TOWARDS THE CREATION OF ACOUSTIC MODELS FOR STRESSED JAPANESE SPEECH

Kozo Okuda, Tomoko Matsu and Satoshi Nakamura

ATR Spoken Language Translation Research Laboratories
{kokuda,tnatsu,nakamura}@slt.atr.co.jp

ABSTRACT

In error recovery utterance, the user using the speech recognition system changes his or her speaking style to aid the system in recognizing the speech. However, this change leads the mismatch between the acoustic models and reduces the performance of the system. This degradation causes a serious problem of speech recognition for a dialog system or a speech translation system. In error recovery utterance in Japanese, the occurrence of syllable-stressed speech increases. In syllable-stressed speech, each syllable is uttered slowly and emphasized. The characteristics of each syllable are strongly altered by this modification and the speech recognition performance is reduced. This paper investigates how to create acoustic models robust in recognizing error recovery utterances, especially syllable-stressed speech. In this paper, we propose an acoustic modeling method for syllable-stressed speech by combining existing acoustic models. Our results indicate that the proposed method improves the system performance. Furthermore, the method does not need any expansion of the recognition dictionary or explicit model selection.

1. INTRODUCTION

The performance of current speech recognition systems is improved by using statistical approaches and large speech databases. However, there are many types of speaking styles in spontaneous speech. In an error recovery utterance, the user changes his or her speaking style to aid the system in recognizing the speech. If the speaking style is changed, the performance of the speech recognition system becomes degraded[1]. Therefore, systems must be made especially robust to error recovery utterances because current speech recognition systems can not avoid errors.

In making an error recovery utterance, a user typically speaks more clearly and slowly. In Japanese, the occurrence of syllable-stressed speech increases. Previous studies[2][3] have demonstrated that acoustic models made or adapted by error recovery utterance data can reduce the word error rate for typical error recovery utterances. However, these studies have not mentioned effects on syllable-stressed utterances.

In this paper, we investigate how to create acoustic models robust against syllable-stressed speech. In syllable-stressed speech, each syllable is uttered slowly and emphasized. The characteristics of each syllable are strongly altered by this modification and the speech recognition performance is reduced. To cope with this problem, we propose an acoustic modeling method by combining existing acoustic models. We also assume that the syllable-stressed speech is uttered continuously in a manner between a continuous utterance and an isolated syllable utterance. Therefore, we use vowel models, which are succeeded by silence in conventional triphone models. For 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 the input utterance. The proposed method has an advantage in that the collection of additional training data is not necessary.

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

2. EXPERIMENTAL SETUP

2.1. Baseline system

In this paper, we use an LVCSR system called ATRSPREC [4]. 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[5]. These HMMs are trained using data uttered by 167 male (about 2 hours) and 240 female (about 3 hours) subjects using the travel arrangement task database[6] collected at our laboratories.

For language modeling, we use a multi-class composite N-gram[7]. This language model is trained with the same training data set as the acoustic model. The number of ‘from’ classes of the preceding word is 700, and the number of ‘to’ classes of the succeeding word is 700. The size of the recognition lexicon is 27K words.
The influence of syllable-stressed utterances on a speech recognition system was investigated. We conducted experiments using a baseline system. Because our recognition system is for continuous speech, we collected 20 sentence data spoken normally and consciously syllable-stressed from each of 5 males and 5 females. The results are shown in Table 1. The performance for the normal utterances is about 80%. However, for the 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 adaptations, we used all data because the amount of data from each subject was 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 adaptations improved the performance for syllable-stressed utterances. However, the performance was still low compared with that for 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 this characteristic. Therefore, adaptations cannot improve the performance enough for practical use.

3. Features of syllable-stressed speech in error recovery

In this section, we describe our collection of error recovery utterances and investigate the occurrence frequency of syllable-stressed speech in error recovery utterances.

3.1. Collection of syllable-stressed speech in error recovery utterances

In order to collect error recovery utterances realistically, we simulated recognition errors. The mean word accuracy of our recognition system is over 90%, although the performance degrades to 60% with the worst speaker. Therefore, we simulated a word error rate of about 40%. In the words containing errors, 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 inputted the words indicated on a display to the system using a microphone. If a recognition error occurred, the system indicated only the occurrence of the error, and the subjects respeared their speech until the recognition was successful. In this way, we collected 210 words from each of the five subjects.

3.2. Features of syllable-stressed speech

We investigated the acoustic likelihood of error recovery utterances and compared each of them 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 logarithmic acoustic likelihood degrades from the normal speech. In the worst case, the logarithmic acoustic likelihood of about 20% of all speech included in error recovery speech was degraded by more than 20% of that of the normal speech.

