Pronunciation Variant Analysis using Speaking Style Parallel Corpus

Hideharu Nakajima, Izumi Hirano†, Yoshinori Sagisaka, and Katsuhiro Shirai†

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto, 619-0288 Japan
{nakajima,sagisaka}@slt.atr.co.jp, and
† School of Science and Engineering, Waseda University
3-4-1 Okubo Shinjuku-ku Tokyo, 169-8555 Japan
{izumi,shirai}@shirai.info.waseda.ac.jp

Abstract
To improve the recognition accuracy for spontaneous conversational speech, we collected a corpus to study how spontaneous conversational speech differs from read style speech. The corpus consists of two parts: 1) spontaneous conversational speech and 2) read speech with the same word transcriptions as the conversational speech. In word and phone recognition experiments, it was confirmed that, for the Japanese language, the recognition of spontaneous speech is harder than that of read speech. By comparing of recognition results, we found that, both in the occurrence of errors appearing with speaking style changes, and in the types of pronunciation variants, there are differences that depend on the linguistic categories that misrecognized words belong to. We confirmed that linguistic categories also affect pronunciation variants that deteriorate the recognition accuracy.

1. Introduction
It has been reported that by comparing the recognition accuracy for spontaneous conversational speech (CS) to that for read speech (RS), one can see that CS recognition is harder than RS recognition under the condition of different speaking styles even though the word transcriptions are identical [1]. CS recognition is indeed difficult, but some parts of utterances can be recognized while other parts of the same utterances cannot. At misrecognized parts, the actual pronunciations might be different from the standard pronunciations in lexica. It is necessary to capture speech variations at these parts to recognize them correctly. Subsequently, to improve the CS recognition accuracy, we need to confirm whether linguistic categories affect pronunciation variations and features such as the language model likelihood and speech rate. Several pronunciation models have also been proposed [2][3][4].

For English speech recognition, linguistic categories have definitely been considered for modeling, e.g., a function word dependent phone model [5]. For Japanese spontaneous speech recognition, in contrast, we need to confirm whether linguistic categories affect pronunciation variations and errors. To date, however, there has been few attempts at analyzing spontaneous speech variants in terms of linguistic categories. In this paper, to capture speech variations, we analyze speech recognition errors in terms of linguistic categories using speech data with different speaking styles. In section 2, differences of speech recognition results are shown, after explaining the speech data used in our recognition experiments. In section 3, differences of errors between CS and RS are analyzed and error characteristics are summarized.

2. Corpus and its characteristics

2.1. Summary of corpus collection
We collected a corpus to consist of two kinds of speech with differences only in speaking style. Here, we describe an overview of the collection.

Firstly, CS was collected (the details are described in Appendix A). The conversations are simulated conversations between hotel clerks and customers. We had already collected this kind of conversation corpus for speech translation research [6]. Next, CS was transcribed at the word level.

Subsequently, in the CS transcriptions, we deleted utterances consisting of only frequent expressions e.g., “hai (yes).” Then, since hesitations are not included in the corpus for language model training, utterances including hesitation were eliminated. Then, long utterances were divided into several shorter parts. Next, these screened transcriptions were read by the same speaker as at the CS collection stage, and the read speech was recorded.

Finally, both the CS and RS phone transcriptions were made.

This corpus gives two types of speech in which the speaking style differs but the word level transcriptions are identical. Hereafter, we use this data for a pronunciation variant analysis. The total number of utterances is 140, and the total number of words is 2,515.

2.2. Corpus characteristics based on recognition performance
To compare CS and RS in terms of the recognition accuracy, focus was on word recognition using N-gram word statistics and phone recognition without phonetic constraints. The conditions of these experiments and the employed models are further described in Appendix B.

Word and phone accuracies are shown in Table 1. Under word recognition, there is a 31 point difference in word accuracy and a 17.8 point difference in phone accuracy between RS and CS recognition. Under phone recognition, there is a 23.4 point difference in phone accuracy. Hence, like the results reported before [1], including phone recognition, it was confirmed that CS recognition is more difficult than RS recognition.

