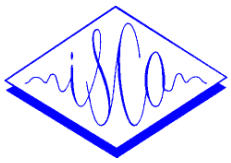

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

A Prototype for a Speech-to-Text Transcription System

Toshiaki Tsuboi and Noboru Sugamura

Nippon Telegraph and Telephone Corporation
Human Interface Laboratories
Yokosuka-shi, KANAGAWA, 238-03 JAPAN

ABSTRACT

A prototype for a speech-to-text transcription system is described. This system recognizes continuous phrasal speech and transcribes it in Japanese text. This paper outlines methods for acoustic and linguistic processing, and describes the system configuration and results of performance evaluation tests. As a text is spoken phrase by phrase, it is recognized by a word-spotting method using a continuous dynamic programming technique. High frequency words and CVs in continuous phrasal speech are detected using established CV and word templates. The CV and word candidates are converted to phrase candidates using a word dictionary, inflection table, post positional word dictionary, compound word table, and phrase syntactic pattern table. Frequent phrase co-occurrence patterns are used to select feasible phrase candidates. A performance evaluation test is carried out for Japanese X-ray CT scanning reports. Conversion accuracies of 80% and 65% are obtained for normal and abnormal medical findings, at input speeds of 100 chinese characters/minute and 50 chinese characters/minute respectively. These input speeds equal those of a professional transcriber and a novice transcriber after 20 days of training.

INTRODUCTION

An effective speech recognition system is needed to remove the current bottleneck in information processing systems: the manual input of text. Moreover, such a system is a necessity in situations that require the user's hands or eyes to be used for other purposes. Since Japanese consists of about 100 Consonant-Vowel syllables (CVs), each of which corresponds to a kana character, isolated CVs are the input units in the Japanese speech recognition systems proposed to date [1]. However, it is difficult to utter isolated CVs and input speed is slow. Additionally, in contrast to written English, written Japanese does not consist of simple word sequences. Since a number of affixes, inflections and post positional words, which are particles and auxiliary verbs, can be tightly associated with words so that they are difficult to divide, it is impossible to utter the components separately. Therefore, text input by large vocabulary word recognition is not suitable for Japanese as it is for English [2],[3]. We have, therefore, developed a prototype Japanese text input system that recognizes continuous phrasal speech. Speaking in phrases is not completely natural, but it is, obviously, much easier and faster than speaking in isolated CV units.

Instead of aiming at a general word-processor having a comprehensive vocabulary, our prototype system addressed a specific field: medical reports from a doctor analyzing X-ray CT scans. The CT scan analysis requires the doctor's complete attention. Our prototype system allows medical reports to be transcribed automatically.

This paper describes system concepts and its configuration, methods for acoustic and linguistic processing, hardware architecture, and results of a performance evaluation test.

System Concepts

Recognition units

Since a phrase consists of words, affixes, inflections and post positional words, a large number of phrase variations are possible. These cannot be recognized using phrase templates. Word templates are useful for detecting key words in the phrase, since high frequency words cover most words in the documents of a specific field. The remaining components are recognized by CV templates. The combination of these two templates effectively recognizes continuous phrasal speech [4].

Template training

Usually, a speech recognition system requires training with a complete vocabulary. However, it is tedious to read out thousands of words. Therefore, we developed a training method that uses speech data uttered during regular text input. Initially, the user utters only 118 syllables and high-frequency words for template generation. As regular text is input and corrected, the training method uses corrected phrases to enhance the effectiveness of the stored CV templates [5].

Voice storage

There are several types of speech-to-text transcription systems. Usually, the recognition results must be corrected immediately after each utterance. In these systems, only the speaker can effectively correct the text. Our system digitizes the speech and stores it together with the transcription. Thus anyone can correct the transcribed text later at any other networked site hearing the stored speech.

Selection and correction of recognition results

There are two ways of correcting erroneous phrases. One is by speech and the other is with a keyboard. Respeaking the phrase seems to be simpler but the system may not respond correctly. Our method uses keyboard or mouse commands to indicate the incorrect phrase and to precisely input or select the correct phrase.

SYSTEM CONFIGURATION

The system consists of two main parts: the speech recognition equipment which uses specially designed hardware; and an ordinary MC68030-based Unix workstation. The system configuration is shown in Fig.1.

