Data Collection and Evaluation of Speech Recognition for Motorbike Riders


Graduate School of Information Science, Nagoya University

Abstract
Speech recognition should be as an eyes-free and hands-free interface. To realise this technology, we need to clarify acoustics in a helmet and determine how much high-level riding noise affects captured speech data. This paper describes the acoustics in a helmet and transfer functions of the microphone position. We constructed a data collection system and collected the speech data of motorbike riders on city roads and express highways. Speech recognition experiments were conducted and we obtained a recognition rate high of 83.1%.

1. Introduction
Riding a motorbike requires more care than driving a car. Even when a rider idles his/her motorbike, such as when he/she waits at a red light, button operations are inconvenient because the rider needs to remove his/her gloves to push the buttons. Therefore, for motorbike riders, an eyes-free and hands-free interface is required for operating information appliances, such as a cellular phone and a route navigation system. Thus, speech recognition is a very important technology.

This study investigated the feasibility of speech recognition for motorbike riders. On a motorbike, riders are exposed directly to high-level noises such as wind noise, engine noise, and road noise. It is known that exposed noise level is varied by various factors such as speed, riding position, and helmets [1, 2]. In order to realise speech recognition on a motorbike, we first need to investigate how much such factors degrade conventional speech recognition performance.

Riders must put on a helmet when they ride motorbikes. Helmets are designed to reduce noise level, however, we need to clarify how much this reduction contributes to speech recognition using microphones inside the side of helmets. Moreover, we need to investigate acoustics in a helmet, because a helmet has a very small cavity.

In this paper, we measured acoustics in a helmet to determine microphone positions for collecting riders’ speech corpus. We then collected the speech data uttered by motorbike riders riding on a highway. We also provide an analysis of the corpus and the results of our speech recognition evaluation.

2. Acoustics in a Helmet
Acoustic features inside a helmet have unique characteristics because a helmet has only a small cavity when a rider puts it on. Since a microphone must be located close to the mouth, directivity of a sound source, that is, the mouth, influences significantly the acoustic feature between the sound source and the microphone. We determined an appropriate microphone position for riders’ speech recognition by measuring the acoustic transfer function inside a helmet.

Figure 1 shows ten positions for installing a small microphone. Acoustic transfer functions between an artificial mouth of a head-and-torso simulator (B&K 4128) and every microphone were measured by using a Swept Sine signal [3]. A full-face helmet ARAI RAPIDE-OR was used.

Figure 2 shows the magnitude responses. The following seven positions are inappropriate when riders put on the helmet; the #1, #2, and #3 positions have no space to attach a microphone, the microphone sank into the buffer material for positions #6, #9, and #10, and the #7 microphone was blown strongly by the riders’ breath.

All the remaining responses (#4, #5, and #8) differ due to reflection of proximity between the mouth and the microphone. Two sharp dips in #5 show that there are small cavities in a helmet; however, the responses fluctuated within −20 dB in the speech band. These responses were not good, but were adequate for speech recognition.

Our recording system for the motorbike has only two input channels. We chose #4 and #8 for data collection, because they had better responses that were flatter and had no dips in the speech band. We refer to the #8 microphone as the mouth microphone and the #4 as the nose microphone.
3. Speech Corpus of Motorbike Riders

3.1. Collection System

Figure 3 depicts the outline of our speech collection system. This system is designed to collect the repetition speech. The subjects repeat a word, a short phrase, or a short sentence, as instructed by the operator using a cellular phone. Speech is collected by both the mouth microphone and the nose microphone simultaneously.

The Yamaha XJR-400, a motorbike of 400 cc displacement, was used for data collection (Figure 4). It has no wind screen. We prepared three sizes of the helmet to fit the size of the subject heads.

3.2. Corpus

We collected 50 subjects’ (49 males and one female) speech. Riding experiences ranged from five months to 25 years. Text for speaking included 50 Japanese phonetic balanced sentences excerpted from the ATR corpus, 22 Japanese words, and 5 Japanese short sentences. The phonetic balanced sentences were used as the training set. When data was collected, they were divided into short phrases to memorize easily in repetition.

The other phrases were designed for evaluation, and included device controlling commands (e.g., “cancel”) and information retrieving commands (e.g., “I’d like to listen Billy Joel’s ‘Just the Way You Are’”). In each session, the subject uttered the training set twice and the evaluation set twice. The corpus was constructed by checking all the collected data and discarding faulty utterances.

Data collecting consisted of two sessions; a city road session and express highway session. In each session, the same text set was used. In the city road session, the subject rode on a city road at a speed ranging from 0 kph to 60 kph, often stopping at red signals. On the other hand, in the express highway session, the subject rode on an express highway at a speed ranging under 100 kph. In this session, the motorbike did not stop except when stuck in traffic.

Table 1 shows the statistics of the constructed corpus.
This corpus was classified by microphone and riding condition into four categories.

