A COMPARATIVE STUDY ON ACOUSTIC AND LINGUISTIC CHARACTERISTICS USING SPEECH FROM HUMAN-TO-HUMAN AND HUMAN-TO-MACHINE CONVERSATIONS

Toshiyuki Takezawa, Fumioi Sugaya, Masaki Naito †, and Seiichi Yamanoto
ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
E-mail: toshiyuki.takezawa@lt.atr.co.jp
† currently with KDD R&D Laboratories, Inc.

ABSTRACT

Speech translation and dialogue systems must accept conversational speech. In this paper, we discuss acoustic and linguistic characteristics based on results of speech recognition experiments using speech from human-to-human and human-to-machine conversations. Conversational speech inputs to machines consist of frozen expressions such as greetings and yes/no statements, and informative individual expressions like numerical data such as dates and telephone numbers. The former has a lower perplexity and acoustic characteristics close to spontaneous speech. The latter has a higher perplexity and acoustic characteristics close to read speech. Each utterance or each inter-pausal unit can be classified into the former or the latter. This new knowledge will help future research on speech translation and dialogue systems.

1. INTRODUCTION

Speech translation and dialogue systems must accept conversational speech. In this paper, we discuss acoustic and linguistic characteristics based on results of speech recognition experiments using speech from human-to-human and human-to-machine conversations.

The characteristics of bilingual travel conversation data through human interpreters differ from those of Japanese monolingual travel conversation data [1]. Since there is no measure to indicate differences between acoustic and linguistic characteristics, we first present numerical data for acoustic characteristics based on speech recognition results using an acoustic model of read speech and an acoustic model of spontaneous speech.

Second, we discuss similarities and differences between human-to-human and human-to-machine conversational speech data. As human-to-machine conversational speech data, we use speech data collected in a dialogue experiment through the English/Japanese ATR-MATRIX speech translation system that was built by ATR Interpreting Telecommunications Research Laboratories [2].

Section 2 describes the concept and techniques of the speech recognition experiments. The discussion in this paper is based on the experiments. Section 3 presents characteristics of human-to-human conversational speech data and discusses differences between direct and indirect communications. Section 4 presents characteristics of human-to-machine conversational speech data in comparison with those of human-to-human communications. Section 5 provides discussions, and section 6 gives our conclusions.

2. SPEAKING STYLE CLASSIFICATION BASED ON MAXIMUM LIKELIHOOD CRITERION

Two kinds of speaking style-dependent acoustic models are used as a tool to investigate acoustic characteristics. One is an acoustic model of spontaneous speech and the other is an acoustic model of read speech. The former model is trained using conversational speech involving direct communications between two Japanese speakers, i.e., Japanese monolingual travel conversation database [3]. The latter model is trained using read speech involving 503 phonetically-balanced Japanese sentences.

Each utterance can be classified into spontaneous or read speech by selecting the most suitable speaking style dependent acoustic model based on the maximum likelihood criterion during the speech recognition process. In this method, the input speech is recognized by both acoustic models and, from among the results of both models, the result and the model with the maximum likelihood is selected.

To reduce the computational cost, based on the assumption that the most suitable model remains unchanged during a single utterance or a single inter-pausal unit, we implement our speech recognition with the most suitable acoustic model as follows.

First, we start with two different hypotheses corresponding to two kinds of acoustic models, and start frame-synchronous Viterbi search. For each frame, these hypotheses grow separately for each acoustic model, and the hypotheses with lower likelihoods, among the hypotheses for all acoustic models, are pruned by using the beam search technique. At the end of the input speech, the most suitable acoustic model and the recognized result are obtained by selecting the hypothesis with the maximum likelihood from among the hypotheses for all acoustic models.

In our speech recognition experiments, we physically define the inter-pausal units by segments separated by silent periods. We assume the lengths of the silent periods to be more than or equal to 700 ms.

