Analysis of Drivers’ Speech in a Car Environment

Tomoyuki Kato, Jun Okamoto, and Makoto Shozakai

Information Technology Laboratory, Asahi Kasei Corporation
Atsugi AXT Main Tower 22F, Okata 3050, Atsugi, Kanagawa, 243-0021, Japan
{kato.tr, okamoto.jd, shozakai.mb}@om.asahi-kasei.co.jp

Abstract

In order to accelerate the promotion of speech recognition systems to the public, understanding characteristics of speech in real environments is one of the most important issues. This paper reports variations of speech characteristics in a car environment. To analyze speech characteristics in the specific environment, a corpus, recorded carefully in terms of equality of utterances and conditions for whole set of speakers, is necessary. We created a new corpus named ‘Drivers’ Japanese Speech Corpus in a Car Environment (DJS-C)’ composed of utterances of words useful for the operation of in-vehicle information appliances. Analysis of the DJS-C corpus shows that differences in speech characteristics are diverse among drivers and change with driving conditions. Quantitative analysis and speech recognition experiments show that performance degrades due to Distance between Phonemes, Uniqueness of speech, and speech recognition experiments show that performance degrades due to Distance between Phonemes, Uniqueness of Speaker’s Voice, and SNNR.

Index Terms: speech recognition, speech corpus, car environment, qualitative analysis

1. Introduction

Dramatic improvements in automatic speech recognition (ASR) techniques have enabled ASR systems to be used practically in real environments. There are ASR applications in many fields, particularly in-vehicle information appliances. Currently, the usage rate of ASR is not high, so performance in actual environments has seen little change for study. Using ASR in a car environment is one of the most important and promising scenarios. Although various noise and stress given to driver in a car environment may cause speech to sound different from that in a silent and stress-free environment, there is little knowledge of the integrated effect of speech characteristics in a car environment. Additionally, although the performance of ASR systems tends to be highly dependent on diverse factors such as physical characteristics of individual speakers, task difficulties, and ambient noise, there is no corpus available to analyze these factors specifically in a car environment. Therefore in this paper, our new large-scale corpus - recorded in both stationary and moving vehicles - is described. This corpus is designed to analyze the real speech characteristics in a car environment, and to analyze task dependency specific to in-vehicle information appliances. This paper reports variations of speech characteristics in the car driving situation and correlates each speaker’s recognition performance and his/her speech characteristics.

In the next section, our corpus is introduced. In Section 3, analysis of speech characteristics is presented. In Section 4, correlations between recognition performance and speech characteristics are discussed. Finally, conclusion of this paper is presented in Section 5.

2. Speech corpus

In this section, an overview of the corpus developed for this study is described.

2.1. Overview of Drivers’ Japanese Speech Corpus in a Car Environment (DJS-C)

To elucidate the variation of recognition performance by noises, by speakers, and by vocabularies, a corpus by which variations of noises and speakers can be analyzed is necessary. Although there are some corpora recorded in running vehicles, for example CIAIR [1] in Japan, there is no corpus which is suitable to analyze these variations due to incomplete control of utterances and driving conditions.

To analyze the real speech characteristics objectively, we compiled the DJS-C, a corpus of 260 speakers who were recorded while sitting in the driver’s seat of a running vehicle. The recording environment and utterance list remained unchanged for the duration of recording. A summary of the corpus is shown in Table 1.

2.2. Definition of driving actions

Three driving conditions are defined in DJS-C: idling car (Idle), running at urban street (City): 0–60 km/h, running at express highway (High): about 100km/h. For safety, subjects were divided into two classes, General and Professional. Recording with General subjects was limited to Idle and City only. Subject categories are shown in Table 2.
2.3. Categories of recording utterances

Recording utterances are designed, for the operation of in-vehicle information appliances. DJS-C consists of eight categories:

1. Hands-Free Command
2. Car Navigation Command
3. Voice-Tag
4. Continuous Digits
5. Category of Point of Interest
6. Point of Interest
7. Address
8. Name of Artist

2.4. Recording parameters

Recording parameters are shown in Table 3. The high-pass filter was used to prevent clipping due to amplitude overflow generated by uneven road surface.

