TR-I-0057

Word Spotting Method Based on HMM Phoneme Recognition

HMM音韻認識に基づくワードスポッティング

Takeshi KAWABATA, Toshiyuki HANAZAWA
and Kiyohiro SHIKANO

川端 豪　花沢 利行　鹿野 清宏

Nov. 21, 1988

ABSTRACT

A new technique for detecting and locating keywords in continuous speech using HMM(Hidden Markov Model) phoneme recognition is proposed. HMM word models are composed of HMM phone models trained on an isolated word database. Because the speaking rate between isolated words and continuous speech is different, phoneme spectra and durations change considerably. An HMM consists of several states and arcs. Each arc has output probabilities for each VQ code. In order to cope with the spectral changes, the output probabilities are smoothed with the probabilities of their spectral neighbor codes. In order to cope with the duration changes, HMM state duration parameters are shifted according to a 2nd-order duration calibration curve. The calibration curve is obtained from a speaking rate ratio of continuous speech to isolated words. The word detection rate for 8 keywords in 25 sentences uttered by one speaker was 98.4%. Accurate word spotting is accomplished using the HMM output probability smoothing technique and the state duration control mechanism taking the speaking rate into account.

ATR自動翻訳電話研究所

1. INTRODUCTION

This paper describes a new word spotting method based on HMM phoneme recognition.

For realizing an interpreting telephony system, large vocabulary and continuous speech recognition technologies are necessary.

In order to construct a large vocabulary system, we use an HMM phoneme recognition approach. Using this approach, any word models can be composed of only 30~ phone models.

The HMM phone models are trained on an isolated word database. Because the speaking rate between isolated words and continuous speech is different, phoneme spectra and durations change considerably. Therefore, this paper proposes two correction mechanisms for adopting the HMM phoneme recognition technique to continuous speech recognition. In order to cope with the spectral changes, the output probabilities are smoothed with the probabilities of their spectral neighbor codes. In order to cope with the duration changes, HMM state duration parameters are shifted according to a 2nd-order duration calibration curve obtained from a speaking rate ratio of continuous speech to isolated words.

The most accurate word spotting is accomplished using the HMM output probability smoothing technique and the HMM state duration control mechanism taking the speaking rate into account.

2. WORD SPOTTING METHOD BASED ON HMM PHONEME RECOGNITION

2.1 SCHEMATIC DIAGRAM

Figure 1 shows a schematic diagram of our word spotting system. HMM phone units are trained on an isolated word database using a forward-backward algorithm. HMM word units are made of HMM phone units according to their phonetic representation. Using these word models and the duration controlled Viterbi algorithm, the system spots keywords in continuous speech.

However, there is a problem. The training data is isolated, and speech for recognition is continuous. Because of the difference in speaking rate, phoneme spectra and durations vary considerably.

For this reason, we implemented two correction mechanisms in our system. One is a smoothing mechanism for HMM probabilities, and the other is a duration control mechanism taking the speaking rate into account.

2.2 TRAINING OF HMM PHONE UNITS

In this system, Japanese phonemes are divided into two classes, transient or stationary.

The model of a transient phoneme has 3 loops for representing its time structure. And the model of a stationary phoneme has only 1 loop. Table 1 and Fig.2 shows the phoneme classes and their HMM structures. Each loop is strongly related to an acoustical event. The duration of each loop is limited with minimum and maximum values.

All phone units are trained on an ATR isolated word database (5,240 words) using the forward-backward (Baum-Welch) algorithm. The speech is sampled at 12kHz and transformed to VQ code sequences using 12th order LPC analysis and a 256 point Hamming window shifted every 3ms. The codebook size is 256, and PWLR spectral distance measure is used.

2.3 SMOOTHING OF HMM OUTPUT PROBABILITIES

In order to cope with the spectral changes, output probability smoothing is carried out in our system.

Each loop of a Hidden Markov Model has an output probability table that includes probabilities for 256 VQ codes. The probability of each code is trained independently using the F-B algorithm. Even if two codes have similar spectra, their probabilities are sometimes very different. The situation is often observed especially when the number of training samples is small. The same problems occur for spectrum changes in continuous speech.

As shown in Fig.3, we smooth the output probabilities with their spectral neighbor codes in each table according to the equation (1), where 'd' means the PWLR spectral distance between code i and code j. A function 'f' depends on the spectral distance measure. In this case, we chose an exponential function.

