A MEDICAL REHABILITATION DIAGNOSES TRANSCRIPTION METHOD THAT INTEGRATES CONTINUOUS AND ISOLATED WORD RECOGNITION

Shoichi Matsunaga, Yoshiaki Noda, Katsutoshi Ohtsuki, Eiji Doi, Tomio Itoh*

NTT Cyber Space Labs., and Kameda General Hospital*  
1-1, Hikari-no-oka, Yokosuka-shi, Kanagawa 239-0847 Japan  
1344, Higashi-machi, Kamogawa-shi, Chiba 296-8602 Japan*  
E-mail: mat@ntspch.hil.ntt.co.jp

ABSTRACT
This paper describes a practical dictation system that is able to compensate for the lack of a large text database and reports on the results of field tests in which the system was used to make medical rehabilitation diagnosis reports in a hospital. Our dictation system uses two recognition engines: continuous speech recognition and isolated word (including connected words) recognition engines. When a user makes a report, the two recognition engines can be selectively operated in a single window through voice commands or mouse click. The system operates on common personal computers under the Windows O.S. A field test conducted in noisy therapeutic work rooms that compared the performance of speech input system comparing to conventional keyboard input system for the medical rehabilitation field, demonstrated the effectiveness of the dictation system.  

1. INTRODUCTION
More and more doctors are using electronic medical-chart systems for their diagnoses, and speech recognition would eliminate many troublesome input operations that currently have to be done by keyboard. Faster diagnoses descriptions would also lead to significant cost reductions. Trials have already been conducted on automatic transcription of diagnoses in the field of X-ray CT[1, 2].  

Current state-of-art dictation approaches use elaborate language models generated by huge text data [3]. For the speech recognition task of a Japanese newspaper, for example, a lot of articles over a five-year period were used[4]. In the medical field, however, there are many cases where such a large database is not available. Furthermore, the reports include not only patient and illness descriptions but also tables and charts. Conventional approaches based on restriction higher-order n-gram language models tend to perform poorly in these respects, and therefore, a more efficient and practical dictation system is needed.

To address the above problems, we are designing a practical dictation system that uses continuous speech and word recognition for medical rehabilitation diagnoses transcription in a medical-chart system, which is a typical task performed by rehabilitation therapists. The continuous speech recognition engine, which uses a stochastic language model, is designed to recognize sentences relating to diagnoses, and the word recognition engine is designed to recognize very-low-frequency proper nouns, abbreviated phrases, and items in the tables. This hybrid architecture’s dictation capability is more efficient than those of conventional systems that use only continuous speech recognition. It has functions of noise adaptation, speaker adaptation, an expandable vocabulary, and word look-up lexicons, and the user can also select the correct transcription from the N-best candidates. The system operates on common personal computers under a Windows O.S. It has been field tested in a therapeutic work room at Kameda General Hospital, where it was used to make rehabilitation reports for an electronic medical chart system.

In this paper, we evaluate the speech input function of this system by comparing it with a conventional keyboard input in a medical rehabilitation environment. Some factors that affect the time it takes to make the report, and others that tend to decrease recognition performance are also discussed.

2. MEDICAL REHABILITATION AND DIAGNOSES
The aim of medical rehabilitation is the patient’s recovery from impairment. After each rehabilitation examination in the hospital, a therapist makes a medical report on the examination and progress of the patient’s recovery. There are three main rehabilitation therapies at Kameda Hospital, each with its own type of specialist: physiotherapy (PT), occupational therapy (OT), and speech therapy (ST). PT is a treatment involving exercises such as stretching. In OT, a patient recovers from his or her impairment by doing interesting tasks, such as tying string or sanding. The major affliction treated by ST is aphasia. The medical reports written by the
therapists are highly itemized, and this characteristic makes them very different from newspaper articles or transcriptions of broadcast news. Physical rehabilitation reports, in particular, contain many tables filled with figures that indicate the degree of a patient's movement, such as range of motion (ROM) or manual muscle test (MMT).

Our target is the creation of a useful speech recognition transcription system for these rehabilitation reports.

3. SYSTEM CONFIGURATION

The system configuration of the rehabilitation diagnoses transcription system is shown in Figure 1. The system operates on common personal computers under the Windows O.S. and can take speech, keyboard and mouse input. Speech input is intended to be the main input method. The speech recognition prototype (REX), which was developed by NTT Cyber Space Labs., is highly sophisticated and can be used for various purposes (dictation with stochastic language model, or with transition grammar based on BNF description, or word recognition[5]). It can operate on various computer platforms.