The recognition equipment generates CV and word lattices from continuous phrasal speech with the spotting method which is based on a continuous dynamic programming (DP) technique [6]. Initially, word templates and multiple CV templates are extracted from initial training speech samples to cover utterance variations. A few hundred words, which are the most often used in the intended task, are selected and uttered by the speaker. Using these words, multiple CV syllable templates are automatically generated based on the contents of the words -- i.e., combinations of CV syllables [7]. CVs and high-frequency words in continuous speech are detected using CV and word templates. CV and word-spotting results are sorted according to the degree of similarity between the templates and the input. The most probable boundaries between CVs are determined by the

global spectral deviation using LPC cepstral parameters. CV and word lattices are generated by combining CV and word spotting results and selected CV boundaries [4],[7].

The workstation is equipped with a word dictionary, an inflection table, a post positional word dictionary, a compound word table, and a phrase syntactic pattern table. It uses the CV and word lattices generated by the recognition equipment to produce phrase translation candidates [8]. Next, if one pair of adjacent candidate phrases matches with a frequent phrase co-occurrence pattern, such phrase candidates take precedence over other candidates. Finally, the ten highest scoring phrase candidates are filed, and the system adds the most feasible one to the transcription. After the entire text is input orally, recognition results are corrected by listening to the digitized speech. If a phrase conversion result is incorrect, the right one from the ten phrase candidates is selected, or corrected with the keyboard [9]. Through this correction technique, the system identifies the speech content and a new CV lattice is re-calculated using segmentation. CV lattices and correct CVs are compared. Based on this comparison, consistently weak templates are eliminated and more reliable CVs are extracted from the input speech and added to the CV templates. Stored CV templates are automatically updated as necessary to maximize the accuracy of CV recognition [5].

ACOUSTIC PROCESSING

In the acoustic processing stage, continuous phrasal speech is recognized by the word spotting method that uses continuous DP. First, since Japanese is composed of about 100 CV syllables, CV syllables in continuous speech are detected using CV templates. In each input phrase, a large number of CV syllables are inaccurately detected. Generating word candidates is very complicated, since the number of CV syllable combinations is enormous. Therefore, segmentation positions are utilized as supplementary information. Segmentation positions are determined by global spectral deviations using LPC cepstrum parameters. Some CV syllable candidates can be extracted using CV syllable locations in a segment, so meaningless CV syllable combinations are eliminated. The CV lattice is thus obtained.

Each field of writing has its own specialized vocabulary. Many words are used sparingly while a few are used often. The high frequency words are collected to form the word templates for the field. The templates allow words in continuous speech to be detected. The CV and word lattices are generated as results of the above processes. Duration and vowel candidates for each segment are added to the CV lattice.

LINGUISTIC PROCESSING

In the linguistic processing stage, the two lattices are converted into text as follows:

(A) All CV syllable combinations produced from the CV lattices are matched to a word dictionary.

(B) The word candidates from the word lattice are evaluated in the CV syllables' recognition order in the CV lattice.

(C) The word candidates from the CV and word lattices are checked grammatically using an inflection table, post positional word dictionary, compound word table, and phrase syntactic pattern table.

(D) An evaluation function is used to select the most feasible translation result. Each phrase candidate is also evaluated using the frequent phrase co-occurrence patterns.

The details of these processes are described below.

A. Word Matching

All CV syllable combinations produced from the CV lattices are matched against a word dictionary. Word combinations that perfectly match some dictionary items and those that perfectly match except for one or two syllables, are selected as word candidates. Therefore, even if there are deletion, insertion and substitution errors, correct word candidates can be extracted. Because a phrase may contain compound words, word extraction continues to the end of the CV matrices. After word matching, the evaluation value of each word candidate is calculated using recognition order and a penalty for unmatched syllables.