Since the phone accuracy of CS is lower than that of RS in phone recognition results, pronunciations in CS must vary significantly. Even when a lexicon is used, the word recognition phone accuracy of CS is lower than that of RS (in the second line of Table 1). That is, a lexicon is unable to describe pronun-
3. Variant analysis using aligned corpus

To find factors for improving the CS recognition accuracy, we make clear i) where pronunciation variations occur in identical word sequences in RS and CS, ii) the relationship between linguistic category and occurrence of pronunciation variation, and iii) what types of pronunciation differences exist between CS and RS. For this purpose, we align CS with RS at both the word and phone transcription levels. We also use word recognition results and the phone sequences obtained from both CS and RS phone recognition results (at section 2.2). Our procedure is as follows:

**Step 1:** Align word-level recognition results of CS with those of RS along with word-level transcriptions.

**Step 2:** Divide phone transcriptions of both CS and RS into words.

**Step 3:** In each style (RS and CS), divide phone-level recognition results into words.

**Step 4:** Align phone transcriptions with phone-level recognition results.

**Step 5:** Classify words recognized correctly in RS but misrecognized in CS into the following two types:

(a) Type A variations, when they have different phone transcriptions between CS and RS.

(b) Type B variations, when they have the same phone transcriptions but the CS phone recognition result is different from the CS phone transcriptions (phone error).

The procedure is also illustrated in Figure 1. The above analyzing process proceeds from the center to the upper part (or lower part) of Figure 1. The dashed lines indicate alignment boundaries at both the word and phone levels.

### Table 1: Comparison of read speech (RS) and conversational speech (CS) in terms of word and phone recognition accuracy

<table>
<thead>
<tr>
<th>Type</th>
<th>Type of Acc.</th>
<th>Word Acc.</th>
<th>Word Acc.</th>
<th>Phone Acc.</th>
<th>Phone Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>RS [%]</td>
<td>84.7</td>
<td>CS [%]</td>
<td>53.7</td>
<td>RS-CS [point]</td>
</tr>
<tr>
<td>Phone</td>
<td>Phone Acc.</td>
<td>92.1</td>
<td>74.3</td>
<td>17.8</td>
<td></td>
</tr>
</tbody>
</table>

To analyze Part 2 and Part 4 in which words are misrecognized in CS sufficiently, resulting in a low word accuracy. To overcome this problem, we must capture pronunciation variations in CS more.

Table 3 shows what types of pronunciation differences exist between CS and RS. For this purpose, we align CS with RS at both the word and phone transcription levels. We also use word recognition results and the phone sequences obtained from both CS and RS phone recognition results (at section 2.2). Our procedure is as follows:

**Step 1:** Align word-level recognition results of CS with those of RS along with word-level transcriptions.

**Step 2:** Divide phone transcriptions of both CS and RS into words.

**Step 3:** In each style (RS and CS), divide phone-level recognition results into words.

**Step 4:** Align phone transcriptions with phone-level recognition results.

**Step 5:** Classify words recognized correctly in RS but misrecognized in CS into the following two types:

(a) Type A variations, when they have different phone transcriptions between CS and RS.

(b) Type B variations, when they have the same phone transcriptions but the CS phone recognition result is different from the CS phone transcriptions (phone error).

The procedure is also illustrated in Figure 1. The above analyzing process proceeds from the center to the upper part (or lower part) of Figure 1. The dashed lines indicate alignment boundaries at both the word and phone levels.

### 3.1. Where pronunciation variations occur

To know where a pronunciation variation occurs, we use the result of the first step in the above-mentioned procedure. From the result, the whole aligned corpus can be divided into the following four parts as in Table 2: i) recognized correctly in both RS and CS (Part 1), ii) recognized correctly in RS but misrecognized in CS (Part 2), iii) recognized in RS but recognized correctly in CS (Part 3), and iv) misrecognized in both RS and CS (Part 4). To improve the CS recognition accuracy, we need to analyze Part 2 and Part 4 in which words are misrecognized in CS sufficiently, resulting in a low word accuracy. To overcome this problem, we must capture pronunciation variations in CS more.