Consequently, VQ codes that have similar spectra have similar probabilities. This operation increases the spectrum robustness of HMM phone models.

2.4 DURATION CONTROL METHOD TAKING THE SPEAKING RATE INTO ACCOUNT

In order to cope with the duration changes, we implement a duration control mechanism taking the speaking rate into account.

Figure 4 shows a scatter plot for the change of acoustical-event duration between fast and slow utterances. As the speaking rate increases, the duration-change distribution is shifted lower. This figure shows that the state duration of a long event changes more than that of a short event.

We approximate the center of the distribution using a 2nd order least square fit curve with a boundary condition as follows.

$g(0) = 0, \quad g'(0) = 1$

Because this 2nd order curve has only 1 free parameter, the curve is determined uniquely from the speaking rate ratio between training and test speech samples.

As previously mentioned, our Hidden Markov Models are duration controlled with minimum and maximum loop limits. The system detects the speaking rate of input speech and corrects these duration control parameters according to the 2nd order calibration curve calculated from the speaking rate.

3 WORD DETECTION EXPERIMENTS

The proposed word spotting system is tested using 25 sentences uttered by one male speaker with 2 utterance manners, a phrase-wise utterance and an utterance with no restrictions.

The system detects 8 Japanese keywords appearing a total of 64 times in the 25 sentences.

The word detection scores are shown in Table 2 and 3. For both utterance manners, word detection rates are 98.4%. There are 5 false alarms for the phrase wise utterance, and 15 false alarms for the utterance with no restrictions. The performance is high enough to make a continuous speech recognizer.

The 2nd table shows the effects of two correction mechanisms. Without any corrections, the detection rate is not particularly high. There are many false alarms. Using the probability smoothing, the detection rate is somewhat improved. Using the duration calibration technique taking the speaking rate into account, the detection rate is improved considerably. The detection rate is high and there are few false alarms.

4 CONCLUSION

The most accurate word spotting is accomplished using the HMM output probability smoothing technique and the state duration control mechanism taking the speaking rate into account. The recognition results show that the duration control is the most important technique for applying the HMM phoneme recognition to continuous speech recognition.

Using these techniques, we intend to construct a phone based continuous speech recognizer with a large vocabulary.

REFERENCES

[1] Hanazawa, T., Kawabata, T. and Shikano, K.: "Discrimination of Japanese Voiced Stops Using Hidden Markov Models", Proc. ASJ Fall Meeting, 1-5-10, pp.19-20 (Oct. 1987)

[2] Takeda, K., Sagisaka, Y. and Katagiri, S.: "Acoustic-Phonetic Labeling in a Japanese Speech Database", Proc. ASJ Spring Meeting, 2-5-10, pp.69-70 (Mar. 1987)

[3] Levinson, S.E., Rabiner, L.R., Sondhi, M.M.: "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition", BSTJ, Vol.62, No.4, pp.1035-1074 (April 1983)