To compensate the small size of the text database, the system adopts word recognition as well as continuous speech recognition. The continuous speech recognition engine, which uses a stochastic language model and a word lexicon, is designed to recognize sentences and phrases relating to diagnoses, and the word recognition engine is designed to recognize very low-frequency words and items that occur in the tables of the diagnoses. The word recognition engine has seven lexicons (names of therapists, doctors, diseases, and so on). The two recognition engines can be selectively operated in a single window through voice commands or mouse. The system accepts voice commands such as “cancel the last utterance”, “stop voice input”, “write on next line”, etc., and can be easily applied to other tasks by changing the word lexicons and the language model.

3.1. Language Model

Many low-frequency words must be able to be recognized by both engines because strict division of the vocabulary makes it harder for uncustomed users to use the system. To set the language model and lexicons, morphological analysis was carried out using about 500 Japanese diagnosis reports; 60% were PT related, 30% were OT related, and 10% were ST related. The amount of the training data was almost equal to that of a Japanese newspaper. The main aim of this analysis was to detect the word boundaries because a Japanese sentence is written without spacing between words. From the analysis, we ended up with about 4700 different words. All words, except for incorrect words generated from incorrect analyses, personal names and very-low-frequency words, were registered in the lexicon of the continuous speech recognition engine. To compensate for the small text database and the therapists’ punctuated utterances while diagnosing, we decided that a word bigram model was appropriate for the system.

3.2. Acoustic Model and Adaptation

Gender-dependent speaker-independent phoneme hidden Markov models (HMMs) and speaker-dependent HMMs for specified therapists were generated. Each model consisted of both context-dependent (triphone) and context-independent (monophone) models and was generated using speech data from phoneme-balanced sentences and technical sentences extracted from PT diagnoses reports.

A supervised speaker-adaptation function was also included in the system; the training sentences were the most frequently occurring sentences, from 10 to 70 depending on the user, and were extracted from three kinds of therapeutic reports. Users were able
Table 1: Word coverage and perplexity for each lexicon

<table>
<thead>
<tr>
<th>lexicon</th>
<th>coverage(%)</th>
<th>perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>690 words</td>
<td>90.4</td>
<td>16.4</td>
</tr>
<tr>
<td>1011 words</td>
<td>93.9</td>
<td>16.7</td>
</tr>
<tr>
<td>2173 words</td>
<td>98.1</td>
<td>16.9</td>
</tr>
</tbody>
</table>

to select the desired acoustic model from among the prepared gender-dependent speaker-independent and speaker-dependent models, and they could activate the adaptation function to generate a better model.

3.3. N-best Candidates

When there are multiple high-likelihood candidates for each utterance, the editor-window displays the candidate with best likelihood as a blue character sequence. If a user points to a part of this string, the N-best candidates appear. The user can listen to the utterance to determine whether there is a correct candidate among them. After selection of the correct candidate, the editor-window displays it black. If there are no correct strings in the N-best candidates, the user must make another utterance or input it by keyboard.

4. DICTATION OF REHABILITATION DIAGNOSES

Dictation experiments to evaluate the performance of our system were carried out using utterances spoken by therapists and staff. Two test sets were prepared; test set A consisted of 100 PT diagnoses sentences uttered by a therapist and two staff in a quiet room. Test set B consisted of one PT diagnosis report and one ST diagnosis report uttered freely in a noisy therapeutic work room. Each utterance was composed of therapeutic sentences, isolated words, and voice commands uttered through a head-set microphone.

Three dictation systems, whose respective vocabularies in the continuous speech recognition engine were 690, 1011, and 2173, served for this test. The lexicons of the word recognition engine were the same in each system, consisting of about 550 words in all. The five-best candidates were generated in these systems.

4.1. Preliminary Test

The word coverage and perplexity of test-set A for each lexicon is shown in Table 1. In calculation of the perplexity, word pairs including out-of-vocabulary words were disregarded.

Each HMM of the acoustic models had Gaussian distributions in continuous mixture density. Speaker-dependent models, each having four mixtures per state, and two types of speaker-independent gender-dependent models (8-mixture triphones and 16-mixture monophones (model-1), and 4-mixture triphones and monophones (model-2)) were prepared. Acoustic features were 16-th order LPC cepstra, delta-cepstra and a delta-power.

The word accuracies were 92.3% (690-word), 92.7% (1011-word), and 91.9% (2173-word) when model-1 was used. Model-2's performance was decreased at the expense of processing time for recognition. Recognition accuracy and processing time ratio are shown in Table 2, where processing time ratio in the 690-word system of model-1 is set to 1.

When speaker-dependent acoustic models were used in the 2173-word system, a 95.3% word accuracy was achieved. These results show that a 2173-word vocabulary is sufficient for practical use.

4.2. Field Test

To analyze the usability of our system, a field test was carried out in noisy therapeutic work rooms. The use of functions of input-gain adjustment, speaker-model selection, speaker adaptation, and noise adaptation were optional, but the subjects were told that use of these functions before the dictation procedure was highly recommended for they would produce better transcription results.