Table 1: Statistics of the Corpus

<table>
<thead>
<tr>
<th>Riding condition</th>
<th>city road</th>
<th>express highway</th>
</tr>
</thead>
<tbody>
<tr>
<td>balanced sentences</td>
<td>35274 (20.4h)</td>
<td>22878 (12.9h)</td>
</tr>
<tr>
<td>words</td>
<td>6152 (2.8h)</td>
<td>5942 (2.6h)</td>
</tr>
<tr>
<td>total</td>
<td>41426 (23.2h)</td>
<td>28820 (15.5h)</td>
</tr>
</tbody>
</table>

3.3. Noise of the Speech Data

We examined the influence of noise, which includes wind, engine, road, and other noise. For analysis, we estimated the SNR (Signal-to-Noise Ratio) of each utterance. For SNR estimation, we assume the log-power distribution of the utterance as a two-mixture Gaussian; lower is noise and higher is speech. Then, the SNR of the utterance is calculated as the difference of their averages. Gaussians are estimated using an EM (Expectation Maximization) algorithm[4].

Table 2 shows the statistics of the average SNR of the four types of speech. The highest average SNR was 21.8 dB, achieved for a city road session captured by the nose microphone, and the lowest average SNR was 11.0 dB, obtained for an express highway session captured by the mouth microphone. In both sessions, the average SNRs of the data captured by the mouth microphone were lower, because the noise was caused by turbulence at the gap between the helmet and neck[1, 2].

Table 2: Statistics of the SNR [dB]

<table>
<thead>
<tr>
<th></th>
<th>city road</th>
<th>express highway</th>
</tr>
</thead>
<tbody>
<tr>
<td>nose</td>
<td>21.8</td>
<td>15.1</td>
</tr>
<tr>
<td>mouth</td>
<td>20.0</td>
<td>11.0</td>
</tr>
</tbody>
</table>

Figures 5 and 6 show the distributions of the utterances’ SNR for the city road session and express highway session, respectively. The range of all the distributions was broad, because both the power of the riders’ voice and the speed of the motorbike ranged widely. In Figure 5, although there was only one peak in the distribution of the nose microphone, there are two peaks in the distribution of the mouth microphone, because the mouth microphone was more influenced by wind noise than the nose microphone. The lower distribution mainly consisted of the data uttered during riding and the higher mainly consisted of the data uttered during stopping. In the express highway session, there is only single distribution for the data captured by the mouth microphone, because the motorbike rarely stopped.

By human checking of the lower SNR samples, we confirmed that most noise was categorized into riders’ breathing and wind noise, and there were only a few due to engine, road, and the other noise from the motorbike.

4. Speech Recognition on Motorbike

We conducted speech recognition evaluation to confirm that the collected data had enough quality for the training of acoustic models for speech recognition.

4.1. Acoustical Preprocess

To avoid the effect of noise, cutting off the lower frequency is effective. Searching for the most effective cutoff frequency, we tried the seven low frequency cutoffs at every 50 Hz from 0 to 300 Hz. (Figure 7, other condition is described in Section 4.2). The best result was achieved at 250 Hz. Although the original signal is sampled at 16 KHz, the bandwidth was limited to the range from 250 Hz to 8000 Hz.
Fixed to 12 MFCC (Mel frequency cepstral coefficients), for the HMM (hidden Markov model) acoustic model were average of the experiments. Experiment sets, and the results were calculated as the average of the experiments.

Throughout the recognition experiments, the features for the HMM (hidden Markov model) acoustic model were fixed to 12 MFCC (Mel frequency cepstral coefficients), 12 ΔMFCC, and Δlog power. MFCC is normalized using CMN (cepstral mean normalization). The basic structure of the HMM is also fixed as three-state continuous density triphones that share 500 states with 32 Gaussian mixture components. All triphones have a simple left-to-right topology except for a short pause that has a transition from the start state to the final state. HTK (hidden Markov model toolkit) [5] was used as the decoder. The evaluation task was word recognition with 1000 word vocabulary.

The baseline models were trained using the matched condition data set, for example, the mouth microphone city roads model was trained from the data captured from the mouth microphone in the city road session. The MVN (Mean and Variance Normalization) models were trained from the data that was processed by additional variance normalization[6]. Then, the merged models were trained from the data of both sessions.

Figures 8 and 9 show the evaluation results of the mouth microphone and the nose microphone, respectively.

In each session, the merged model of the mouth microphone marked the highest recognition rate though its average SNR was lower than the nose microphone. The highest recognition rates were 83.1% in the city road session and 77.6% in the express highway session. At that time, a 25.6% error reduction was achieved in the city road session and 21.4% was achieved in the express highway session.

5. Conclusion

Exploring the feasibility of speech recognition for motorbike riders, we analyzed the acoustics in a helmet, and then confirmed that microphone position inside the helmet affects transfer characteristics. We collected the speech data of motorbike riders who rode on city roads and express highways, and then analyzed acoustics characteristics of the data. Finally, we conducted a speech recognition experiment, and obtained an 83.1% recognition rate for the city road session and 77.6% for the express highway session. Throughout the experiment, the microphone position inside of the helmet affected not only SNR of the data but the speech recognition performance as well.

6. References