An Evaluation of Discriminative Training for Hidden Markov Models in a Real-Environment Speech-Oriented Guidance System

Denis Babani*, Tomoki Toda*, Hiroshi Saruwatari*, Kiyohiro Shikano*
* Nara Institute of Science and Technology, Takayama 8916-5 Ikoma city, Nara Prefecture, Japan
E-mail: {babani-d.tomoki.sawatari.shikano}@is.naist.jp

Abstract—This paper presents experimental evaluations of Maximum Mutual Information discriminative training of the acoustic model in a real-environment speech-oriented guidance system "Takemaru-kun".

I. INTRODUCTION

We have developed a real environment speech-oriented guidance system, "Takemaru-kun" [1], based on SDS (Spoken Dialog System) techniques consisting of large-vocabulary continuous speech recognition (LVCSR), inquiry classification, and text-to-speech (TTS) components. Since the first day of operation, this system has been used as a source of spontaneous Japanese speech data.

In this work we investigate the benefits of using Maximum Mutual Information (MMI) training [2] using such real-environment, spontaneous speech data recorded by the "Takemaru-kun" system. The evaluation is done by changing various initial and training conditions, and then comparing the word accuracy rate. The experimental results show that MMI training yields significant improvements in word accuracy of "Takemaru-kun" system compared with the conventional Maximum Likelihood Estimation (MLE) training.

II. SPEECH-ORIENTED GUIDANCE SYSTEM "TAKEMARU-KUN"

"Takemaru-kun" is a speech-guidance system installed in North community center of Ikoma city, Nara prefecture, Japan. The purpose of this system is to handle queries related to the agent, general information and about surrounding area. Since the installation day, "Takemaru-kun" system has been recording user inquiries; however until now we have only the first two years completely transcribed by humans. These utterances have been labeled as speech, noise or partially speech and have been subjectively classified into five groups related to age of the speaker (i.e. preschool, lower grade school children, higher grade school children, adults and elderly persons).

Utterances recorded by "Takemaru-kun" are usually short in length. Being in a real environment, speech data usually is not clean, it contains background noise, and even speech overlapping between multiple speakers. For this reason, this corpus is adequate for evaluating MMI performance in real case scenarios.

III. EVALUATION

A. Experimental Conditions

Evaluations were conducted separately for each speaker group. A common dictionary with 58k words was created in order to have zero OOV (out of vocabulary) words for training utterances. Acoustic models were built from scratch for each speaker group using their corresponding training utterances. All acoustic models consisted of 3-state left-to-right triphone HMMs, with GMMs as output probability density. The acoustic feature vector was a 25-dimensional vector including ΔE (energy), 12 MFCC and 12 ΔMFCC.

<table>
<thead>
<tr>
<th>Group</th>
<th>MLE</th>
<th>MMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>75.36</td>
<td>77.45</td>
</tr>
<tr>
<td>Elderly</td>
<td>45.16</td>
<td>45.62</td>
</tr>
<tr>
<td>Preschool</td>
<td>44.16</td>
<td>45.46</td>
</tr>
<tr>
<td>Lower grade</td>
<td>61.11</td>
<td>63.82</td>
</tr>
<tr>
<td>Higher grade</td>
<td>67.14</td>
<td>68.75</td>
</tr>
<tr>
<td>Total</td>
<td>62.30</td>
<td>64.54</td>
</tr>
</tbody>
</table>

B. Evaluation Results

The first comparison was done by increasing language scale factor parameter, which gave significant improvements in word accuracy. This setting is responsible for including more acoustically competitive hypotheses in word lattices and improving the discriminability of the acoustic model. During this evaluation we observed that a 2-gram language model yielded better word accuracy than a 1-gram one. These results are not consistent with [2]. The type of language model used in lattice generation is strictly connected with generalization of MMI training. However, since the performance of the acoustic model in "Takemaru-kun" data is not high enough, stronger language constraints are still effective for improving quality of word lattices.

Results of changing some parameters in model update, such as I-smoothing parameter and the acoustic scale factor, showed no significant differences in word accuracy for all speaker groups. These results have also been observed in the other speaker groups.

Table 1 shows the increased word accuracy by MMI training from MLE training in every speaker group. We can observe that MMI training yields word accuracy improvements around 2% absolute in total.

IV. CONCLUSIONS

Based on results of this evaluation, we can say that MMI training gives significant improvements in word accuracy even in real environment spontaneous speech data.

V. ACKNOWLEDGMENT

The authors are grateful to Dr. Erik McDermott for the invaluable advices concerning discriminative training techniques.

REFERENCES


Proceedings of the Second APSIPA Annual Summit and Conference (Student Symposium), page 8, Biopolis, Singapore, 14-17 December 2010.