimated as

\[
\tilde{S}(f,t) = |Y(f,t)| - |\bar{N}(f)|.
\]

If the assumption of stability of the noise does not apply, the estimated spectrum \( \tilde{S}(f,t) \) eventually become negative. To avoid this, the following method is often used[6].

\[
|\tilde{S}(f,t)|^2 = \max \left\{ |Y(f,t)|^2 - \alpha |\bar{N}(f)|^2, \beta |\bar{N}(f)|^2 \right\}
\]

where \( \alpha \geq 1 \) and \( 0 < \beta < 1 \).

The main problem of applying SS to the speech recognition by the small robots is that the noise changes when the motion of the robots changes. The motion of a robot often changes during speech. If this happens, \( \bar{N}(f) \) cannot be used as an estimation of the noise anymore.

To solve this problem, we first make three assumptions: (1) the kind of motions, its starting time and duration are known to the system, (2) the spectrum of the noise for a motion is stable during the motion, and (3) noise signals for all motions can be obtained beforehand. Under these assumptions, we can prepare the noise spectrum \( \bar{N}_m(f) \) for a motion \( m \). Then we can suppress the internal noise by performing the following process.

\[
|\tilde{S}(f,t)|^2 = \max \left\{ |Y(f,t)|^2 - \alpha |\bar{N}_m(f)|^2, \beta |\bar{N}_m(f)|^2 \right\}
\]

where \( m(t) \) denotes the motion at time \( t \). We call this method ‘a motion-dependent spectral subtraction (MDSS).’

To investigate the performance of MDSS, we carried out word recognition experiments. The conditions of the experiments are shown in Table 3. The speech with internal noise was prepared by superimposing the internal noise over the clean speech. We used three kinds of motions: pan, tilt and walk. ‘Pan’ was to move the head horizontally, ‘tilt’ was to move the head vertically, and ‘walk’ was to walk using four legs. The boundaries of the motions were given manually.

In this experiment, we compared three kinds of noise suppression techniques. ‘Sub1’ method is the conventional SS method that estimates the noise spectrum using the beginning part (50ms) of the input signal. ‘Sub2’ method is a kind of SS that uses the average spectrum of noises of all motions as the noise spectrum to be subtracted. ‘MDSS’ is the proposed method. In addition to these methods, we also observed ‘normal’ condition in which no noise suppression was employed.

The recognition results for each motion is shown in Figure 2 when SN ratio was 5 dB. In this figure, ‘reference’ means the recognition result for the clean speech. From this result, it was found that the proposed method (MDSS) gave the best performance for the ‘pan’ and ‘tilt’ noise. However, all noise suppression method including MDSS did not work for the ‘walk’ noise. The reason is that the walk noise is unstable in the motion.

Next, we investigated the performance of the proposed method for various SN ratios. In this experiment, ‘pan’ and ‘tilt’ noises were used. Figure 3 shows the result. This result shows that the proposed method gave the best performance for all SN ratio. An especially big improvement was obtained for lower SN ratio (\(-5 \sim 5 \) dB).

4. Noise prediction using neural network

According the previous experiment, the motion-dependent spectral subtraction was partly successful, but it did not work for the ‘walk’ noise. The reason of the failure is that the assumption (2) did not hold, i.e., the ‘walk’ noise was not stable.
To suppress unstable noise, we must predict the spectrum of the noise frame-by-frame.

To achieve this, we constructed a neural network that predicts the spectrum of the internal noise from the status of joints. As the ERS-210 has sensors that measure the angles of all joints[8], we used the sensor outputs to estimate the noise spectrum. Figure 4 shows the joints considered in this experiment. A sensor gives the angle of the joint in $\mu$rad at every 32 ms. We converted the sensor data into angle velocities by calculating the difference of the two contiguous sensor outputs.

Figure 5 shows the neural network for noise estimation. There are 75 units in the input layer and temporally five contiguous sets of the angle velocities are given to the units. There are two hidden layers in the network, and the output layer outputs 24 filterbank coefficients and a power of the frame.

First, we investigated the optimum number of units in the hidden layers. We used the same number of units in the two layers, and observed the mean square error (MSE) for various networks with a different number of hidden layers. The training data was 330 seconds of noise data and corresponding sensor data consisted of 80 kinds of motions. The evaluation data was 60 seconds of noise data that was not used for training of the network. Figure 6 shows MSE vs. learning epochs for networks that have 30–200 units in one hidden layer. This result shows that increasing the number of hidden units does not improve the performance of prediction, which seems to be caused by overfitting to the training data. From this result, we chose the number of units in a hidden layer to be 30 and the number of epochs to be 10,000.