In the beginning stage of the experiments, all subjects were unfamiliar with speech recognition. They did not pay attention to selection of acoustic models and were concerned about the limited word coverage. Most subjects did not use the speaker-adaptation function even though it could yield better performance because they were forced to utter more than ten sentences for adaptation. In the latter stage of the experiments (three months later), recognition accuracy had increased as therapists had learned to use the word recognition function in the parts of the reports where it should be used. Test-set B was a collection of speech data from the latter stage. The subjects used the 2173-word-vocabulary system and their speaker-adapted models trained from model-2.

PT data in this set was uttered by a skilled physical therapist (male), who had used this system several times before. The ST data was uttered by a skilled speech therapist (female) who had no experience with this system but had the ability of blind keyboard typing. About three percent of the utterances were recognized by the word recognition engine, and the rest were recognized by the continuous speech recognition engine. The PT report contained an equal amount of sentences and tables, whereas, the ST report was mostly sentences. For further comparison, another subject, who was not a therapist but who was familiar with speech recognition, uttered both reports. Input speed (characters per second) for each report is shown in Table 3 and Table 4.
Table 2: Word accuracies and processing time ratio for speaker-independent models

<table>
<thead>
<tr>
<th></th>
<th>acoustic model</th>
<th>model-1</th>
<th>model-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>lexicon</td>
<td>accuracy(%)</td>
<td>proc.</td>
<td>accuracy(%)</td>
</tr>
<tr>
<td>690 words</td>
<td>92.7</td>
<td>1</td>
<td>85.0</td>
</tr>
<tr>
<td>1011 words</td>
<td>92.7</td>
<td>1.1</td>
<td>87.4</td>
</tr>
<tr>
<td>2175 words</td>
<td>91.9</td>
<td>1.3</td>
<td>85.0</td>
</tr>
</tbody>
</table>

Table 3: Input speed for PT reports (characters/sec)

<table>
<thead>
<tr>
<th>subject</th>
<th>speech</th>
<th>keyboard</th>
</tr>
</thead>
<tbody>
<tr>
<td>physical therapist</td>
<td>0.31</td>
<td>0.38</td>
</tr>
<tr>
<td>non-therapist</td>
<td>0.44</td>
<td>0.34</td>
</tr>
</tbody>
</table>

Table 4: Input speed for ST reports (characters/sec)

<table>
<thead>
<tr>
<th>subject</th>
<th>speech</th>
<th>keyboard</th>
<th>keyboard*</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech therapist</td>
<td>0.42</td>
<td>0.98</td>
<td>0.23</td>
</tr>
<tr>
<td>non-therapist</td>
<td>0.65</td>
<td>0.57</td>
<td>–</td>
</tr>
</tbody>
</table>

In both experiments, the therapists’ keyboard input speed was higher than their speech input speed. This is because both therapists were still unfamiliar with the system and familiar with the typing of technical terms. In contrast, the non-therapist’s speech input was superior to keyboard input. Note that in the PT dictation task, the speech input of the non-therapist was faster than the physical therapist’s keyboard input. Even though the number of subjects was small, this result clearly shows the usefulness of the system. In the ST dictation task, the keyboard input speed of therapist was excellent because of her blind typing ability. To erase the influence of her skill, another input test, using a keyboard that had a different character arrangement, was carried out. The entry in Table 4 that is marked with an asterisk indicates that voice input speed was superior to this keyboard input speed, which shows the promising ability of the system.

The main causes of recognition errors in the field tests were: overflow of input gain, utterance out of rehabilitation task vocabulary, the use of inappropriate acoustic models, inappropriate utterances such as with laughing, and loud speech near the subject when he or she was using the system.

The most time consuming problem was subject’s repeating of the utterance to get correct results when there were no correct words in the lexicon of the continuous engine (utterance of out-of-vocabulary words). If the subject did not change the input method to word recognition mode or keyboard input, inputs had to be repeated several times. Experience in using the system was needed to overcome this problem.

5. CONCLUSION

This paper described a practical dictation system for medical rehabilitation diagnosis reports that is able to compensate for the lack of a large text database and reported on the results of field tests. Our dictation system used two recognition engines: continuous speech recognition and isolated word (including connected words) recognition engines. To make a report, both recognition engines can be selectively operated in a single window. Experimental results demonstrate the potential utility of the system, however, large-scale recognition experiments, including detailed analyses, are necessary. We think that appropriate usage of both keyboard and voice inputs will enable therapists to make faster reports after they become familiar with the system.

6. ACKNOWLEDGMENTS

The authors are especially grateful to Dr. Toshitada Kamada, chairman of Kameda General Hospital, who permitted our experiments in the hospital. We wish to thank the aid of Mr. Yoshiaki Yoshida, Mr. Tetsuya Sakino, at Kameda Medical Information Laboratory, and Mr. Tomokazu Yamada, Mr. Takatoshi Jitsuhiro, at NTT Cyber Space Laboratories.

REFERENCES