IMPROVING THE SPEAKER-DEPENDENCY OF SUBWORD-UNIT-BASED ISOLATED WORD RECOGNITION

Takuya Koizumi, Shuji Taniguchi, and Kazuhiro Kohtoh

Dept. of Information Science, Fukui University
3-9-1 Bunkyo, Fukui-shi, 910-8507 Japan

ABSTRACT

This paper deals with a subword-unit-based isolated word recognition system with enhanced speaker-independency. The subword is defined as a part of word whose central portion has rather stationary or time-invariant short-time spectra with its portions near its ends having rapidly varying short-time spectra. In this system each isolated word is decomposed into a sequence of subwords, each of which is identified by means of a particular semi-continuous hidden Markov model that is named a subword HMM. Each isolated word is recognized by a particular set of concatenated subword HMMs that is designated as a word HMM. Subword boundaries within a word are detected by finding peaks of the magnitude of delta cepstral vectors obtained from the word. The system attains average word recognition rates over 87 % for a number of Japanese words uttered by ten native male speakers.

1. INTRODUCTION

In speech recognition, commonly used units of recognition are phoneme, word etc.. As is well known, however, these are not ideal units of recognition, for one finds inherent difficulty in using them.

Another choice of unit is an acoustic unit called “subword” which is defined as a part of word whose central portion has rather stationary or time-invariant short-time spectra with its portions near its ends having rapidly varying short-time spectra. By using this subword as a unit of recognition it is possible to resolve the problem of coarticulation that might otherwise deteriorate the performance of any phoneme-based speech recognizer. Decomposing a word into a sequence of subwords will be an easy task, if one uses the delta cepstrum which is capable of detecting spectral change effectively. Subword boundaries within a word are detected simply by finding peaks of the magnitude of delta cepstral vectors obtained from the word.

Large vocabulary isolated word recognition requires a large amount of training data proportional to the vocabulary size to characterize each individual word model. A subword-unit-based approach [1]~[5] is a more viable alternative than the word-based approach to overcome the problem of the training data size, since different words can share common segments in their representation in the former approach.

It has been previously found that a subword-unit-based isolated word recognizer, in which subword units are identified by means of discrete hidden Markov models (DHMMs), is capable of attaining higher recognition accuracies than conventional word-based isolated word recognizers in the absence of background noise. In fact, recognition accuracies achieved with this subword-unit-based system are found to be over 98 % in speaker-dependent applications where utterances of ten speakers are used for recognition as well as training. However, the performance of this system is speaker-dependent and it more or less deteriorates in speaker-independent applications where input words to be recognized are uttered by a speaker whose utterances have not been used to train the DHMMs within the system, since the DHMMs trained using isolated words uttered by some speakers tend to incorrectly identify those uttered by some other speakers. An effective way to cope with this speaker-dependency is to replace the DHMMs with semi-continuous HMMs (SCHMMs). The SCHMMs will show less speaker-dependencies in their ability of identifying subwords than the DHMMs do, unless the spectral change of each subword due to the change of speaker is beyond certain limits.

The ability of the subword-unit-based isolated word recognizer employing the SCHMMs in speaker-independent applications has been investigated by an isolated word recognition experiment using a number of Japanese words uttered by ten native male speakers.

In what follows, the isolated word recognition system will be described in detail, then the performance of this system will be compared with that of a similar subword-unit-based system employing the DHMMs.

2. DESCRIPTION OF THE SYSTEM

2.1. Word Data

Ten native male speakers uttered 75 different Japanese city names six times in a quiet environment and the resulting word data containing 4,500 isolated words were equally divided into two parts in either of the following two methods, A and B, to use one part for training the system and the other part for testing the system.

Method A: 2,250 words uttered three times by ten speakers are used for training the system, and 2,250 words uttered by the same ten speakers for testing.

Method B: 2,250 words uttered by five speakers are used for training, and 2,250 words uttered by five other speakers are used for testing.

