Effects of Time-Scale Modification on the Performance of Speech Recognition

Jin Ah Kang1 and Seung Ho Choi2

1 School of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea
jinari@gist.ac.kr
2 Department of Electronic and IT Media Engineering
Seoul National University of Science and Technology, Seoul 139-743, Korea
shchoi@seoultech.ac.kr

Abstract. In this paper, we describe the effects of time-scale modification on the performance of speech recognition. Speaking rate normalization is achieved by selecting a scaling factor of time-scale modification. In addition, we describe how speech data are classified depending on speaking rate, and then discuss the variations of the ASR performance for each class.

Keywords: Speech recognition, time-scale modification, speaking-rate normalization, rate of speech, Gaussian mixture classifier

1 Introduction

Speaking rate influences the performance of automatic speech recognition. For example, when fast speech comes in automatic speech recognition (ASR) systems, the word error rate tends to be increased because of increases in deletion and substitution errors [1][2]. Similarly, the word error rate for slow speech is also increased due to the large insertion errors even if the deletion and substitution errors are smaller than those for normal speech [2].

Thus, a rate of speech (ROS) can be measured by using one of signal-based approach, alignment-based approach, and GMM-based approach [3][4][5]. Since ROS is measured without regarding to ASR, the performance of ASR systems using speaking rate normalization rely upon the detection accuracy of ROS.

2 Distribution of Rate of Speech

We constructed a connected digit recognition system, where each digit string was composed of eleven Korean digits and the number of digits in a string varied from 3 to 7 [6]. A block of consecutive samples corresponding to 20 ms was used a frame. Each frame was multiplied by a Hamming window, and consecutive frames were spaced 10 ms apart. Therefore, the frame rate of baseline system was 100 Hz. Each frame was represented by 12 mel-frequency cepstral coefficients (MFCCs) and an
energy parameter, where MFCCs were derived from 22 mel-spaced filterbanks [7]. We applied a cepstrum mean normalization [8] to the MFCCs and also normalized the energy parameter so that the maximum value of energy in an utterance was 0 [9]. Finally, we added the first and second derivatives of MFCCs and energy to the feature vector. As a result, the feature vector dimension was 39.

In order to show the effect of speaking rate on the ASR performance, we defined an ROS as the number of phonemes per second. Fig. 1 displays the distribution of ROS with a solid curve. To classify the test data set according to ROS, we applied a Gaussian mixture classifier to the ROS distribution, where the number of Gaussian mixtures was three. By using the expectation-maximization (EM) algorithm, we obtained three Gaussian parameter sets and displayed three probabilities with dash-dotted curves in Fig. 1. The Gaussian mixture classifier provided decision boundaries, which were depicted with dashed lines in the figure, so we could divide the data set into three groups as indicated in Fig. 1. As a result, the numbers of utterances for Group 1, Group 2, and Group 3 were 761, 1241, and 600, respectively.

3 Effect of Time-Scale Modification on ASR Performance

The effect of speaking rate on WER was investigated by breaking up the WER of the baseline system into deletion, substitution, and insertion errors. Table 1 shows the WER for each group. The first and the second columns show the number of utterances of each group and the average ROS, respectively. As shown in the table, Group 1 had the smallest average ROS, which means that the utterances belonged to Group 1 has been spoken fast on the average. In addition, Group 1 showed the largest WER be-
cause of large deletion and insertion errors. On the other hand, the utterances in 
Group3 were spoken fast rather than those in Group1 and Group2, while Group3 had 
the lowest WER due to the smallest insertion errors. In general, insertion errors in- 
crease when speech is spoken slowly, and they are larger than deletion errors. On the 
other hand, fast spoken utterances cause more deletion errors. In this point of view, 
utterances in Group1 were spoken most slowly, and the speaking rate increased for 
Group2 and Group3. It is expected that we can further reduce the insertion errors for 
Group1 by increasing ROS and reduce the deletion errors for Group3 by decreasing 
ROS.

4 Conclusion

We have described the effects of time-scale modification on the performance of 
speech recognition. First, to classify the test data set according to rate of speech, we 
applied a Gaussian mixture classifier. Then, we discussed the variations of the ASR 
performance for each class.

Acknowledgments. This work was supported in part by the National Research Found- 
ation of Korea (NRF) grant funded by the MSIP (Ministry of Science, ICT&Future 
Planning) (No. 2012-010636).

References

recognition systems. In: Proceedings of IEEE International Conference on Acoustics, 
2. Nanjo, H., Kawahara, T.: Speaking-rate dependent decoding and adaptation for spontane- 
(2002)

Table 1. Comparison of word error rate (WER) of the baseline system according to differ- 
ent speaking rates.

<table>
<thead>
<tr>
<th>Speaking Group</th>
<th>No. Utterance</th>
<th>Avg. ROS</th>
<th>WER (%)</th>
<th>Errors (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>761</td>
<td>3.349</td>
<td>6.69</td>
<td>2.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.28</td>
</tr>
<tr>
<td>Group 2</td>
<td>1241</td>
<td>4.412</td>
<td>4.91</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.51</td>
</tr>
<tr>
<td>Group 3</td>
<td>600</td>
<td>5.474</td>
<td>4.75</td>
<td>1.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.65</td>
</tr>
<tr>
<td>Average</td>
<td>2602</td>
<td>4.346</td>
<td>5.36</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.52</td>
</tr>
</tbody>
</table>