### 3.2. Linguistic categories and pronunciation variants

To know whether there is a difference in the occurrence of errors in each linguistic category when the speaking style changes from RS to CS (in Part 2), we calculated the deterioration rate as the error rate in each linguistic category in CS word-level recognition. Table 3 shows to what extent words, that are correct in RS recognitions, are misrecognized in CS word-level recognition for each linguistic category. In Table 3, the left column is for content words or function words as higher categories than the part of speech. The middle column is for the part of speech of each word. The value of the line “MEAN” is the mean value of the deterioration rate for each category. Table 3 shows that errors occur more in content words than in function words when the speaking style is different (RS v.s. CS). This finding suggests that pronunciation variants occur more in content words than in function words in CS.

### Table 3: Detetioration rate (deletion and substitution error) for each linguistic category (part of speech, and content word or function word) when the speaking style changes from RS to CS

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>Adjectival determiner</th>
<th>Conjunction</th>
<th>Adjective</th>
<th>Pronoun</th>
<th>Noun</th>
<th>Verb</th>
<th>Adverb</th>
<th>Interjection</th>
<th>MEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Word</td>
<td>76.9</td>
<td>59.1</td>
<td>53.1</td>
<td>51.3</td>
<td>41.1</td>
<td>41.6</td>
<td>31.6</td>
<td>14.9</td>
<td>36.6</td>
</tr>
<tr>
<td>Function Word</td>
<td>47.8</td>
<td>45.1</td>
<td>31.5</td>
<td>21.0</td>
<td>12.2</td>
<td>28.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 3.3. Pronunciation variation

To know the variant types in each category in Table 3, we use phone sequences obtained from the phone recognition results of those words classified into “Type B variations” at with step 5 in the above-mentioned procedure. We analyze three-tuples of the phone in each word, particularly the tuples with phone deletion or substitution errors at their centers. Since the phone sequences obtained from the phone recognition results are aligned with each word, the part of speech and linguistic category such as the content word or function word that every three-tuples belongs to are obvious. Hence, the classification of the three-tuples along with the linguistic categories is easy.

The total number of extracted tuples involving phone substitution is 43 (35 for content words and eight for function words). The total number of extracted tuples involving phone deletion is 21 (19 for content words and two for function words). The following tendencies can be observed:

- **Phone Substitution**
  - For three-tuples in nouns of content words, 75% of the phone substitutions occur at phones between vowels. (There were no tendencies in other parts of speech, except for nouns.)

- **Phone Deletion**
  - Regardless of the linguistic category, over 90% of the phone deletions occur at phone $X_{j}$ in sequences such as “$YX_{j}X_{k}$” and “$X_{j}X_{k}Y$” where $X_{j}$ and $X_{k}$ belong to one of three classes among the $\{j,w,r\}$ class (semi-vowel class), $\{q,n,g,t\}$ class, and vowel class, and Y belongs to any class.

Examples are shown in Appendix C. This tendency shows at least two kinds of pronunciation variants: i) variants dependent on the linguistic category and ii) variants common through all linguistic categories.

### 4. Discussion

Through the above analysis, it was found that

1. At the both word and phone levels, CS recognition is harder than RS recognition (by comparing the RS and CS accuracies in Table 1).
2. Pronunciation modeling is necessary for CS (by comparing CS and RS in terms of word recognition phone accuracy in Table 1).
3. There are proportionally more errors in content words than in function words (from Table 3).
4. There are two kinds of pronunciation variants: one dependent on the linguistic category and the other common through all linguistic categories (section 3.3).

The fourth in particular suggests that we should consider linguistic categories when some pronunciation model learns variants, and when we apply the learned pronunciation model to generate pronunciation variants from canonical pronunciations. It also suggests that we can take another approach that is different from the proposed approach, where only one model or one rule set is learned for the whole variant and the model or the rule set is used equally to generate pronunciation variants irrespective of the linguistic category.

Additionally, the third discovery suggests that we need to cope with variants in content words more than in function words. Since the number and places (document) of occurrences of content words are more limited, it might be difficult to have a sufficient amount of data to learn variants in content words automatically.

### 5. Summary and future work

Our analyses showed that there are more pronunciation variants in content words than in function words when the speaking style changes, and that it is necessary to include the factor of linguistic category when modeling pronunciation variants because in pronunciation variant types there is differences that depend on the linguistic categories. In the future, we are going to collect...
a larger speaking style parallel corpus to verify our discoveries described above and to confirm effects in CS recognition.

