TEXT INPUT USING SPEAKER-ADAPTIVE CONNECTED SYLLABLE RECOGNITION

Yoichi Takebayashi, Hiroyuki Tsuboi, Shouichi Hirai, Hiroshi Matsura and Tsuneo Nitta

TOSHIBA Corporation, 1 Komukai-Toshiba-cho, Saiwai-ku, Kawasaki 210, JAPAN

This paper describes a speech recognition system for large vocabulary text input. The system recognizes connected Japanese syllables by both continuous pattern matching and speaker-adaptation based on the Multiple Similarity(MS) method. The recognition algorithm consists of syllable boundary detection, vowel and consonant recognition and lexical verification. The reference pattern vectors adapt to each speaker by K-L expansion through covariance matrix modification. Recognition experiments on a 17,877 word Japanese vocabulary showed 92.6% accuracy for 10 male 4,400 phrase utterances.

INTRODUCTION

While many speech recognition systems have been developed in the last decade, few word recognition systems have been accepted for text input application owing to poor accuracy and limited vocabulary. DP matching is a prevailing technique for word pattern matching, but it's not practical enough except for speaker-dependent small-vocabulary word recognition. The Multiple Similarity(MS) word pattern matching method is extremely powerful, but limited to a speaker-independent small vocabulary[1]. Likewise, the multi-template method is not applicable to a large vocabulary. Several word recognition systems based on probabilitistic model have been developed[2], but they require a lot of computation for large vocabulary recognition. On the other hand, the phoneme or syllable based recognition methods, the syntactic methods, are absolutely required for both continuous speech recognition and practical large vocabulary recognition[3]. However, the accuracy of the phonological units has been insufficient due to no effectual training algorithms. While rule-based speech recognition method is being studied to achieve full use of speech knowledge[4], automatic learning is still a open problem in AI research.

In this paper, an approach to achieving a large vocabulary word recognition system is first described. Then the proposed system is presented concerning acoustical and phonetic and lexical representations, continuous pattern matching and speaker adaptation. Finally experimental results are shown.

APPROACH

In order to attain a practical large vocabulary word recognition system or voice-activated word processor, we focus on the following points as design concepts:

a) High recognition accuracy
b) Strong and automatic speaker adaptation mechanism
c) Ease of utterance for novice users
d) Hardware realization and LSI implementation.

Taking these points into account, we have developed new connected syllable recognition and speaker adaptation methods[5]. The recognition system--consisting of syllable segmentation, vowel recognition and consonant recognition--employs MS calculation and acoustic labeling on a time-continuous frame by frame basis. We introduce a promising MS based approach because of the reliability and accuracy of the MS method in speaker-independent word recognizers[1] and character readers[6]. Conventional pattern matching methods, like DP matching, are so sensitive to pattern variation that they cannot be applied to syllable recognition. Rule-based phoneme recognition systems are being developed to utilize speech-specific knowledge. However, the learning mechanism (automatic knowledge acquisition) is still poor to date. Hence the pattern recognition oriented approach is much more promising than the rule-based one for implementing the speaker adaptation mechanism. The continuous MS matching is suitable for hardware realization. Also the vowel and consonant pattern vectors are reasonably represented by considering their inherent properties. In addition, our MS based adaptation method has a huge capacity to represent the phoneme pattern variability in detail--a large degree of freedom, therefore it is robust and reliable in regard to pattern variation and distortion. While the connected syllable approach is restrictive, the continuous MS matching and adaptation methods are applicable to further continuous speech recognition research. The recognition system demands user's cooperation, that is, clear utterance for connected syllable recognition. Our main purpose for developing this system is that novice users can input a lot of data more comfortably and efficiently by using this recognizer than keyboard.

Figure 1 shows a newly developed recognition system.