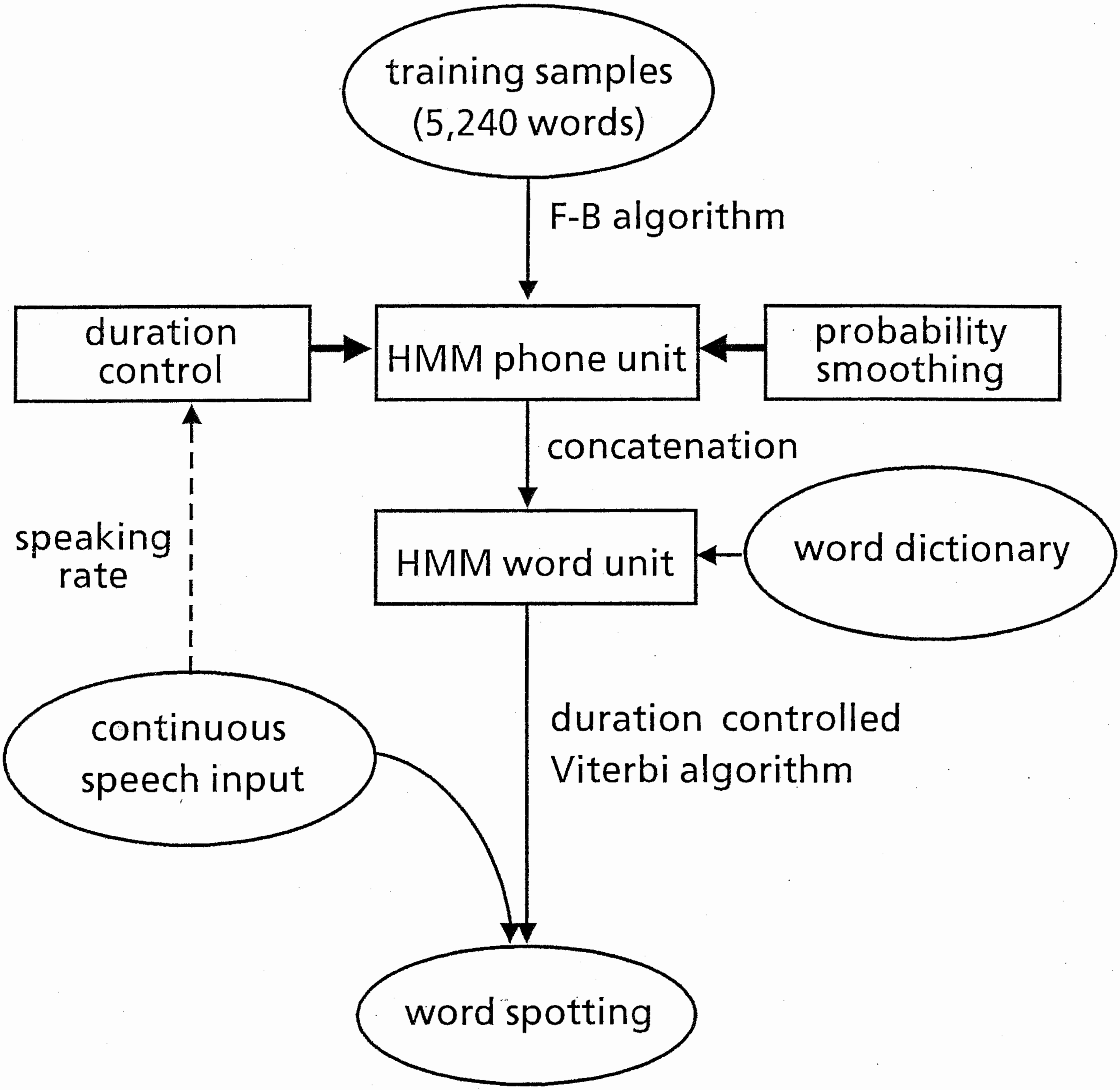
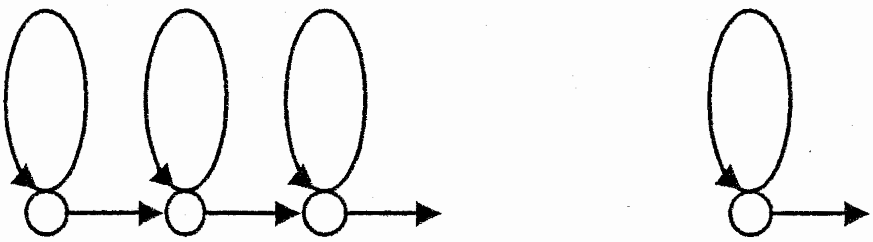


Fig.1 Schematic diagram of phoneme-based HMM word spotting

TABLE 1　PHONE UNITS

phone unit	loops
/b/,/d/,/g/,/p/,/t/,/k/,/m/, /n/,/ng/,/r/,/z/,/ch/,/ts/, /y/,/w/	3
/i/,/e/,/a/,/o/,/u/,/N/, /ii/,/ee/,/aa/,/oo/,/uu/,/ei/,/ou/, /s/,/sh/,/h/	1