Acknowledgments
We thank all the member of ATR Spoken Language Translation Research Laboratories for the data collection and tagging.

6. References

A. Details of the corpus construction
A.1. Collection of conversational speech
Simulated conversations between hotel clerks and customers were collected. These conversations are of the same type as the conversations that ATR has been collecting. The used language is Japanese. Twenty-seven words (like television, air conditioner, checkout, postal mail, etc.) were given as topic words to the speakers in a conversation. The number of speakers was two, and the total length of the conversation was about 36 minutes. The total number of utterances, the number of interjections, and the number of hesitations are listed in Table 4.

Table 5: Acoustic feature analysis
<table>
<thead>
<tr>
<th>Feature Parameter</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling Frequency</td>
<td>16 kHz</td>
</tr>
<tr>
<td>Frame Shift</td>
<td>10 ms</td>
</tr>
<tr>
<td>Frame Length</td>
<td>20 ms (Hamming)</td>
</tr>
<tr>
<td>Pre-emphasis</td>
<td>0.98</td>
</tr>
<tr>
<td>Logpower</td>
<td>Δlogpower</td>
</tr>
<tr>
<td>MFCC</td>
<td>12 dim.</td>
</tr>
<tr>
<td>ΔMFCC</td>
<td>12 dim.</td>
</tr>
<tr>
<td>Mean Cepstrum</td>
<td>Normalized power</td>
</tr>
</tbody>
</table>

Table 6: Extracted three-tuples (SD means a deleted phone)
<table>
<thead>
<tr>
<th>Substitutions</th>
<th>Deletions</th>
</tr>
</thead>
<tbody>
<tr>
<td>(noun) i, k, a ⇒ i, f, a</td>
<td>c, r, e</td>
</tr>
<tr>
<td>(noun) i, b, o ⇒ i, p, o</td>
<td>h, a, i</td>
</tr>
<tr>
<td>(noun) i, h, o ⇒ i, r, o</td>
<td>a, r, i</td>
</tr>
<tr>
<td>(noun) o, n, o ⇒ o, u, o</td>
<td>a, w, a</td>
</tr>
<tr>
<td>(noun) o, t, o ⇒ o, t, i</td>
<td>a, ng, g</td>
</tr>
<tr>
<td>(noun) o, q, t ⇒ o, s, d</td>
<td>a, q, t</td>
</tr>
<tr>
<td>(noun) o, t, o ⇒ o, s, d</td>
<td>h, a, i</td>
</tr>
<tr>
<td>(noun) u, r, u ⇒ u, s, d</td>
<td>r, i, j</td>
</tr>
<tr>
<td>(noun) f, u, r ⇒ f, s, d</td>
<td>s, d, r</td>
</tr>
</tbody>
</table>

A.2. Collection of read speech
As read speech, we collected two kinds of speech as follows i) Read without emotions: utterances in word transcriptions were randomly presented to eliminate contextual effects, and then the speakers read them without adding any emotional expression, and ii) Read Conversational: the speakers read word transcriptions like talking to their partners according to information on transcription situations in an actual conversation. Since there was not a big difference between recognition accuracies ((i) and (ii)), we used the former speech data for analysis purposes in this research.

B. Experimental conditions for recognition
For both acoustic and language model training, the ATR spoken language database [6] was used. For the acoustic model for both CS and RS (word and phone) recognitions, read-style phoneme balanced speech data (235 speakers) in the database was used. The acoustic model is a shared-state context-dependent (tri-phone) HMMs produced with the ML-SSS algorithm [8]. Parameters of acoustic feature analysis are listed in Table 5. As our language model for word recognitions, the multiclass composite n-gram [9] trained with the above-mentioned database was used. This model is a kind of class n-gram and has connected words consisting of frequent sequences of words in the lexicon. The model has been confirmed to achieve a high accuracy (equal to or more greater than a word trigram) with a few parameters. The total number of lexicon entries was 13,732, and the training data size was 1,606,951 words. In phone recognition, phonetic constraints were not used. As decoders, we used ATRSPREC [7], which uses multi-pass search.

C. Example of pronunciation variants
(Extracts)