2.5. Corpus Size

Corpus sizes are shown in Table 4. Total size represents hours of audio recorded.

3. Analysis of speech characteristics

In order to clarify acoustic differences between each utterance, each speaker, and each driving condition, physical parameters are measured for each utterance. Utterances recorded at the driver side’s microphone are used for this analysis. Each utterance was down-sampled to 16 kHz. To measure model-based physical parameters, acoustic models were created. The whole set of Japanese phonemes (IPA) are listed in Table 5. In this paper, an acoustic model structure is a mono-phone HMM expressed by a single Gaussian distribution. The number of monophones is 43 - whole set of Japanese phonemes and additional silence model - and each HMM has a left-to-right topology with 3 states. The acoustic feature vectors consist of 12 MFCCs, 12 delta-MFCCs, and 1 delta-log power.

### Table 3: Recording Parameters

<table>
<thead>
<tr>
<th>Place of Microphones</th>
<th>Number of Channel</th>
<th>Driving condition</th>
<th>Total size [h]</th>
</tr>
</thead>
<tbody>
<tr>
<td>map lamp (2), visor, steering wheel shaft</td>
<td>4 channels</td>
<td>High</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>City</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Idle</td>
<td>40</td>
</tr>
</tbody>
</table>

### Table 4: Corpus Size

<table>
<thead>
<tr>
<th>Driving condition</th>
<th>Speakers</th>
<th>Utterances per speaker</th>
<th>Total size [h]</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>20 males</td>
<td>270</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>20 females</td>
<td>270</td>
<td>6</td>
</tr>
<tr>
<td>City</td>
<td>130 males</td>
<td>270</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>130 females</td>
<td>270</td>
<td>40</td>
</tr>
<tr>
<td>Idle</td>
<td>130 males</td>
<td>300</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>130 females</td>
<td>300</td>
<td>40</td>
</tr>
</tbody>
</table>

### Table 5: Japanese Phonemes (IPA)

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Consonant</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, i, u, e, o, a:, i:, u:, e:, o:</td>
<td>N, w, y, j, my, ky, by, gy ny, by, ry, py, p, t, ts, ch, b, d, dy, g, z, m, n, s, sh, h, f, r, q</td>
</tr>
</tbody>
</table>

### Table 6: Calculation Methods

<table>
<thead>
<tr>
<th>Speech Power</th>
<th>Maximum average power in every 10 ms frame with spectral subtraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNNR</td>
<td>The power ratio of signal (speech plus noise) and noise</td>
</tr>
<tr>
<td>Speaking Speed</td>
<td>Number of Moras in one second</td>
</tr>
<tr>
<td>Distance between Phonemes</td>
<td>Distance between 5 vowels in speaker-dependent acoustic model</td>
</tr>
<tr>
<td>Uniqueness of Speaker’s Voice</td>
<td>Distance between speaker-dependent acoustic model and speaker-independent acoustic model</td>
</tr>
</tbody>
</table>

3.1. Methods for physical parameter calculation

Calculation methods of physical parameters are shown in Table 6. Details of each method are described in this section.

3.1.1. Speech Power

Speech Power is calculated as the maximum power of every 10 ms frame average. To reduce the effects of noise, Speech Power is calculated after noise suppression by spectral subtraction.

3.1.2. SNNR

Speech in this corpus is recorded in a real environment, and so it is impossible to calculate SNR (Signal to Noise Ratio), because signal is a mixture of both speech and noise. Thus, SNNR (Speech-plus-Noise to Noise Ratio) is calculated instead. Signal is pre-emphasized before being calculated by following formula:

\[
s_n = s_n - ks_{n-1}
\]

where \( k \) is the pre-emphasis coefficient of 0.97.

3.1.3. Speaking Speed

Speaking Speed is measured as the number of Moras spoken per second.