2.2. The Decomposition of Words into Subword Units

The decomposition (segmentation) of input words into subword units is accomplished through the use of delta cepstral analysis. The delta cepstrum shows a sharp peak at a point where the short-time spectrum of speech makes an abrupt change. Suppose that the delta cepstral analysis is applied to an input word, producing several sharp peaks that are above an appropriate threshold prescribed beforehand. A subword unit is defined as follows: A part of the word between two consecutive delta cepstral peaks including the falling edge of the first peak and the rising edge of the second peak will be characterized by its peculiar short-time spectra such that its central portion has rather time-invariant short-time spectra and its portions near its ends have rapidly time-varying short-time spectra. This part of the word call we a "subword." The delta cepstral analysis is applied to each frame of a word and the sum of the first through 6th coefficients squared of the delta cepstrum, which some authors called "dynamic measure"[6], is calculated for each frame. The value of the sum shows a peak around a boundary between two successive subwords, therefore, if a peak of those sums which is above an appropriate threshold is detected, the position of the peak may be considered as a boundary between two consecutive subwords.

2.3. Isolated Word Recognition

A block diagram of the subword-unit-based isolated word recognition system is depicted in Figure 1. Each input word undergoes spectral and delta cepstral analyses and segmentation based on the delta cepstral analysis, and is transformed into a sequence of preliminary subwords.

In training phase, a secondary spectral analysis is applied to each preliminary subword of the sequence to generate a sequence of 48-dimensional power spectral vectors. Each of 48-dimensional power spectral vectors belonging to the resulting sequence is classified into 50 groups by a process called "segment quantization", which is a vector quantization with 50 quantizing levels, using K-means clustering algorithm. In this segment quantization a mean of vectors belonging to one group or cluster is chosen as a representative vector of that group. As a result of this quantization each sequence of preliminary subwords corresponding to a word is transformed into a new sequence which will be called a "subword sequence." This process produces a subword dictionary which is a collection of subword sequences representing words. Each subword sequence undergoes the second vector quantization with 200 quantizing levels, producing a subword code sequence. A set of subword code sequences representing a particular subword is used to

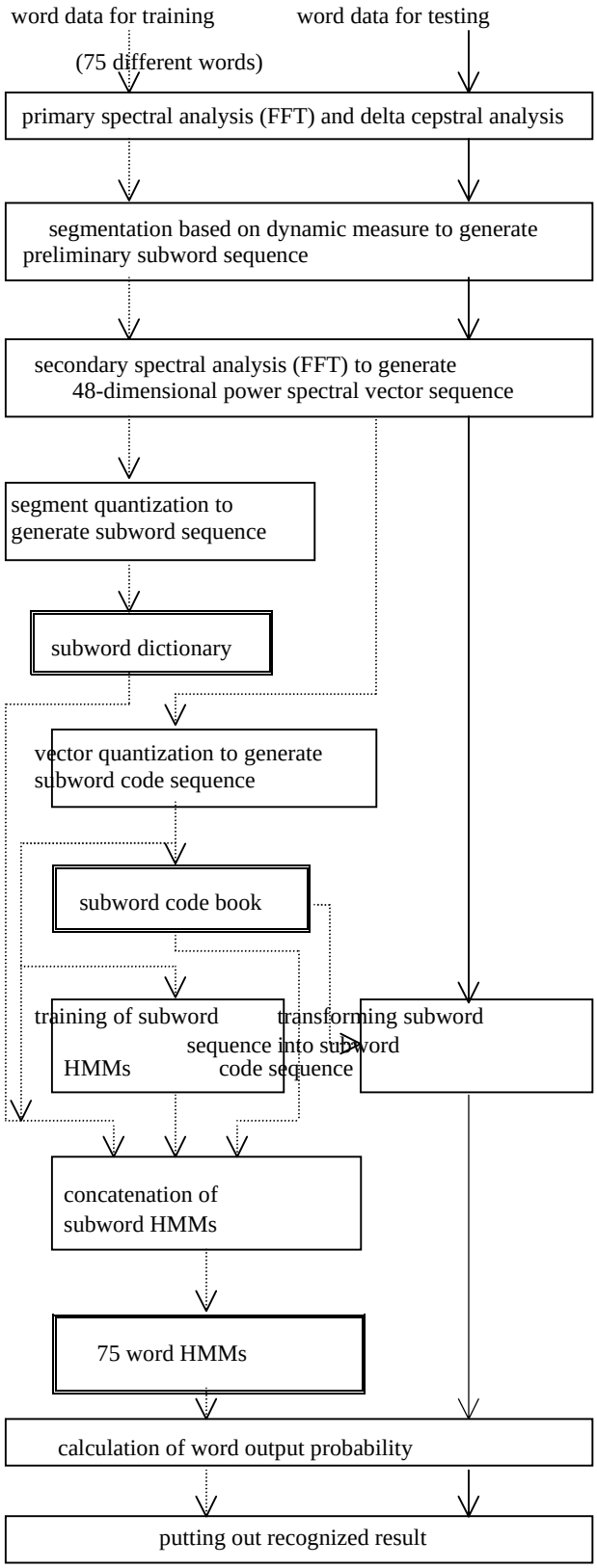

Figure 1: The block diagram of the subword-unit-based isolated word recognition system.