3. CHARACTERISTICS OF HUMAN-TO-HUMAN CONVERSATIONS

3.1. Data collection

ATR Interpreting Telecommunications Research Laboratories have collected conversational speech data for speech translation research [4, 3, 1]. Two kinds of data collections were adopted as shown in Figure 1. One involved direct communications between two Japanese speakers, i.e., Japanese monolingual travel conversation database [4, 3]. The other involved indirect communications between
Table 1. Summary of monolingual and bilingual travel conversation databases

<table>
<thead>
<tr>
<th>Conversation style</th>
<th>Monolingual (J-J)</th>
<th>Bilingual (J-E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of collected conversations</td>
<td>892</td>
<td>618</td>
</tr>
<tr>
<td>Speaker participants</td>
<td>499</td>
<td>71</td>
</tr>
<tr>
<td>Interpreter participants</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>Total number of utterances</td>
<td>22,874</td>
<td>301,984</td>
</tr>
<tr>
<td>Total number of Japanese words</td>
<td>191,135</td>
<td>391,509</td>
</tr>
<tr>
<td>Utterances including one filled pause or more</td>
<td>42%</td>
<td>24%</td>
</tr>
<tr>
<td>Utterances including one self-repair or more</td>
<td>6%</td>
<td>3%</td>
</tr>
<tr>
<td>Average unit length (per utterance)</td>
<td>35 morae</td>
<td>30 morae</td>
</tr>
</tbody>
</table>

Table 2. Automatic selection of acoustic models for human-to-human conversational speech data

<table>
<thead>
<tr>
<th>Acoustic model</th>
<th>Speaking style</th>
<th>Direct communications</th>
<th>Indirect communications</th>
<th>Read speech from conversational text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spontaneous speech</td>
<td>84.0%</td>
<td>64.8%</td>
<td>35.3%</td>
<td></td>
</tr>
<tr>
<td>Read speech</td>
<td>15.4%</td>
<td>35.2%</td>
<td>64.7%</td>
<td></td>
</tr>
<tr>
<td>Test set perplexity</td>
<td>21.4</td>
<td>18.4</td>
<td>18.4</td>
<td></td>
</tr>
<tr>
<td>Speech recognition accuracy</td>
<td>86.0%</td>
<td>89.9%</td>
<td>94.8%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Two kinds of data collections of human-to-human conversations

A Japanese speaker and an English speaker through interpreters, i.e., bilingual travel conversation database [4, 1]. The task involves travel conversations between a tourist and a front desk clerk at a hotel. This task was selected because of its familiarity to people, and its expected use in future speech translation systems. Table 1 shows the summary of the monolingual and bilingual travel conversation databases.

The frequency of filled pauses and the frequency of self-repairs of the bilingual travel conversation database are less than those of the monolingual travel conversation database. Several constraints may cause these significant differences. Again, the Japanese monolingual conversation database involves direct communications between two Japanese speakers, but the bilingual database involves indirect communications. Accordingly, the time for turn around is much longer

Figure 2. Coverage of acoustic models for human-to-human conversations

For the bilingual case than for the monolingual case, such that the speakers in the bilingual case may have more than enough time to consider what to say for, and how to say, their next utterances.

3.2. Experiments and results

From the Japanese monolingual travel conversation data, we used 533 utterances (6342 words) by 17 males and 25 females. From the bilingual travel conversation data through human interpreters, we used 330 utterances (1665 words) by Japanese native speakers (not interpreters) comprising eight males and 15 females. We conducted an additional experiment using read speech data from conversational text, that is, 295 utterances (4000 words) of ten males and ten females. Here, the speech data and speakers were not used to train acoustic models.