Figure 2 shows a spectrogram of utterances that degrade the acoustic likelihood. The circled area indicates the intermediate part of /jizai/. Compared with the normal utterance, the continuity of the syllable is unnatural and the acoustic characteristics are varied. In the syllable-stressed utterances, the changes in the characteristics of the co-articulation of each phoneme occur as follows.

<table>
<thead>
<tr>
<th>Table 1: Word accuracy using baseline system.</th>
</tr>
</thead>
<tbody>
<tr>
<td>speaker</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>sp1</td>
</tr>
<tr>
<td>sp2</td>
</tr>
<tr>
<td>sp3</td>
</tr>
<tr>
<td>sp4</td>
</tr>
<tr>
<td>sp5</td>
</tr>
<tr>
<td>sp6</td>
</tr>
<tr>
<td>sp7</td>
</tr>
<tr>
<td>sp8</td>
</tr>
<tr>
<td>sp9</td>
</tr>
<tr>
<td>sp10</td>
</tr>
</tbody>
</table>
Each syllable is uttered in a manner close to that of an isolated syllable utterance. Due to this change, the acoustic characteristics of each vowel become like those of a vowel succeeded by silence.

Each syllable is uttered in a 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 is degraded.

4. ACoustIC ModelING METHOD FOR SYLLABLE-STRESSED SPEECH

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

- Vowel triphone models succeeded by silence for a syllable that is like an isolated syllable.
- Left context-dependent vowel biphone models for changes in acoustic characteristics at co-articulation areas between syllables.

For a complete isolated utterance, each model is supplemented with a one-state pause model, which can be skipped for smooth continuity with the vowel triphone models succeeded by silence. Models preceded by silence are added. These models and the conventional triphone models compose a single acoustic model by taking a multi-path approach. Figure 3 shows an example of this proposed model t-a+k and a-k+i. During decoding, the path that has the highest likelihood is implicitly selected for the input utterance. With 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 model selection.

5. RECOGNITION EXPERIMENTS

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

- Using the acoustic model of the baseline system for the conventional triphone models.
- Expanding the vowel triphone models succeeded by silence in the acoustic model of the baseline system for the vowel triphone models succeeded by silence.
- Creating a left context-dependent biphone model by using the same training data of the baseline system’s acoustic model for the left context-dependent vowel biphone models. This model is a gender-dependent model (1,400 states in total) with five Gaussian mixture components per state.

5.1. 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 the performance more than model adaptation does. However, the word accuracy of speaker 5 was under 40%. The reason was the mismatch of the acoustic model of the baseline system. This mismatch led to the degradation of the proposed method’s performance.

5.2. Recognition experiments using the proposed acoustic model with model adaptation

To remove the mismatch of the acoustic model, we applied model adaptation to the proposed model. 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 results of the recognition experiment using this acoustic model. The proposed acoustic model with model adaptation improves the performance of speaker 5, and the word accuracies of speakers 1 and 10 are over 80%.

6. DISCUSSION

The results of the recognition experiment described above indicate the effect of our proposed method. In this section, we conducted several additional experiments to investigate the effect of additional models combined into the proposed acoustic model. Table 2 shows results for normal utterances. These results show that the proposed acoustic model does not degrade the performance for normal utter-

![Figure 3: Example of the proposed acoustic model.](image)

![Figure 4: Word accuracy of syllable-stressed utterance using the proposed acoustic model.](image)
Table 2: Word accuracies of normal utterances using the proposed model.

<table>
<thead>
<tr>
<th>speaker</th>
<th>baseline</th>
<th>proposed model</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp1</td>
<td>75.8%</td>
<td>80.8%</td>
</tr>
<tr>
<td>sp2</td>
<td>78.6%</td>
<td>79.3%</td>
</tr>
<tr>
<td>sp3</td>
<td>87.3%</td>
<td>88.0%</td>
</tr>
<tr>
<td>sp4</td>
<td>78.6%</td>
<td>84.8%</td>
</tr>
<tr>
<td>sp5</td>
<td>80.6%</td>
<td>80.0%</td>
</tr>
<tr>
<td>sp6</td>
<td>81.3%</td>
<td>75.7%</td>
</tr>
<tr>
<td>sp7</td>
<td>73.8%</td>
<td>74.4%</td>
</tr>
<tr>
<td>sp8</td>
<td>74.4%</td>
<td>77.9%</td>
</tr>
</tbody>
</table>

Figure 6: Effect of each model combined into the proposed model.

Figure 7: Effect of one-state pause model.

7. CONCLUSIONS

We described an acoustic model that is robust against syllable-stressed utterances in error recovery. This proposed acoustic model improved the performance for syllable-stressed utterances without needing to collect training data of syllable-stressed speech. However, some data could not improve the 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 data for these factors. We will then develop acoustic models that are robust against other speaking style variations.

REFERENCES