train a three-state semi-continuous HMM, which will be called a subword HMM. The final process of the training phase is to form a chain of subword HMMs (concatenated subword HMMs[7]) for each of the 75 different words, which will be called a word HMM.

In testing or recognition phase, an unknown word, after the primary spectral and delta cepstral analyses, the segmentation, and the secondary power spectral analysis, is transformed into a sequence of codes referring to the subword code book generated in the training phase, which is then put into each of the 75 word HMMs. Word output probability is calculated for each word HMM and a word represented by a word HMM which has yielded the largest word output probability is put out as a recognized result.

By the way, the training of a word HMM requires sequences of codes with an appropriate length determined by the number of states of the HMM. It is not desirable that code sequences used in the recognition phase are longer than those used in the training phase. To avoid such an undesirable situation it does not suffice to provide a single word HMM for each word, but it is necessary to provide six word HMMs (concatenated SCHMMs) with 8, 10, 12, 14, 16, and 18 states, which are appropriate for the recognition of a word made up of 17, 21, 25, 29, 33, and 37 subwords, respectively, for each of the 75 different words. One of the six word HMMs provided for each word will be chosen by the number of subwords included in an incoming word to be recognized. This method of coping with the variability of the number of subwords has turned out to be an effective means of improving the recognition accuracy of the system, and will be called a multiple HMM scheme. An example is shown in Figure 2.

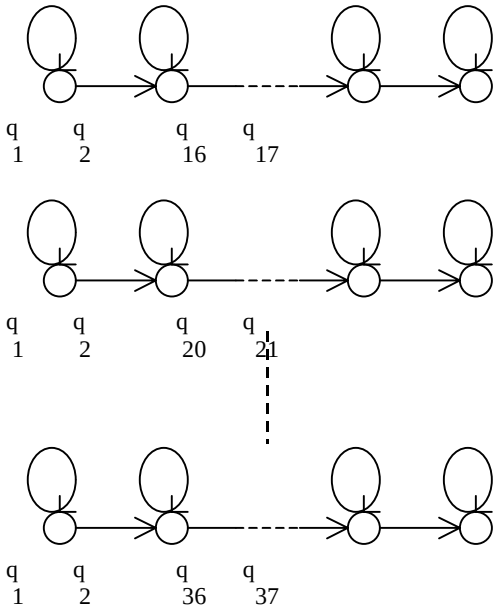

Figure 2: An example of the multiple HMM scheme.

3. WORD RECOGNITION EXPERIMENT

A word recognition experiment has been performed using the word data described earlier. Two methods of dividing the word data into two parts were employed: Method A where 2,250 words uttered three times by ten speakers are used for training the system, and 2,250 words uttered by the same ten speakers for testing and Method B where 2,250 words uttered by five speakers are used for training and 2,250 words uttered by five other speakers for testing. The former is intended for investigating the ability of the system in speaker-dependent applications and the latter in speaker-independent applications. A result of the experiment is shown in Table 1, where average recognition accuracies of a similar system employing the DHMMs as the word HMMs are shown for comparison. Table 2 shows word by word recognition rates of the system employing the SCHMMs for the 75 Japanese words uttered by the ten speakers.

Word HMMs	Method A	Method B
DHMMs	96.5 %	83.2 %
SCHMMs	98.5 %	87.8 %

Table 1: Average word recognition rates for the different types of word HMMs and word data.