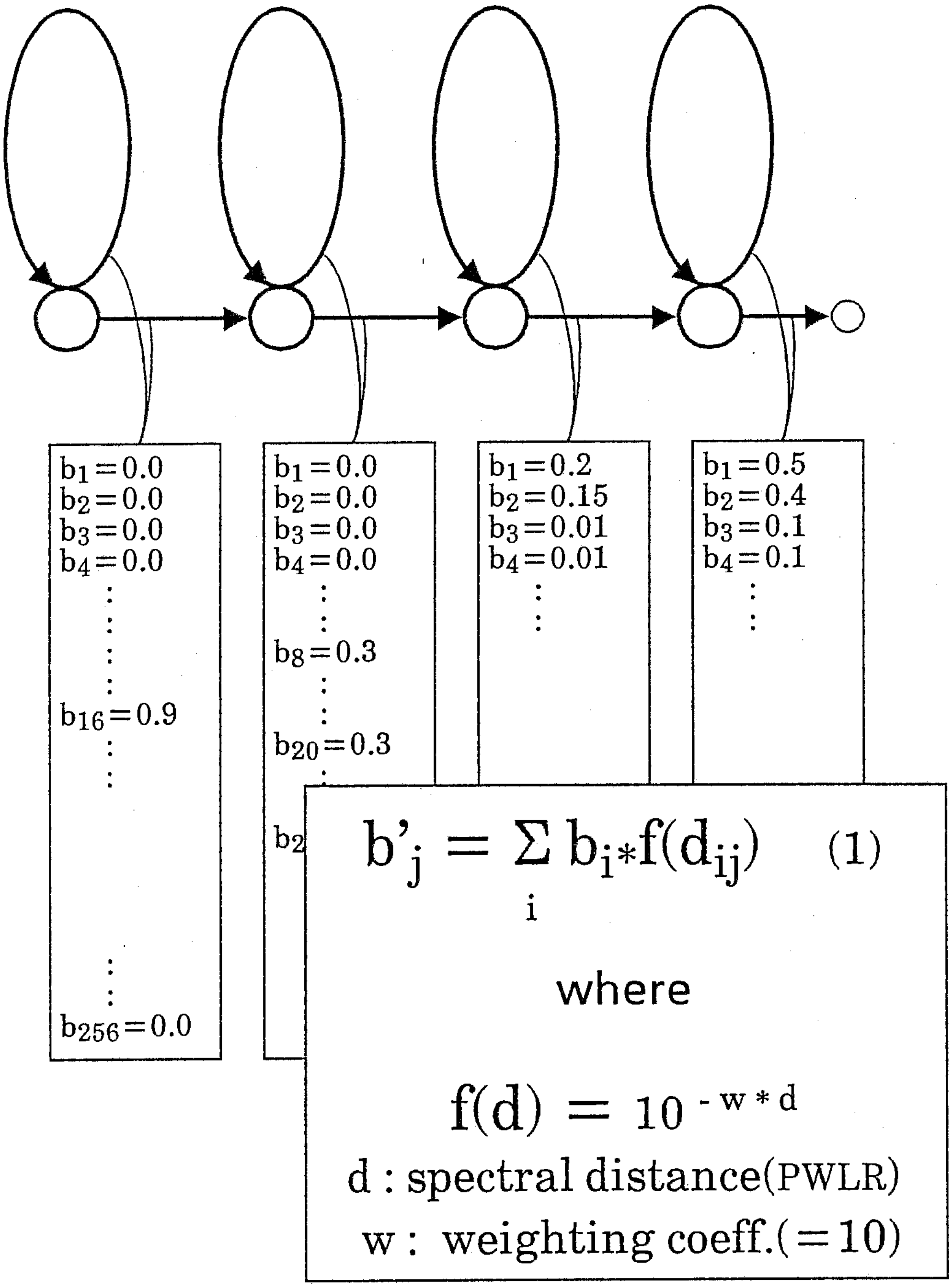


(a) 3-loop model　　(b) 1-loop model

Fig.2 Phone HMMs

◎ Each loop is strongly related to an accoustical event.

◎ The duration of each loop is simply controlled with two parameters. (loop_{min}, loop_{max})



$$b'_j = \sum_i b_i * f(d_{ij}) \qquad (1)$$

where

$$f(d) = 10^{-w*d}$$

d : spectral distance(PWLR)

w : weighting coeff.(=10)

Fig.3 Smoothing of HMM output probabilities ($b_{i\ (i=1..256)}$)