3.1.4. Distance between Phonemes

This is a model-based parameter, and it expresses the readiness of speech of a speaker. In this paper, the mutual distance between each phoneme is measured by using the Bhat-charyya distance \([2, 3]\). The mutual distance \( D(i, j) \) between phoneme \( i \) and phoneme \( j \) is defined by the following formula:

\[
D(i, j) = \frac{1}{S} \sum_{s=1}^{S-1} \sum_{l=1}^{L-1} d(i, j, s, l)
\]

where \( S \) represents the number of states in the acoustic model of phoneme \( i \) and \( j \). \( L \) stands for the number of dimensions of the acoustic feature parameters. \( d(i, j, s, l) \) is the distance between the \( l \)-th dimension’s Gaussian distribution of the \( s \)-th state in phoneme \( i \) and \( j \). Since this corpus contains various environmental noises brought by running car, distances were calculated only for 5 vowels (a, i, u, e, o).
3.1.5. Uniqueness of Speaker’s Voice

This is also a model-based parameter, and it expresses difference of speech characteristics between a specific speaker and general speakers. The Uniqueness of Speaker’s Voice is measured as the distance between speaker-dependent (SD) acoustic model and speaker-independent (SI) acoustic model. The distance \( D(i, j) \) between acoustic models \( i \) and \( j \) is defined by the following:

\[
D(i, j) = \sum_{k=1}^{K} d(i, j, k)w(k) / \sum_{k=1}^{K} w(k) \tag{3}
\]

where \( d(i, j, k) \) denotes mutual distance between acoustic unit \( k \) in acoustic model \( i \) and acoustic unit \( k \) in acoustic model \( j \). \( w(k) \) represents occurrence frequency for acoustic unit \( k \). \( K \) is the total number of acoustic units.

3.2. Results of analysis

3.2.1. General Speakers

Analysis results of General Speakers are shown in Table 7. In Idle, differences in Speech Power and SNNR between females and males are larger than those in City. This suggests the Lombard Effect [4, 5, 6] appeared in City, due to being noisier and thus having a lower SNNR. Standard deviation in the Speech Power of females is large and it can be expected that differences in lung-power among speakers affect this. Irrespective of conditions, Speaking Speed of females is slower than that of males. Yet, gender is not a factor for Distance between Phonemes and Uniqueness of Speaker’s Voice. In City, because acoustic model parameters are impacted by noise, Distance between Phonemes and Uniqueness of Speaker’s Voice become smaller.

3.2.2. Professional Speakers

Analysis results of Professional Speakers are shown in Table 8. The results in Idle and City are similar to those of General Speakers. In High, SNNR is considerably lower than other conditions.

4. Correlation between recognition performance and physical parameters

4.1. Experimental conditions

Table 9 shows experimental conditions. In this section, the High running condition is discussed. Acoustic models for recognition are trained based on leave-one-out cross validation, that is, except evaluation subjects, all speech data is used as training data. A language model is a simple word network grammar. The word recognition accuracies of all speakers are shown in Figures 1 and 2.

4.2. Analysis of low recognition performance speakers

Figure 3 shows histograms of physical parameters in males, where green bars are histograms of the speaker encircled in Figure 1, whose recognition performance is the lowest in males. These histograms show 3 aspects of this low recognition performance speaker: Speaking Speed is fast, Distance between Phonemes is short, and Uniqueness of Speaker’s Voice is relatively high.

Figure 4 shows histograms of physical parameters in females, where green bars are histograms of the speaker encircled in Figure 2, whose recognition performance is the lowest in females. These histograms show 3 aspects of this low recognition performance speaker: Distance between Phonemes is very
along with car speed becoming faster. From analysis of the speakers’ speech. For instance, Speech Power becomes louder especially in females. The changes of speech characteristics by stance, variance of Speech Power by speaker is very large, especially shown by analysis of performance. The variances of the speech characteristics are statistically significant for females more than in males.

This paper explored analysis of the speech characteristics in a running car and the observations made here will empower researchers to improve recognition performance in this unique application of ASR systems.

6. Acknowledgments

This research is supported by the METI: “Development of Fundamental Speech Recognition Technology,” and is supervised by Prof. Furui at Tokyo Institute of Technology and Prof. Kobayashi at Waseda University.

7. References