Table 2 shows experimental results, such as the coverage of automatically selected acoustic models, test set perplexity, and speech recognition accuracy. Figure 2 illustrates the coverage of acoustic models for human-to-human con-
Characteristics of Human-to-Machine Conversations

4.1. Data Collection

As human-to-machine conversational speech data, we used speech data collected in a dialogue experiment through the English/Japanese ATR-MATRIX speech translation system that was built by ATR Interpreting Telecommunications Research Laboratories [2]. Figure 3 illustrates the data collection. The dialogue experiment was carried out between Japanese speakers (three males and two females) and an English speaker (one female). Each Japanese speaker participated in three sessions. We selected the first session because this session was expected to have the most typical characteristics of human-to-machine conversations. The speakers sometimes uttered the same or similar expressions several times when the first trial utterance had some recognition errors. In the speech recognition experiment, we used only the first trial utterance for each turn of the Japanese speakers; the experiment contained 110 utterances (955 words). Here, the speech data and speakers were not used to train acoustic models.

4.2. Experiments and Results

Table 3 shows experimental results, such as the coverage of automatically selected acoustic models, test set perplexity, and speech recognition accuracy.

Asynchronous models were changed for 10.0% of the utterances. In this respect, the human-to-human and human-to-machine speech data differed. An acoustic model of spontaneous speech was selected in 52.7% of the utterances and an acoustic model of read speech was selected in 37.3% of the utterances. This human-to-machine conversational data had similar characteristics to the human-to-human indirect communication speech data, that is, the bilingual travel conversation data through human interpreters, from the viewpoint of the rate of acoustic model selection.

5. Discussions

The perplexity, or entropy per utterance, was used as a tool to investigate linguistic characteristics.

Figure 4 shows a histogram of an entropy distribution in speech recognition experiments involving direct speech communications, i.e., the Japanese monolingual speech database. The horizontal line indicates the entropy per utterance. This entropy value is calculated by the transcribed utterance and rounded from a fraction to an integer. The vertical line indicates the frequency normalized by percentage because the total number of utterances differed from experiment to experiment. Two kinds of bars are shown in the bar graph. One is for utterances classified into spontaneous speech by the experiment. The other is for utterances classified into read speech by the experiment. Utterances with a lower entropy tend to consist of frozen expressions such as greetings, yes/no statements, and other frequently used expressions. Utterances with higher entropy tend to consist of informative individual expressions like numerical data such as dates and telephone numbers, or sometimes unfamiliar content words.

Figure 5 shows a histogram of an entropy distribution of indirect speech communications, i.e., the bilingual travel conversations through human interpreters. Figure 6 shows a histogram of an entropy distribution of read speech from conversational text. Figure 7 shows a histogram of an entropy distribution of human-to-machine conversations.

An acoustic model of spontaneous speech tended to be...
selected for the direct human-to-human communications as shown in Figure 4. An acoustic model of read speech tended to be selected for the read speech from conversational text (except utterances with a lower entropy of less than three) as shown in Figure 6. The indirect communication data, that is, the bilingual travel conversations through human interpreters, had an intermediate characteristic, i.e., it was positioned between the direct communication data, that is, the Japanese monolingual conversations, and the read speech data from conversational text. The human-to-machine conversational data had similar characteristics to the human-to-human indirect communication speech data.

According to Figure 7, an acoustic model of spontaneous speech was selected for all utterances with a lower entropy of less than three and an acoustic model of read speech tended to be selected for utterances with a higher entropy of more than three.

The information in these figures is summarized as follows. Conversational speech inputs to machines consist of frozen expressions such as greetings and yes/no statements, and informative individual expressions like numerical data such as dates and telephone numbers. The former has a lower perplexity and acoustic characteristics close to spontaneous speech. The latter has a higher perplexity and acoustic characteristics close to read speech. Each utterance can be classified as one of the former or one of the latter.

6. CONCLUSIONS

Speech translation and dialogue systems must accept conversational speech. In this paper, we discuss acoustic and linguistic characteristics based on results of speech recognition experiments using speech from human-to-human and human-to-machine conversations. Conversational speech inputs to machines consist of frozen expressions such as greetings and yes/no statements, and informative individual expressions like numerical data such as dates and telephone numbers. The former has a lower perplexity and acoustic characteristics close to spontaneous speech. The latter has a higher perplexity and acoustic characteristics close to read speech. Each utterance can be classified as one of the former or one of the latter. This new knowledge will help future research on speech translation and dialogue systems.

REFERENCES