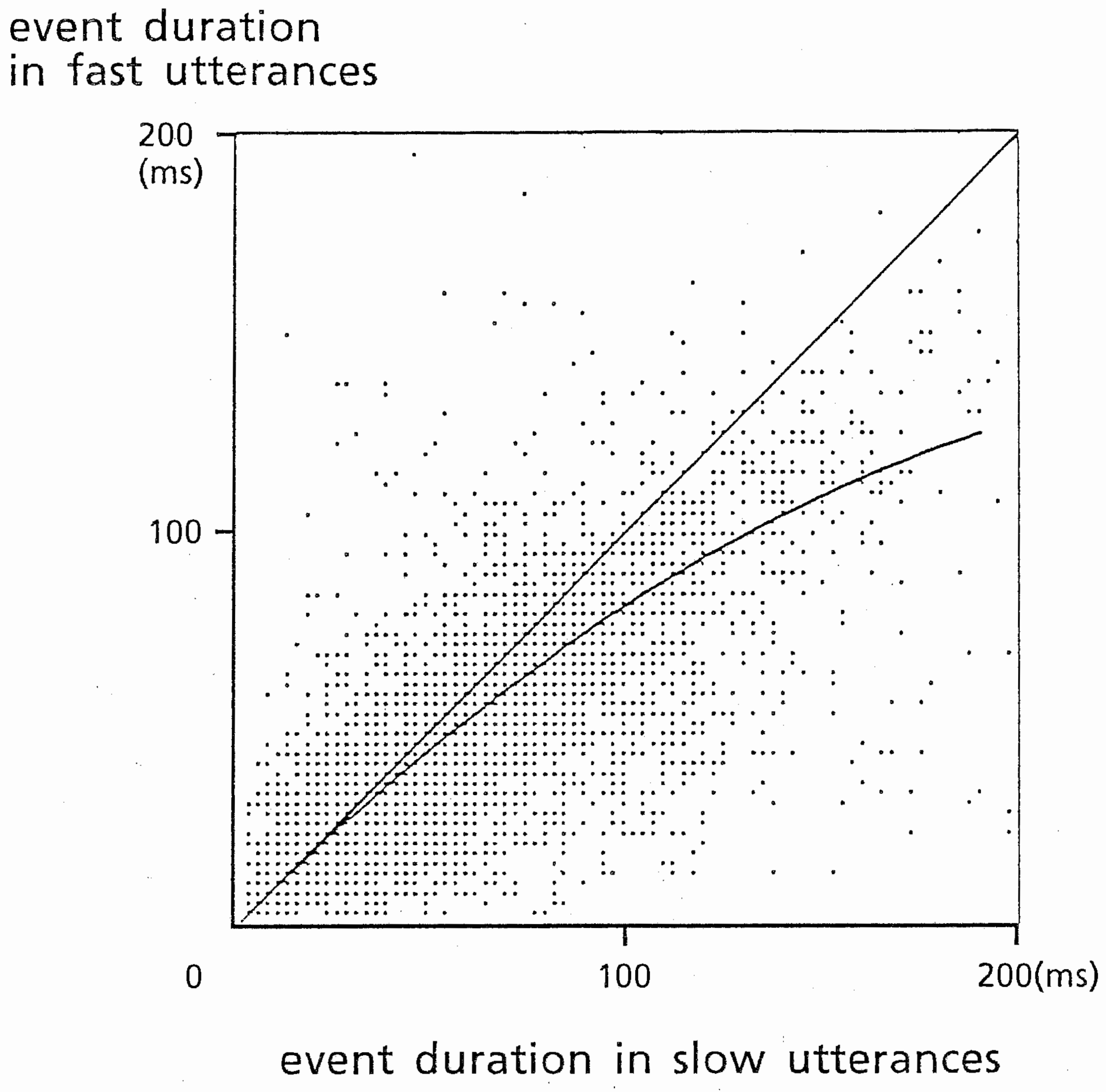


Fig.4 Scatter plot for the change of event duration between fast and slow utterances and its 2nd order least square fit curve

[Word detection experiment]

Speech materials are 25 sentences uttered by a male speaker with two utterance manners.

(A) Phrase-wise utterance

(B) Sentence (no restrictions)

Keywords to be detected are /kaigi/, /kokusai/, /deNwa/, /saNka/, /moushikomi/, /tsuuyaku/, /hoNyaku/ and /touroku/. They appear a total of 64 times in above sentences.

TABLE 2 WORD DETECTION SCORE 1

utterance	detection rate(%)	False alarms
(A) Phrase-wise	98.4	5
(B) Sentence	98.4	15

TABLE 3 WORD DETECTION SCORE 2

techinique		detection rate(%)	False alarms
probability smoothing	duration control		
-	-	87.5	54
○	-	89.1	48
-	○	98.4	8
○	○	98.4	5

(using phrase-wise utterances)