Words	Method A	Method B
Fukui	100.0	76.7
Katsuyama	100.0	83.3
Ohno	100.0	90.0
Sabae	100.0	90.0
Takehu	100.0	93.3
Tsuruga	96.7	83.3
Obama	100.0	100.0
Kanazawa	100.0	73.3
Komatsu	100.0	100.0
Hakui	100.0	90.0
Suzu	100.0	80.0
Wajima	96.7	100.0
Mattoh	96.7	93.3
Nanao	96.7	86.7
Kaga	90.0	80.0
Toyama	100.0	83.3
Takaoka	100.0	96.7
Uozu	90.0	70.0
Namerikawa	100.0	100.0
Himi	93.3	93.3
Tonami	96.7	86.7
Oyabe	100.0	90.0
Kurobe	100.0	100.0
ShiNminato	100.0	96.7
Gihu	100.0	93.3
Ohgaki	96.7	86.7
Kagamigahara	96.7	96.7
Tajimi	100.0	93.3
Kani	100.0	86.7
Seki	100.0	83.3

Mino	100.0	86.7
Mizunami	100.0	93.3
Toki	96.7	83.3
Nakatsugawa	96.7	90.0
Hashima	100.0	86.7
Minokamo	100.0	100.0
Takayama	93.3	76.7
Ena	93.3	76.7
Itoigawa	100.0	96.7
Ojiya	100.0	100.0
Kashiwazaki	100.0	100.0
GoseN	100.0	93.3
Sanjoh	100.0	83.3
Shibata	100.0	93.3
Joh-etsu	100.0	93.3
Shirane	100.0	90.0
Tsubame	100.0	100.0
Tohkamachi	100.0	100.0
Tochio	100.0	100.0
Toyosaka	100.0	100.0
Nagaoka	96.7	96.7
Niigata	100.0	80.0
Niitsu	100.0	96.7
Mitsuke	100.0	96.7
Murakami	100.0	96.7
Ryohtsu	96.7	63.3
Iida	93.3	80.0
Iiyama	96.7	66.7
Ina	90.0	43.3
Ueda	100.0	83.3
Okaya	96.7	86.7
Komoro	100.0	100.0
Saku	96.7	66.7
Shiojiri	100.0	96.7
Suzaka	100.0	70.0
Suwa	100.0	83.3
Chino	96.7	76.7
Nagano	96.7	63.3
Matsumoto	100.0	100.0
Beppu	100.0	90.0
Sapporo	100.0	100.0
Kamo	100.0	73.3
Ohmachi	100.0	90.0
Nakano	100.0	70.0
Komagane	100.0	96.7

Table 2: Word by word recognition rates (%) of the system employing the SCHMMs for the 75 Japanese words uttered by the ten speakers.

The systems utilize no linguistic knowledge about input words, so these recognition rates have been obtained solely by the so-called bottom-up processing of the input words.

4. CONCLUSIONS

Findings from the result of the experiment can be summarized as follows:

1. The subword-unit-based isolated word recognition system employing either the DHMMs or the SCHMMs attains high average recognition accuracies in speaker-dependent applications.
2. The performance of the subword-unit-based isolated word recognition system employing the DHMMs deteriorates considerably in speaker-independent applications.
3. The subword-unit-based isolated word recognition system employing the SCHMMs attains over 4 % higher recognition accuracies in speaker-independent applications than the similar system employing the DHMMs.

5. REFERENCES

1. Svendson, T., Paliwal, K.K., Harborg, E., and Husøy, O. “An Improved Subword Based Speech Recognizer,” *Proc. ICASSP*: 108–111, 1989.
2. Svendson, T., and Soong, F.K. “On the Automatic Segmentation of Speech Signals,” *Proc. ICASSP*: 77–80, 1987.
3. Sugamura, N., and Furui, S. “Large Vocabulary Word Recognition Using Pseudo-Phoneme Templates,” *Trans. IEICE Japan, J65–D*: 1041–1048, 1982.
4. Koizumi, T., Fukuyama, A., Mori, M., and Taniguchi, S. “Speech Recognition Based on Subword Units,” *Proc. the 3rd Joint Meeting of ASA and ASJ, Hawaii*: 1129–1134, 1996.
5. Mori, M., Koizumi, T., Fukuyama, A., and Taniguchi, S. “Speech Recognition Based on Subword Units,” *Trans. IEE Japan, 118–C*: 520–527, 1998.
6. Sagayama, S., and Itakura, F. “On Individuality in a Dynamic Measure of Speech,” *Proc. Spring Conf. Acoust. Soc. Japan*: 3-2-7, 589-590, 1979.
7. Maruyama, K., Hanazawa, T., et al. “English Word Recognition Using HMM Phone Concatenated Training,” *Tech. Rep. IEICE Japan, SP88–119*: 1989.