ACOUSTIC MODEL FOR ROBUST SPEECH RECOGNITION OF STRESSED
JAPANESE SPEECH

Kozo Okuda, Tomoko Matsui and Satoshi Nakamura

ATR Spoken Language Translation Research Laboratories
2-2-2, Hikaidai, Seika-cho, Soraku-gun, Kyoto, Japan
{kokuda, tmatsui, nakamura}@slt.atr.co.jp

ABSTRACT
In making an error recovery utterance, the users of a speech recognition system utter more clearly and slowly. In addition, the occurrence of syllable-stressed speech increases in Japanese. This paper investigates a method that is robust in recognizing syllable-stressed speech uttered for error recovery. In syllable-stressed speech, each syllable is uttered slowly and emphasized. The characteristics of each syllable is strongly altered by this modification and thereby the speech recognition performance is reduced. To cope with these problems, we propose a new recognition method. In this paper we propose an acoustic modeling method for recognizing the syllable-stressed speech by combining existing acoustic models. By our method, it is not necessary to collect additional training data. Our results indicate that the proposed method improves performance. Furthermore, the method does not need any expansion of the recognition lexicon or explicit selection of the models.

1. INTRODUCTION
There are many types of speaking styles in spontaneous speech, such as speech to convey stress, emotion and error recovery. If the speaking style is changed, the performance of a speech recognition system becomes degraded. In a speech recognition system for an intelligent human-machine interface, it is crucial to achieve robustness against speaking style variations. The system must be especially robust with regard to error recovery utterances because current speech recognition systems can not avoid errors. In an error recovery utterance, the user changes his or her speaking style to help the system recognize the speech. The user typically speaks more clearly and slowly. In addition, the occurrence of syllable-stressed utterances increases in Japanese.

Previous studies\cite{1,2} demonstrated that acoustic models made or adapted by the error recovery utterance data reduce the word error rate for typical error recovery utterances. However, these studies did not mention the effect on syllable-stressed utterances.

In this paper, we investigate a method that is robust against this syllable-stressed speech. In syllable-stressed speech, each syllable is uttered slowly and emphasized. The characteristics of each syllable is strongly altered by this modification and thereby the speech recognition performance is reduced. To cope with this problem, we propose an acoustic modeling method for recognizing the syllable stressed speech by combining existing acoustic models. Also we assume that the syllable-stressed speech is uttered in the manner between a continuous utterance and an isolated syllable utterance. Thus we use vowel models, which are succeeded by silence in conventional triphone models. For detecting the change in acoustic characteristics between syllables, we use left context dependent vowel biphone models. These models and conventional triphone models compose a single acoustic model by a multi-path approach. During decoding, the path that has the highest likelihood is implicitly selected for an input utterance. The proposed method has an advantage that collection of additional training data is not necessary.

This paper is organized as follows. In the first section, we explain our baseline system used in this paper. Next we report the frequency and the features of the syllable-stressed speech in error recovery. In addition, we report the results of recognition experiments on syllable-stressed speech using our baseline system. After that we propose a method to create acoustic models that are robust against the syllable-stressed utterances. In the last section we report recognition experiments using our proposed method.

2. BASELINE SYSTEM
In the recognition experiment, we use our recognizer called ATRSPREC\cite{3}. For acoustic modeling, a 25-dimensional feature vector (12-dimensional mel-cepstral coefficients, 12-dimensional first order derivatives of mel-
cepstral coefficients and 1-dimensional first order derivative of logarithmic power) is computed with a 20 ms window length and a 10 ms frame shift. The baseline acoustic models are gender-dependent shared-state HMMs (1,400 states in total) with five Gaussian mixture components per state[4]. These HMMs are trained using data uttered by 167 male (about 2 hours) and 240 female (about 3 hours) subjects using the travel task database collected in our laboratories.

For language modeling, we use multi-class composite N-gram[5]. This language model is trained with the same training data set as the acoustic model. The number of “from” class of the preceding word is 700, and the number of “to” class of the succeeding word is 700. The size of the recognition lexicon is 27K words.