B.Evaluation of Word Spotting Results

As described above, an evaluation value is attached to each word candidate obtained from the CV lattice. To compare each word candidate from the CV lattice with those obtained from the

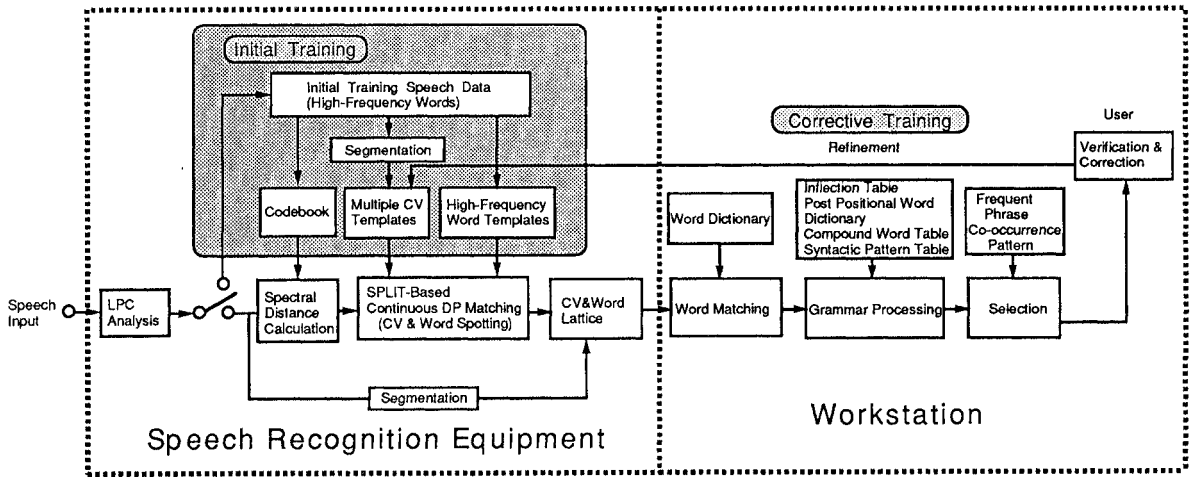

Fig.1 A Speech-to-Text Transcription System Using Continuous Speech Recognition

word lattice on an equal basis, evaluation values must also be attached to the latter candidates. The evaluation process is described below.

Word spotting results are compared to the CV lattices for matching using the value of the matched syllables in the CV lattice to calculate the evaluation values. Unmatched syllables are weighted with a pre-determined penalty. The evaluation value of a word is defined as the sum of its evaluation values and penalties.

C. Grammar Processing

CV syllable combinations extracted from the CV lattice are matched against an inflection table and post positional word dictionary. Post positional words consist of particles and auxiliary verbs. The compound word table and the phrase syntactic pattern table is used to extract only grammatically correct combinations.

D. Selection

An evaluation score is calculated for all words in each phrase candidate. Each phrase candidate is also evaluated with frequent phrase co-occurrence patterns. The evaluation score of a phrase candidate is the sum of the scores of its constituents. Finally the phrase candidate with the highest evaluation value is selected as the most feasible.

If a pair of adjacent phrase candidates is matched with a frequent phrase co-occurrence pattern, the evaluation score of such phrase candidates is increased. It takes precedence over other phrase candidates.

HARDWARE ARCHITECTURE

The hardware architecture of the recognition equipment is shown in Fig.2. A commercially available DSP is used to carry out the LPC analysis and spectral distance calculation in real-time. The controller is a MC68030 CPU. Together with system control, the controller extracts segmentation position and generates CV and word lattices using the CV and word spotting results from six processing elements(PE). A PE mainly consists of a DMA controller, local memory(LM), and 4 gate array LSIs designed specially. Each PE spots 250 CVs or words in parallel. The matching distance calculation for 250 CVs or words is performed within one frame period for real-time recognition. The CV and word lattices are transfered through a SCSI interface to the host computer.

PERFORMANCE TEST

A performance evaluation test was carried out for X-ray CT scanning reports. The dictionaries, tables, and co-occurrence patterns were created by analyzing 1,400 scanning reports which contained a total of about 70,000 phrases. At the time of the tests, the word dictionary had about 2,200 entries, while entries of the stored compound word table, phrase syntactic pattern table, and phrase co-occurrence patterns totaled about 1,700, 1200, and 5,000 respectively. Since 600 highest frequency words cover more than 90% of the words in reports, the number of word spotting templates was set at 600.

One male and female speaker uttered 30 CT scanning reports at their own speed. 20 of the reports had been used to generate the stored patterns. The reports were divided to two types: 15 normal and 15 abnormal medical findings. Each type of report contained about 25 and 50 phrases respectively. The reports were processed by the prototype system (corrective training was not used) and by a human transcriber.