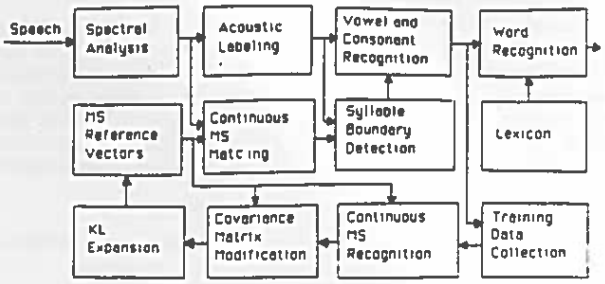

Fig.1 Blockdiagram of Recognition and Adaptation System

RECOGNITION AND ADAPTATION ALGORITHMS

Acoustic Representation

Input signal is converted into a 12-bit digital signal at 12kHz-sampling frequency. Spectral analysis is done by 3 sets of 4-pole digital band-pass filters. These filter outputs are squared and smoothed over 16ms. frames, and converted into logarithmic ones and then sampled at every 8ms. The overall energy is simultaneously obtained every 8ms. The 16-channel filter and 8-channel filter outputs are fed into vowel and consonant MS calculation, respectively. The 4- channel filter outputs are used for acoustic labeling.

The MS method utilizes the structure of pattern variation on pattern space for each category. Therefore parametric analysis, like LPC, is not used. Instead, non-parametric filter bank analysis is used. Modeling in pattern space based on the MS method is more reasonable and effective than that in speech signal for speech recognition.

Continuous Multiple Similarity Method

The MS method has been theoretically derived and experimentally proved to be powerful and effective by several optical character readers [5] and speaker-independent word recognizers[1]. The telephone speech recognizer accomplished a high performance in spite of significant pattern distortion. However, it cannot directly apply to phoneme or syllable recognition, as phoneme or syllable patterns have much less information than word utterance patterns. In order to obtain accuracy, real-time processing and adaptation mechanism, we propose the continuous MS pattern matching method. This method, based on the time-continuous MS calculation every 8 ms., is suitable for hardware realization. The problem in applying the MS method to connected syllable is how to represent the vowel and consonant feature vectors as N-dimensional feature vectors.

Vowel and Consonant Pattern Vector Representation

Each Japanese syllable has either one of five vowels or a syllabic nasal. The vowel is more durable and stable than the consonant. Therefore vowel recognition is a crucial component of all the recognition system. Considering these points, we represent the vowel pattern as a 16-dimensional vector(one frame 16 channel frequency spectrum) for the continuous vowel MS calculation.

As contrast with the vowel, the consonant part is not stable and inherently characterized by time-variant spectral patterns. Therefore we represent the consonant pattern vector as a multiple-frame time-frequency spectrum, not as a

one-frame spectrum. The consonant 64-dimensional vectors, generated by 8-channel frequency spectra over 8 frames, have 128ms. duration and are continuously matched by consonant reference vectors every 8ms.

Acoustic Labeling

Although the continuous MS pattern matching might work considerably well for both vowel and consonant recognition, we also introduce the acoustic labels in order to complement the MS values. A similar acoustic labeling was effectively employed in the telephone speech recognition system[1]. The 4-channel spectrum and overall energy are fed into the labeling processing.

Syllable Boundary Detection

Loose syllable boundaries(start and endpoints) are needed as clues for vowel and consonant recognition, as the highly efficient and stable continuous matching is employed. These points are determined by not only a time series of overall energy, 4-channel spectrum and acoustic label but also syllable duration constraints. Syllable recognition accuracy depends significantly upon the syllable detection performance.

Syllable(Vowel and Consonant) Recognition

Vowel region is estimated by both the loose syllable boundary information and acoustic label sequences. Then vowel recognition is carried out by using a time series of vowel similarities and acoustic labels in the estimated vowel region. A segmented input syllable is classified to one of 6 vowels(/a/,/i/,/u/,/e/,/o/,/N/). The 15 entries of MS reference vectors are prepared for the accurate vowel recognition.

The vowel recognition result focuses the syllable candidates on ones that include the recognized vowel. It can significantly lighten the computation load and also consonant pattern variation based on co-articulation effect. The consonant region is determined by the syllable boundary, vowel recognition result and acoustic labels. Then consonant recognition is realized by using a time series of consonant MS values. The simplest recognition way is where the consonant category with the maximum MS value within the region is regarded to be a recognized consonant(syllable) as a result. The second and the third rank candidates with likelihood are also obtained by using their MS values for lexical verification at next stage.