3. FEATURES OF SYLLABLE-STRESSED SPEECH IN ERROR RECOVERY

In utterances of error recovery, the speaking style changes. The users speak more clearly to help the system to recognize their speech, the duration of each phoneme increases, the length and frequency of silence periods increases, and the pitch rises[6]. In addition, the occurrence of syllable-stressed speech increases in Japanese. Syllable-stressed speech is uttered slowly and emphasized. The characteristics of each syllable is strongly altered by this modification.

In this section, we describe collection of error recovery utterances and investigate the occurrence frequency of syllable-stressed speech in error recovery speech. In order to collect error recovery utterances realistically, we simulated recognition error. The average word accuracy of our recognition system is over 80%, although the performance degrades to 60% with the worst speaker. Therefore, we simulated a word error rate about 40%. In the words containing error, we set 50% of the words to have two consecutive errors, 25% of the words to have three consecutive errors, and 12.5% of the words to have four and five consecutive errors. The subjects input the words indicated on a display to the system using a microphone. If the recognition error occurs, the system indicates only the information of the occurrence of an error, and the subjects rephrase their speech until the recognition is successful. In this way, we collected 210 words from each of the five subjects.

Next, we investigated the acoustic likelihood of the error recovery utterance and compensated it with the first uttered speech to determine the occurrence frequency of syllable-stressed speech. The results are shown in figure 1. In an error recovery utterance, the acoustic likelihood degrades from normal speech. In the worst case, the acoustic likelihood of about 20% of the speeches was degraded by more than 1.2 times that of normal speech included in error recovery speech.

Figure 2 shows a spectrogram of the utterances that degrade acoustic likelihood by more than 1.2 times from normal speech. This figure represents the Japanese utterance of /jizai/. The circled area indicates the intermediate part of /ji/ and /s/. Compared with a normal utterance, the continuity of the syllable is unnatural and the acoustic characteristics are varied.

![Figure 1: Occurrence of syllable-stressed speech in error recovery](image1.png)

4. RECOGNITION EXPERIMENT USING BASELINE SYSTEM

To investigate the influence of syllable-stressed utterances on the speech recognition system, we conducted recognition experiments using the baseline system. Because our baseline system is for continuous speech, we collected 20 sentence data spoken normally and with conscious syllable-stressing from each of 5 males and 5 females. The results are shown in Table 1.

The performance for normal utterance using the acoustic model trained for continuous speech recogni-
Table 1: Word accuracy using baseline system

<table>
<thead>
<tr>
<th>speaker</th>
<th>normal utterance</th>
<th>syllable-stressed utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>baseline</td>
<td>syllable-stressed</td>
</tr>
<tr>
<td>sp1</td>
<td>75.8%</td>
<td>18.6%</td>
</tr>
<tr>
<td>sp2</td>
<td>78.6%</td>
<td>-28.9%</td>
</tr>
<tr>
<td>sp3</td>
<td>87.3%</td>
<td>-44.0%</td>
</tr>
<tr>
<td>sp4</td>
<td>78.6%</td>
<td>-86.9%</td>
</tr>
<tr>
<td>sp5</td>
<td>80.6%</td>
<td>-51.9%</td>
</tr>
<tr>
<td>sp6</td>
<td>81.4%</td>
<td>-27.1%</td>
</tr>
<tr>
<td>sp7</td>
<td>81.3%</td>
<td>-12.0%</td>
</tr>
<tr>
<td>sp8</td>
<td>73.8%</td>
<td>6.2%</td>
</tr>
<tr>
<td>sp9</td>
<td>74.4%</td>
<td>-2.5%</td>
</tr>
<tr>
<td>sp10</td>
<td>75.0%</td>
<td>32.1%</td>
</tr>
</tbody>
</table>

...tion is about 80%. However, for syllable-stressed utterances, the performance is degraded. The reason for this degradation is the mismatch between the acoustic models due to the change in the acoustic characteristics caused by the stress and the change in each phoneme’s duration.