The performance evaluation results are shown in Fig.3. The initial accuracy (Japanese conversion accuracy: correct chinese characters/all chinese characters) of the prototype system varied from 60 to 95% for the normal findings and from 50 to 85% for the abnormal findings. The average conversion accuracy was 80% for normal findings and 65% for abnormal

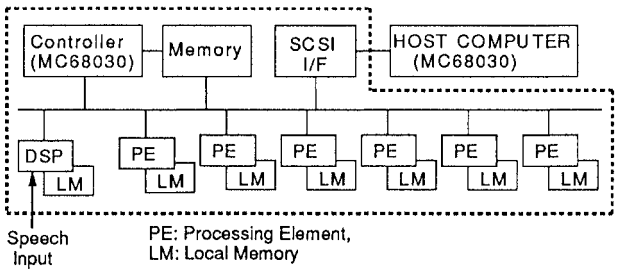

Fig.2 Hardware Architecture of Recognition Equipment

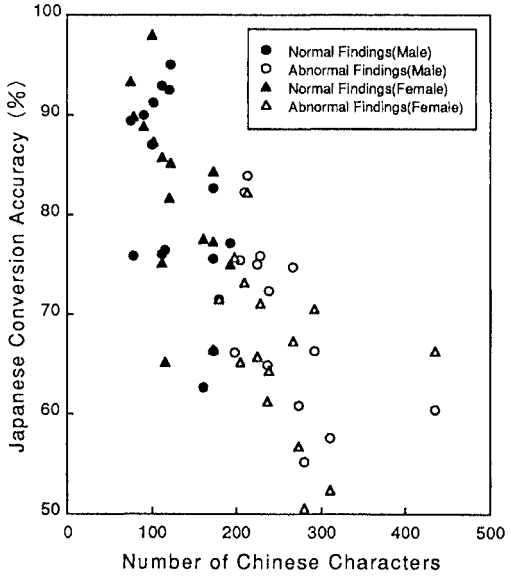

Fig.3 Performance Evaluation Results of Two Speakers

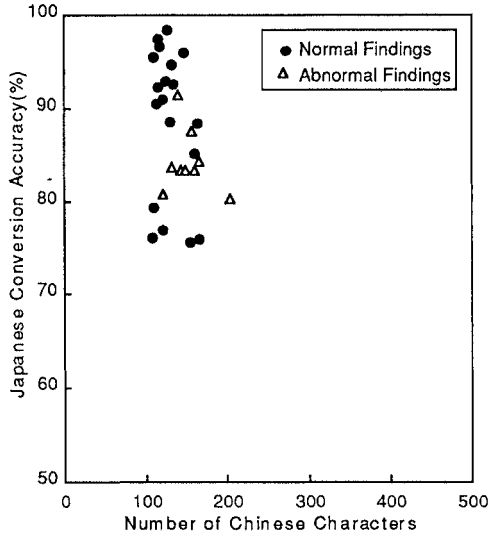

Fig.4 Performance Evaluation Results of a Doctor

findings. The average conversion accuracy was 87% and 75% within the first 10 candidates, respectively. The number of chinese characters varied from 80 to 200 for the normal findings and from 180 to 450 for abnormal findings. Abnormal findings are more complex, so the number of chinese characters is larger. The conversion accuracy is related to the number of chinese characters. A doctor used the system to transcribe 19 normal and 9 abnormal findings (all findings were new). The evaluation of conversion accuracy is shown in Fig.4. His result broadly mirror those of the two speakers.

Fig.5 shows the relation between conversion accuracy and correction speed for the female speaker. Correction of the machine transcription lead to an effective input speed of 60-220 chinese characters/minute for the normal findings and 40-70 chinese characters/minute for the abnormal findings. Input speeds of the professional transcriber, who had 2 years' experience, were 100 and 85 chinese characters/minute for normal and abnormal findings. Input speed for normal findings compare well with that of the professional transcriber. Input speed for abnormal findings compare well with the human transcriber who, after 20 days of training, reached a corresponding input speed of 50 chinese characters/minute.

The effect of iterative training was tested for the female speaker. Japanese conversion accuracy was improved by 2% using 5 normal and 5 abnormal findings. One reason for this limited improvement is that after adding new CV templates from input speech the eliminating process was not performed. Therefore, performance degrading templates remained.