5. Neural prediction based spectral subtraction

Next, the spectral subtraction technique is applied using the predicted noise signal. Here, as the predicted signal is filterbank coefficients, we have to calculate power spectrum from the coefficients.

The filterbank analysis is performed as follows[9]. First, let $S(f)$ be the spectrum of the signal and $W_i(f)$ be the window function of $l$-th channel. The triangular window $W_i(f)$ is calculated as

$$W_i(f) = \begin{cases} \frac{f - f_{lo}(l)}{f_{hi}(l) - f_{lo}(l)}, & \text{if } f_{lo}(l) \leq f \leq f_c(l) \\ \frac{f_c(l) - f}{f_{hi}(l) - f_c(l)}, & \text{if } f_c(l) \leq f \leq f_{hi}(l) \\ 0, & \text{otherwise} \end{cases}$$

(5)

where $f_{lo}(l), f_c(l), f_{hi}(l)$ are the lower bound frequency, center frequency and upper bound frequency of the $l$-th window function, respectively. Then the $l$-th coefficient is calculated as

$$m(l) = \sum_{f = f_{lo}(l)}^{f_{hi}(l)} W_i(f) |S(f)|.$$  

(6)

To estimate the noise spectrum from the filterbank coefficients, we make an assumption that the noise spectrum is uniform within a channel. Let $\bar{N}_l$ be the magnitude of noise spectrum (that is uniform from the assumption). Then

$$\bar{N}_l = \frac{m(l)}{\sum_{f = f_{lo}(l)}^{f_{hi}(l)} W_i(f)}.$$  

(7)
Table 4: Experimental conditions for NPSS

<table>
<thead>
<tr>
<th>Frame shift SN ratio</th>
<th>Motions</th>
</tr>
</thead>
<tbody>
<tr>
<td>8ms</td>
<td>(1) move the head horizontally</td>
</tr>
<tr>
<td>−5 ~ 10dB</td>
<td>(2) move the head horizontally changing its speed</td>
</tr>
<tr>
<td></td>
<td>(3) wave the front left leg</td>
</tr>
<tr>
<td></td>
<td>(4) move the front legs up and down</td>
</tr>
<tr>
<td></td>
<td>(5) move the head and all legs randomly</td>
</tr>
<tr>
<td></td>
<td>(6) tilt the head</td>
</tr>
<tr>
<td></td>
<td>(7) walk on four legs</td>
</tr>
</tbody>
</table>

and the noise-subtracted filterbank coefficient of the \( l \)-th channel is calculated as

\[
| \tilde{S}_l(f) |^2 = \max \left\{ \frac{|Y(f)|^2 - \alpha \bar{N}_l^2}{\beta \bar{N}_l^2} \right\}
\]

and

\[
\tilde{m}(l) = \sum_{f=\text{Lo}(l)}^{\text{Hi}(l)} W_l(f)|\tilde{S}_l(f)|.
\]

We call the proposed method "a neural prediction based spectral subtraction (NPSS)."

We carried out experiments to compare the proposed method to other methods. The experimental conditions are listed in Table 4. The conditions not listed in Table 4 are same as that in Table 3.

Figure 7 shows the word recognition rate for each methods for SN ratio of -5, 0, 5 and 10dB. This result proves that NPSS gave the best result, which is about 10 point higher than the best of the other methods.

The recognition results for each motion are shown in Figure 8. For a stable noise (1), the NPSS gave almost same performance as the MDSS. For unstable noises (2–6), the NPSS gave better results than the MDSS, which proved that the noise prediction using a neural network worked well. However, as the result for (6) indicates, it did not improve the performance for the walking noise. The reason why the NPSS did not work for the walking noise seems that the walking noise contains signals derived from contacts between body parts and the floor.

6. Conclusion

Two new methods are described to suppress internal noise of small robots. One method is MDSS that uses motion-dependent noise spectrum for spectral subtraction. The other method is NPSS that predicts spectrum of noise from angular velocities of joints of the robot. A neural network is used for the prediction. From the experimental result, the MDSS gave good results when the noise in one motion is stable, and the NPSS worked better even when the noise was unstable.

7. References