Next we tried to apply the model adaptation of each subject. In the adaptation, we used all data because the amount of data from each subject is limited and the experiment was conducted by using closed data. We adapted mean values and transition probabilities by using MAP-VFS. The results are shown in Table 1. The adaptation improved the performance for syllable-stressed utterances. However, the performance is still low compared with those of normal utterances.

From the results described above, we observed that a syllable-stressed utterance has a special acoustic characteristic, and the conventional triphone models cannot deal with these characteristics. Therefore, the adaptation cannot sufficiently improve performance for practical use.

5. ACOUSTIC MODEL FOR SYLLABLE-STRESSED SPEECH

In the decoding process of a large vocabulary continuous speech recognition system, the decoder connects the HMM models in accordance with the description of the phonemes in the recognition lexicon. In particular, the context-dependent triphone HMM model improves performance by modeling the co-articulation of each phoneme. However, in the syllable-stressed utterances, the changes in the characteristics of the co-articulation of each phoneme occur as follows.

1. Each syllable is uttered in the manner closely to that of the isolated syllable utterance. Due to this change, the acoustic characteristics of each vowel become like those of a vowel succeeded by silence.

2. Each syllable is uttered in the manner between a continuous utterance and an isolated syllable utterance. Due to this change, the acoustic characteristics between syllables are changed.

Because of these changes, the conventional context-dependent triphone models do not work well and the performance of a recognition system becomes degraded.

This paper propose an acoustic modeling method for recognizing the syllable-stressed speech by combining existing acoustic models. The models combined in this method are as follows.

1. Vowel triphone models succeeded by silence for a syllable that is like an isolated syllable.

2. Left context-dependent vowel biphone models for the change in the acoustic characteristics at the co-articulation area between syllables.

For the complete isolated utterance, each model is supplemented with a one state pause model, which can be skipped. For smooth continuity with vowel triphone models succeeded by silence, models preceded by silence are added. These models compose a single acoustic model by taking a multi-path approach. Figure 3 shows an example of the proposed models t-a+k and a-k+i. During decoding, the path that has the highest likelihood is implicitly selected for an input utterance. By our method, it is not necessary to collect additional training data. Furthermore, this method does not need any expansion of the recognition lexicon or explicit selection of the models.

Figure 3: Example of proposed acoustic model

6. RECOGNITION EXPERIMENTS

We conducted several recognition experiments. The acoustic models used in these experiments were created as follows.

125
1. Conventional triphone models
Use the acoustic model of the baseline system.

2. Vowel triphone models succeeded by silence
Expand the vowel triphone models that are succeeded by silence in the acoustic model of the baseline system.

3. Left context-dependent vowel biphone models
Create a left context-dependent biphone model by using the same training data as the baseline system's acoustic model. This model is a gender-dependent model (1,400 states in total) with five Gaussian mixture components per state.

6.1. Recognition experiments using the proposed acoustic model
Figure 4 shows the results of recognition experiments for syllable-stressed speech. These results show that the proposed acoustic model improves performance more than model adaptation does. However, the word accuracy of speaker 5 was under 10%. The reason is the mismatch of the acoustic model of the baseline system. The mismatch of the acoustic models led to the degradation of the proposed method's performance.

6.2. Recognition experiments using the proposed acoustic models with model adaptation
To remove the mismatch of the acoustic model, we applied model adaptation to the proposed method. The mean values and transition probabilities of the baseline system's acoustic model and the left context-dependent biphone model were adapted to each speaker by using MAP-VFS. Figure 5 shows the results of the recognition experiment using this acoustic model. The proposed acoustic models with speaker adaptation can improve the performance of speaker 5, and the word accuracy of speakers 1 and 10 are over 80%.

7. CONCLUSIONS
We have described an acoustic model that is robust against syllable-stressed utterances in error recovery. This proposed acoustic model improved performance for syllable-stressed utterances without needing to collect the training data of syllable-stressed speech. However, some data could not improve word accuracy above 50%. There are some factors that can not be handled by our proposed method.

In future work, we will investigate these factors and develop a method to improve the performance of the data for these factors. Then, we will develop acoustic models that are robust against the other speaking style variations.

REFERENCES