Development of Preschool Children Subsystem for ASR and Q&A in a Real-Environment Speech-oriented Guidance Task

Tobias Cincarek, Izumi Shindo, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science
Nara Institute of Science and Technology, Japan
{cincare-t,shikano}@is.naist.jp

Abstract

The development of a module for speech recognition and answer generation for preschool children for a speech-oriented guidance system is described. This topic requires extra treatment because the performance is still disproportionally low to children of higher age, there is a growing business demand and only relatively few research on preschool children ASR has been carried out. This is especially true for building practical applications. A real-environment speech database with more than 12,000 utterances of Japanese preschool children and more than 60,000 utterances of school children are employed for system development. The difference between preschool children’s and standard pronunciation is narrowed by introducing uniform reference transcriptions and pronunciation modeling. Furthermore, language and acoustic model are optimized. Final evaluation shows, that the speech-oriented guidance system’s response accuracy can be improved by more than 12% absolute.

Index Terms: preschool children, acoustic model, selective training, pronunciation modeling, Q&A

1. Introduction

Children have already been using video game devices, computers and cellular phones for a long period of time. The group of potential users is becoming younger and speech recognition technology has been started to be employed for such devices in recent years. However, difficulties especially regarding the recognition performance remain. The younger the user, the more difficult is the recognition of speech.

Previous research on children speech recognition has most often been mainly focusing on children of six or more years of age [1, 2, 3, 4], investigating speech characteristics [5, 6] and employing a method for vocal tract length normalization (VTLN) [7, 2, 3, 8, 9]. Differences between adult and children speech are shifts in fundamental and formant frequency which can be explained anatomically. VTLN is employed to reduce the difference between adult and children speech features for recognition with an adult acoustic model. If enough children speech data is available, construction of a child-specific acoustic model would be a different way overcome this problem. However, the main reason for a lower recognition performance is the unstable speech production process (articulation) of speech sounds by children. Spectral and cepstral characteristics of different realizations of the same speech sound of during a certain realization vary remarkably [5]. Consequently, in this paper as an additional approach, the number of possible pronunciations per word is increased in order to cope with different acoustic realizations of the same word.

There are only a few examples for building practical systems for children using speech recognition technology, e.g. [1, 10, 11]. Furthermore, only a few of previous investigations focus on spontaneous speech, e.g. [4], while most of the others are restricted to read speech, e.g. [2, 3, 6, 8]. Speech material employed for system development in this paper are from preschool children. Real-environment, spontaneous speech data was collected with the public speech-oriented guidance system "Takemaru-kun" [11, 12]. After describing the baseline system built with just a few months of real-environment spontaneous children speech data, the system is adapted for preschool children using real speech data which has been collected over two years. Besides more tolerant pronunciation modeling, optimization of the acoustic and the language model are considered.

2. Real-Environment Speech Database

Spontaneous Japanese children speech from the Takemaru database is employed. Takemaru-kun [11] is a speech-oriented dialogue system intended to provide the user information on the weather, news, the surrounding environment, public transportation system, Internet pages, etc. The system is very popular among children, because it is based on an animated character. It is a working system installed in a public place in Nara, Japan. The system collected more than 70,000 speech and noise inputs since November 2002. All data recorded during the first two years are completely transcribed, labeled with tags (e.g. noise) and classified subjectively into speaker groups: preschool children, elementary school children, junior-high school children, adults and elderly persons. Preschool and school children speech data employed for experiments is shown in Table 1. The evaluation data set was selected so that there is one randomly drawn utterance each for the most frequent 1,000 transcripts.

<table>
<thead>
<tr>
<th>Speaker Group / Data</th>
<th>Rel. Share</th>
<th># Utter.</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children Data (P+L+H)</td>
<td>78,824</td>
<td></td>
<td>42.7 hrs</td>
</tr>
<tr>
<td>(P)Preschool Children</td>
<td>14.4%</td>
<td>11,366</td>
<td>6.3 hrs</td>
</tr>
<tr>
<td>(L)Lower Grade Children</td>
<td>64.2%</td>
<td>50,635</td>
<td>28.5 hrs</td>
</tr>
<tr>
<td>(H)Higher Grade Children</td>
<td>21.3%</td>
<td>16,823</td>
<td>8.9 hrs</td>
</tr>
<tr>
<td>Evaluation (Preschool)</td>
<td>1,000</td>
<td>0.6 hrs</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Real-environment, spontaneous speech data collected with the speech-oriented guidance system Takemaru-kun were employed for preschool children experiments.
2.1. Preschool Children Data