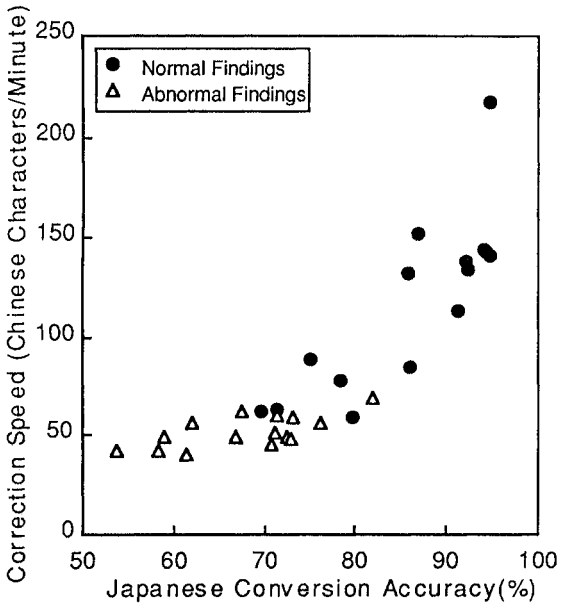

Fig.5 Relation Between Conversion Accuracy and Correction Speed

Conclusion

A prototype of a speech-to-text transcription system has been described. This system recognizes and transcribes continuous Japanese phrasal speech. Average conversion accuracies of 80% and 65% were obtained for normal and abnormal X-ray CT reports, respectively even though the iterative template training system was not employed. Effective input speed for normal findings was almost equal that of a professional transcriber. Tests of the time and fatigue benefits over keyboard transcription methods are currently being conducted at a hospital.

Acknowledgments

The authors would like to express their gratitude to Dr. Tadayuki Maehara, Kanto Teishin Hospital, for providing samples of reports. They would like to thank Dr. Ryohei Nakatsu in NTT Basic Research Laboratories and Mr. Fumihiko Obashi in NTT Intelligent Technology Co. for their useful discussions and important comments. They would like to thank Dr. Hirokazu Sato and Dr. Sadaoki Furui in the Speech and Acoustics Laboratory, NTT for their guidance. They also would like to thank Miss Miyuki Tsunoda and Miss Akemi Mitsuhashi, of NTEC, for their help in arranging data.

References

[1] I.Namiki, H.Hamada, R.Nakatsu: "Japanese Sentence Input Using Speech Recognition," Proceedings of ICTP pp.304-308 (Oct. 1983).

[2] The IBM group: "Experiments with the Tangora 20,000 word speech recognizer," Proceedings of ICASSP 87, pp.701-704 (Apr. 1987).

[3] J.Baker: "DRAGONDICTATE-30K: Natural Language Speech Recognition with 30,000 Words," Proceedings of Eurospeech 89, Vol.2,pp.161-163 (Sep. 1989).

[4] N.Sugamura, T.Tsuboi, R.Nakatsu: "Japanese Text Input System Based On Continuous Speech Recognition," Proceedings of ICASSP 86, pp.1125-1128 (Apr. 1986).

[5] A.Tomihisa, N.Sugamura, R.Nakatsu; "Japanese Continuous Speech Recognition Based on CV Syllable Spotting," Proceedings of Seventh FASE Symposium, pp.561-568 (Aug. 1988).

[6] R.Oka, "Continuous Words Recognition by Use of Continuous Dynamic Programming for Pattern Matching," Trans. of the Committee on Speech Research, S78-20, pp.145-152 (June 1978) (in Japanese).

[7] N.Sugamura: "Continuous Speech Recognition Using Large Vocabulary Word Spotting and CV Syllable Spotting," Proceedings of ICASSP 90, pp.121-124 (Apr. 1990).

[8] T.Tsuboi, A.Tomihisa, N.Sugamura: "Japanese Linguistic Processing for Continuous Speech Recognition," Proceedings of ICASSP 87, pp.805-808 (Apr. 1987).

[9] T.Tsuboi, N.Sugamura: "A Speech-to-Text Transcription System Using Continuous Speech Recognition," Proceedings of AVIOS 90 (Sep. 1990).