Speaker Adaptation

While the proposed speaker adaptive recognition system works without training, recognition accuracy can dramatically increase after the sophisticated adaptation[5]. Most traditional adaptation methods, based on the multi-template technique or some statistical learning method like perceptron or linear discriminants, are not clear and not structural from lack of speech knowledge utilization. In contrast, the MS based adaptation and recognition methods positively utilize the structure of pattern variability. Namely, the speaker-adapted reference vectors of each category represent a specific speaker's essential pattern distribution, to accomplish robustness and reliability. An important problem is how to extract the training pattern vectors from the whole speech pattern. We consistently introduce the continuous MS matching not only for recognition but also adaptation. Training patterns including a vowel or consonant part are approximately extracted in terms of the acoustic labels and syllable boundaries. Then the continuous MS calculation is done on these patterns. Subsequently, the fixed training pattern vectors are extracted from the frames with the greatest MS values. Next, the covariance matrices are modified by these vectors. Finally the K-L expansion of the covariance matrices generates the reference pattern vectors. As the learning progresses, the extracting position can change to successively precise positions. Thus stable and robust reference pattern vectors can be obtained at the adaptation stage as the user utilize the recognizer more and more. Both enormous capacity and knowledge acquisition mechanism are remarkable advantages of the proposed method.

Word/Phrase Recognition

Dealing with only clearly spoken connected syllables, the system ignores possibility of the syllable insertion and deletion. Thus the lexicon for phrase(word) recognition is simply represented using syllables. Word or phrase recognition is carried out by lexical verification between the syllable candidates with their likelihood and the lexicon. For real-time processing, lexical search space reduction is made by using three preceding syllable candidates. As the lexical processing is quite simple, further research is necessary to improve the word recognition performance.

EXPERIMENTAL RESULTS

A 10 male training data set(50 samples per consonant, 100 samples per vowel) was collected for 101 Japanese syllables for adaptation of each speaker. Another test data set including 4,400 phrases(9,230 syllables) was collected for evaluation of large-vocabulary recognition at the speed from 3 to 4 syllables per second. Table 1 shows the accumulated syllable recognition scores for both data sets. The accumulated scores, 97.8% and 100% suggest the stability and robustness due to a large capacity of the MS based method. More than 99.0% vowel recognition accuracies were obtained for the same data. Table 2 gives the phrase recognition score for a 17,877 word vocabulary. While simple lexical matching is used, the phrase(word) recognition score is considerably high because of the high syllable recognition accuracy. The results also demonstrate the reliability of the MS based continuous matching and adaptation.

	Training Data Set (50,500 syllables)	Test Data Set (9,230 syllables)
Best Candidate	98.9%	91.4%
3 Best Candidate	100.0%	97.7%

Table 1 Syllable Recognition Score

(after lexical verification) using 3-best-candidate	Test Data Set (9,230 syllables) (4,400 phrases)
Phrase Recognition Rate	92.6%

Table 2 Phrase Recognition Score

CONCLUSION

A text input recognition system using connected syllable recognition has been developed for novice keyboard users. The system employs both continuous matching and speaker-adaptation based on the MS method. The experimental results have shown that the proposed system is accurate enough to act as a practical voice-activated word processor or large vocabulary data entry system. Since the dominant computation of the MS recognition and learning methods is multiplication-accumulation, a real-time machine can be easily realized by LSI implementation.

REFERENCES

[1] Y.Takebayashi,et all.,"Telephone speech recognition using a hybrid method" Proc.fifth ICPR,pp.1232-1235,1984

[2] F.Jelinek,"The development of an experimental discrete dictation recognizer",Proc.IEEE,vol.73,no.11,pp.1616-1624,1985

[3] M.J.Hunt, et al.,"Experiments in syllable-based recognition of continuous speech",Proc.ICASSP'80,pp.880-883,1980

[4] V.W.Zue,"The use of speech knowledge in automatic speech recognition", Proc.IEEE,vol.73,no.11,pp.1602-1615,1985

[5] H.Tsuboi,et al.,"The connected syllable recognition based on Multiple Similarity method",Proc.ICASSP'86,1986

[6] K.Sakai,et al.,"An optical Chinese character reader",Proc. Third IJCPR,pp.122-126,1976