The preschool children speech data is remarkably different from the speech data of lower grade and higher grade school children. In a preliminary evaluation experiment the recognition performance was 76.9% for higher grade school children and 75.2% for lower grade school children, but only 45.5% for preschool children even when employing an acoustic model trained on collected data of each corresponding speaker group. Different pronunciation of words which appear randomly, a lower speaking rate and differences in sentence construction and word usage are the three most prevalent characteristics. There are also differences among preschool children, since they usually grow up in a different environment (different parents).

3. Baseline System

The major components of the speech-oriented guidance system are a speech recognizer and a response generation module as shown in Figure 1. The acoustic model is a context-dependent phonetic-tied mixture (PTM) model. The initial model is constructed from the JNAS database [13] and then retrained with all available children speech data (cf. Table 1). The baseline language model has been built from one million sentences of text data collected from the web, transcriptions of about 17,000 children utterances and about 6,000 questions devised by humans. The vocabulary size is about 42k and the pronunciation dictionary contains only one possible phoneme sequence per word. Other experimental conditions are uniform. They are given in Table 2.

4. System Optimization

The baseline system has been built using relatively few data after the system has been installed in a public place. In the following it is shown how the system performance can be optimized for preschool children, when two years of real-environment children speech data become available.

4.1. Acoustic Modeling

Several acoustic models built by retraining the baseline model on different children speech data sets are considered. The result is given in Table 3. The best performance is naturally obtained for the matched case (1), i.e., when training the initial model with preschool children data. When employing data from lower grade (2) or higher grade school children (3) the recognition performance degrades by about 5% and 20% absolute, respectively. Adding lower (4) and higher (5) grade data to the preschool children data, performance improves to 51.9% (7), which is 2-3% higher than when using only preschool children data (6) and 5-7% higher than when employing the whole children speech data pool (5) for training.

4.2. Language Modeling

The baseline language model is trained on web data and transcriptions of several thousand children utterances. The performance of this model is compared to five language models which are trained on transcriptions of children utterances only. Acoustic model (7) from the previous section is employed for this
investigation. From Table 4 it is clear that, the transcription-based language models (2)-(6) perform better than the web-based baseline model (1). It has a positive effect when adding lower grade (3) and higher grade school children data (4) to the preschool children data (2) for model training. The reason is likely to be the fact, that in contrast to the lower and higher grade data, the preschool children speech transcripts contain non-standard words. Consequently, a language model (5) trained on normalized preschool children data transcripts was examined finally which yielded the best performance. Here, adding lower grade (6) data to (5) did not improve the performance further.

### 4.3. Pronunciation Modeling

The purpose of pronunciation modeling is to be able to recognize the normalized form of words although their pronunciation has been altered by the preschool children in an idiosyncratic way. For example, preschool children pronounce 'Takemaru' as 'Tachimaru', 'Takebaru' or 'Takemau'. Nevertheless, all three pronunciation variants should be recognized as 'Takemaru'. Therefore, it is proposed to add these pronunciation variants for the word 'Takemaru' in the pronunciation dictionary used during recognition.

In order to understand the idea of the proposed approach in more detail, the reader should be aware, that in the Japanese language there is a one-to-one correspondence between graphemes and speech sounds, i.e. a bijective mapping between Japanese characters and phoneme pairs can be established. Except a few Japanese characters for the 'N' sound and the five vowels 'a', 'i', 'u', 'e', 'o', each character of the Hiragana/Katakana writing system is pronounced as a consonant followed by a vowel. Hiragana/Katakana are the Japanese "alphabets" by which any character, the so-called Kanji to remove meaning ambiguities. In this paper pronunciation modeling is considered at the Hiragana/Katakana level. The preschool children training data is transcribed in two ways:

1. Transcription of how words were actually pronounced
2. Transcription of each word's standard pronunciation

Context-sensitive pronunciation mappings are obtained from the forced alignment of the transcriptions (1) and (2). There are substitution (3 \to 3 characters), deletion (3 \to 2 characters) and insertion (2 \to 3 characters) rules. After discarding rules with an absolute occurrence frequency of one, there were 1,295 substitution, 57 insertion and 35 deletion rules. Extra substitution rules have been derived for word begin and end by ignoring either the left or the right context, respectively. Several examples are given in Table 5. Since readers may be unfamiliar with the Japanese writing system, only phoneme sequences are shown.

