NEW WORD SPOTTING ALGORITHM BASED ON FORWARD DECODING

Jin'ichi Murakami

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
murakami@itl.atr.co.jp

1. INTRODUCTION

In this paper, we describe a new word spotting algorithm based on forward decoding. In past studies, continuous dynamic programming[1] had been popular for word spotting. This algorithm which uses dynamic programming is considered to have a free start point and end point, and is controlled by some criterion; determining this criterion is very difficult.

The proposed decoding algorithm basically performs comparison for the score of the last state of any word HMM. Therefore, it is expected to obtain a higher performance than continuous dynamic programming. Also we describe the problem of Viterbi decoding and explain that this algorithm may resolve the above problem.

2. ERGODIC HMM CONNECTING WORD HMMS WITH WORD BIGRAM PROBABILITIES

Connecting word HMM models to each other with word bigram probabilities, results in the network becoming a large single Ergodic HMM consisting of multiple states and transition probabilities. This Ergodic HMM is shown in Fig. 1.

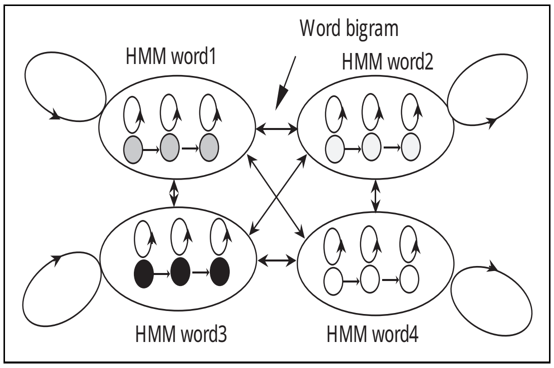

Figure 1: Ergodic HMM connecting Word HMMs with Word bigrams

The transition probabilities from one word HMM to another word HMM correspond to word bigram probabilities and are estimated from the existing text database. The word HMMs are estimated for a word utterance speech database using the Baum-Welch algorithm.

3. DECODING FOR OPTIMAL STATE SEQUENCE

Given the Ergodic HMM parameters and speech data, we consider the problem of estimating the optimal state sequence. To resolve this problem, the Viterbi decoding is a formal technique for finding the single best state sequence. Word sequences are directly estimated to obtain this optimal state sequence.

Now, we propose the new decoding algorithm. We call this algorithm the forward decoding. It is described in the following.

3.1. Forward decoding (New spotting algorithm)

In this algorithm, first, a score is calculated for summing each score of a state. Next, the optimal state sequence is assigned for the best score of each frame. This decoding algorithm is given as follows.

s_t : optimal state sequence at time t;
$\delta_t(j)$: score at time t and state j ;
N : number of states for Ergodic HMM ;
a_{ij} ; transition probability at state i to j;
$b_j(o_t)$; output probability at state j
when the observation vector is (o_t) ;

1. For all states $i \in \{1, ..., N\}$
   $\delta_1(i) = \pi_i \times b_i(o_1)$
   $s_1^* = arg \max_i \delta_1(i)$
2. Calculate $j \in \{1, ..., N\}$ along the time axis for $t = 2, ..., T$ and for all states j, compute:

   $\delta_t(j) = \sum_i [\delta_{t-1}(i) \times a_{ij} \times b_j(o_t)]$
   $s_t^* = arg \max_j \delta_t(j)$
3. The final score is given by :
   $\sum_S P(O, S \mid M) = \sum_j \delta_T(j)$
   $s_T^* = arg \max_j \delta_T(j)$

Finally, if the optimal state sequence is the last state of any word, it is assumed that this word is recognized.

This decoding algorithm is similar to the forward-backward algorithm[5]. However, it can recognize the word for each frame, and therefore, can be regarded as a kind of word spotting algorithm.

4. CONTINUOUS SPEECH EXPERIMENTS

To show the effectiveness of this algorithm, we carried out continuous speech recognition experiments.

4.1. Experimental conditions

We developed a word HMM to connect phone HMMs. To train continuous mixture HMMs, 2,635 word utterances were used. For the training of bigram probabilities, 8,475 sentences of the ATR Dialog Database were used in text-open experiments and smoothed by the deleted-interpolation algorithm [4]. Then the same 8,475 sentences plus 38 test sentences were used in text-closed experiments. In this way, it was possible to perform both the text-open and text-closed experiments using the same test data. To reduce the memory requirements, we used beam pruning. The results of these experiments were evaluated in terms of the word correct rate and word accuracy rate[3]. The experimental conditions are summarized in Table 1.

Table 1: Experimental conditions

phone model	3-state 4-loop continuous mixture HMM
mixture number	max 10 (valid for each syllable)
acoustic parameter	16th order LPC cepstrum + power +Δ power + 16th order Δ cepstrum
frame window	20 ms
frame period	5 ms
training voice	word speech (2,620 words)
# syllable categories	26 syllables
language model	word bigram
training data	8,475 sentences (57,354 words)
vocabulary	435
beam width	4,096
test sentence count	38 sentences

4.2. Results of Experiments

We show the results of our experiments in Table 2. The table shows that forward decoding is superior to Viterbi decoding for the word accuracy rate. However, it is inferior for the word correct rate. These results mean that forward decoding has fewer insertion errors than Viterbi decoding.

Table 2: Experimental results

Forward decoding	text-open	text-closed
word correct	34.0%	36.3%
word accuracy	25.9%	29.0%

Viterbi decoding	text-open	text-closed
word correct	54.1%	56.8%
word accuracy	20.1%	23.6%

5. DISCUSSION (COMBINING VITERBI DECODING AND FORWARD DECODING)

The Viterbi decoding algorithm is a formal technique for finding the single best state sequence. This algorithm compensates the global optimal state sequence for test data. However, beam pruning is normally used because full Viterbi decoding is impossible due to its high computational cost and memory requirements. This beam pruning is based on the schema that the local optimal state for each frame corresponds to the global optimal state sequence for the test data. However, this assumption is sometimes broken. Going back to forward decoding, it is based on the local optimal state for each frame. Therefore, combining viterbi decoding and forward decoding that is instead of beam pruning, may result in a better recognition algorithm.

6. CONCLUSION

In this paper, we describe a new word spotting algorithm that uses forward decoding. This algorithm is based on the Ergodic HMM. It basically performs comparison for the score of the last state of any word HMM. Therefore, it is expected to obtain a higher performance than continuous dynamic programming. Results of experiments showed a 36 % word accuracy rate which is superior to that by Viterbi decoding.

In the future, we plan to make a comparison between this new spotting algorithm and past continuous dynamic programming. In addition, we will try to make a new continuous speech recognition algorithm that combines Viterbi decoding and forward decoding.

REFERENCES

[1] J.R.Rohlicek et al., "Continuous Hidden Markov Modeling for Speaker Independent Word Spotting", ICASSP89, S12.8, pp.627-630 (1989).

[2] Y. Miyazawa, et al., "An All-Phoneme Ergodic HMM for Unsupervised Speaker Adaptation", ICASSP93, pp.2-574,2-577 (1993).

[3] Cambridge University Engineering Department Speech Group and Entopic Research Laboratories Inc.,
"HTK: Hidden Markov Model Toolkit V1.5"
(23 September 1993).

[4] F.Jelinek, "Self-Organized Language Modeling for Speech Recognition", Readings in Speech Recognition, Morgan Kaufmann Publishers, Inc. San Mateo, California, pp.450-506, (1990).

[5] X.D.Huang, et al., "Hidden Markov Models for Speech Recognition", Edinburgh (1990).