The source part of each mapping is compared with every pronunciation dictionary showing the effectiveness of pronunciation modeling. The result is shown in Table 5.

### Table 5: Examples for pronunciation mapping rules.

<table>
<thead>
<tr>
<th>Type</th>
<th>Source</th>
<th>Target</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>sub</td>
<td>d e s u k a</td>
<td>d e s h u k a</td>
<td>303</td>
</tr>
<tr>
<td>sub</td>
<td>d e s u k a</td>
<td>d e c h u k a</td>
<td>162</td>
</tr>
<tr>
<td>ins</td>
<td>N</td>
<td>c h i</td>
<td>N</td>
</tr>
<tr>
<td>sub</td>
<td>N s a i</td>
<td>N sh a i</td>
<td>46</td>
</tr>
<tr>
<td>del</td>
<td>b a</td>
<td>b a - i</td>
<td>43</td>
</tr>
<tr>
<td>sub</td>
<td>d e s u k a</td>
<td>d e q</td>
<td>k a</td>
</tr>
</tbody>
</table>

### 4.4. Response Accuracy

Since the application considered in this paper is a guidance system, it is imperative to evaluate the system’s response accuracy in order to see whether an improvement in speech recognition accuracy transfers to a increase in response accuracy. Response generation is based on a question and answer data base (QADB), which contains Q&A pairs for approx. 300 responses. The questions are taken from the transcriptions of the collected speech data, the corresponding answers are assigned manually by humans. Three QADBs are considered here: (Presch) based on the collected preschool children data, (Child) built from all available children data and (Closed) which contains only the Q&A pairs corresponding to the transcription of the test data. The evaluation result is shown in Table 7. Updating either the language or the acoustic model has equally high positive effects.
Table 7: Response accuracy for different combinations of acoustic model, language model and pronunciation dictionary.

<table>
<thead>
<tr>
<th>System Configuration</th>
<th>Response Accuracy [%] (Presch)</th>
<th>Q&amp;A Database (Child)</th>
<th>(Closed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline System</td>
<td>40.3</td>
<td>38.5</td>
<td>49.1</td>
</tr>
<tr>
<td>AM(7) Update</td>
<td>45.2</td>
<td>43.8</td>
<td>54.5</td>
</tr>
<tr>
<td>LM(5) Update</td>
<td>45.8</td>
<td>46.2</td>
<td>57.1</td>
</tr>
<tr>
<td>LM(5)+Dict. Update</td>
<td>47.8</td>
<td>46.9</td>
<td>56.5</td>
</tr>
<tr>
<td>LM(5)+AM(7) Update</td>
<td>53.2</td>
<td>51.3</td>
<td>61.6</td>
</tr>
<tr>
<td>LM(5)+AM(7)+Dict.</td>
<td>53.4</td>
<td>51.2</td>
<td>61.7</td>
</tr>
</tbody>
</table>

in improving the system’s response accuracy. When updating both models, the absolute improvement of their independent update seem to add up. Editing the pronunciation dictionary has a positive effect, but it is less promising than optimizing the acoustic model by training with normalized utterance transcriptions and selecting additional training utterances from lower and higher grade school children data. There were no further improvements when updating acoustic model and pronunciation dictionary at the same time.

5. Conclusion

In this paper automatic speech recognition and response generation for a real-environment speech-oriented guidance system has been investigated. For improving speech recognition accuracy of preschool children, acoustic, language and pronunciation modeling are considered. For language modeling, a model trained on the normalized transcriptions of more than ten thousand preschool children utterances yielded the highest performance. A previously proposed method for selective training of the acoustic model was applied successfully for augmenting the existing preschool children training data with an appropriate subset of the school children data. Context-dependent pronunciation modeling applied at the Japanese ‘syllable’ level was less promising. Finally, the response accuracy of the system was evaluated. Improvements in recognition accuracy transferred in the same degree to an improvement in response accuracy. In the end a response accuracy of more than 60% was achieved.

6. Acknowledgments

A part of this work is supported by the MEXT e-Society project, Japan.

7. References