PREDICTION OF KEYWORD SPOTTING ACCURACY BASED ON SIMULATION

Yoichi Yamashita

Dep. of Computer Science, Ritsumeikan University
1-1-1, Noji-Higashi, Kusatsu-shi, Shiga, 525-8577 Japan
yama@cs.ritsumei.ac.jp

ABSTRACT

This paper proposes a method of predicting accuracy of keyword spotting in terms of FA count and spotting score of correct detections. A new measure F for predicting the FA count is calculated by simulation of the keyword spotting for phoneme sequences that phoneme-based language model generates. Another measure C for predicting the spotting score of correct detections is obtained from a product of correct recognition probabilities of phonemes. Both correlation coefficients and prediction errors are used to evaluate these measures in comparison with a simple measure of the keyword phoneme length, L. The prediction errors of FA count based on L was 7.71. The measure F reduced the prediction errors by 16%, and it had stronger correlation with the FA count. Furthermore a combined measure of F and L reduced the errors by 23%. On the other hand, L was more effective to predict the spotting score of correct detections than the measure C.

1. INTRODUCTION

Automatic classification of spoken documents is an important technique for efficient retrieval of multimedia data with audio channel. Topic identification (TID) of news speech, which is one of such tasks, selects a topic from given topic candidates based on keywords extracted from input speech[1]-[5]. The keywords must have strong relationship with topics. In addition, the spotting-based TID requires that the keywords should be easy to detect. Prediction of spotting accuracy of a word is necessary to design a keyword set for the spotting-based TID.

In general, the spotting of a short word generates many false alarms (FA's). Short words are omitted from the keyword set of the TID application based on the length of the word[5]. However, the spotting accuracy is dependent on not only the length but also the phonemes composed of the keyword. This paper proposes a method of predicting the accuracy of keyword spotting in terms of FA count and spotting score of correct detections, based on the simulation of the keyword spotting.

2. KEYWORD SPOTTING SYSTEM

The HMM-based recognizer carries out the keyword spotting using two linguistic constraints illustrated in Figure 1. The recognizer uses 26 phonemes as acoustic model including silence. The first constraint (a) allows any sequences of phonemes, and the second one (b) supposes that the last word is a keyword. The existence likelihood of a keyword, $KL(kw)$, is defined as

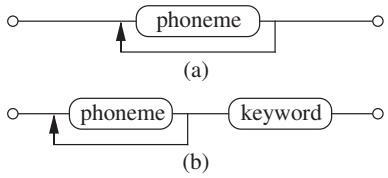


Figure 1. Linguistic constraints for keyword spotting.

$$KL(kw) = \frac{L_{kw}(kw) - L_{ph}(kw)}{length_of_keyword - 2.5} \quad , \qquad (1)$$

where $L_{kw}(kw)$ and $L_{ph}(kw)$ are logarithmic recognition scores from the beginning of the speech to an arbitrary time under constraints (b) and (a), respectively. The *length_of_keyword* is defined as the number of phoneme in the keyword. The denominator is the keyword length modified by subtraction of -2.5 in order to suppress FA's for short keywords. A keyword is detected at the time when $KL(kw)$ is larger than a threshold.

3. METHOD

3.1. Prediction of the FA count

A new measure for predicting the FA count is proposed. The spotting simulation based on the DP matching between a keyword, w, and phoneme sequences generated by the phoneme-based language model, estimates possibility of FA for the keyword. Phoneme trigram is used as the language model. The matching process for each keyword is repeated for K phoneme sequences, $s_k(k = 1, \cdots, K)$, each of which is composed of M phonemes, $ps_{k1}, ps_{k2}, \cdots, ps_{kM}$. In this paper, K=100,000. The probability that a subsequence of s_k is recognized as w, $P(w|s_k)$, corresponds to probability of FA occurrence.

The spotting simulation needs statistics of phoneme recognition of the spotting system. It is characterized in terms of three kinds of probabilities, $S(p_j|p_i)$, $D(p_i)$, and $I(p_j)$. $S(p_j|p_i)$ is a probability that a phoneme p_i is substituted by p_j. $D(p_i)$ is a probability that a phoneme p_i is deleted. $I(p_j)$ is a probability that an inserted phoneme is p_j if any. Here, these probabilities are constrained by

$$\sum_{j=1}^{N} S(p_j|p_i) + D(p_i) = 1, \sum_{j=1}^{N} I(p_j) = 1, \qquad (2)$$

where N is the kind of phonemes. In the simulation N=25 because a phoneme model of the silence is not used.

Let a keyword w composed of phonemes, pw_1, $pw_2, \cdots, pw_L$. L is the phoneme length of w. The DP matching with termination free maximizes $P(w|s_k)$ using $S(p_j|p_i)$, $D(p_i)$, and $I(p_j)$ as matching costs of phonemes as follows.

1) $g(1,1) = S(pw_1|ps_{k1})$,
   $g(1,j) = 0 \quad (j = 2, \cdots, L)$
2) repeat step 3), 4), and 5) for $i = 2, \cdots, L+D$
3) $j1 = \max(1, i-D), j2 = \min(L, i+D)$
4) repeat step 5) for $j = j1, \cdots, j2$
5)

$$g(i,j) = \max \begin{cases} g(i-1,j-1) \times S(pw_j|ps_i) \\ g(i-1,j) \times D(ps_i) \\ g(i,j-1) \times I(pw_j) \end{cases} \quad (3)$$

6) $P(w|s_k) = \max(g(L-D,L), \cdots, g(L+D,L))$

Here, D controls possible occurrences of insertions and deletions. In this paper we tried D=0 and 3. D=0 means neither insertions nor deletions occurs. The probability of FA is given by

$$FP = \sum_{k=1, w \notin s_k}^{K} P(w|s_k). \quad (4)$$

A large FP means that many FA's are expected and a short keyword also invokes many FA's, too. In oder to make a measure comparable with the keyword length,

$$F = -\log FP \quad (5)$$

is used a measure for predicting the FA count of a keyword.

3.2. Prediction of the correct detection score

The probability that a phoneme sequence of a keyword, pw_1, pw_2, $\cdots$, pw_L, is correctly recognized as the keyword w is simply estimated by

$$P(w|w) = \prod_{i=1}^{L} S(pw_i|pw_i). \quad (6)$$

This probability is also transformed into

$$C = -\log P(w|w) \quad (7)$$

for easy comparison with the keyword length. This measure is used to predict the spotting score of correct detections.

4. MODEL TRAINING AND DATA

4.1. Phoneme models for spotting

We used the HTK tools provided by Entropic to obtain the phoneme models. The feature vector is composed of 12 melcepstrum, 12 delta-melcepstrum, and delta-energy. Each phoneme model has 5 states with 4 mixtures and 3 loops. The phoneme models were trained with 13.6 hour phonetically balanced sentences by 64 speakers in the ASJ Continuous Speech Corpus for Research[6]. Accuracy of the phoneme recognizer is 54% for the first 10 news utterances in speech data mentioned later.

Three kinds of phoneme recognition statistics, $S(p_j|p_i)$, $D(p_i)$, and $I(p_j)$, are necessary to the spotting simulation mentioned in 3.1. These probabilities are obtained based on the confusion matrix from a preliminary experiment of continuous speech recognition using the same phoneme models. Clean speech of a TV news program was recognized in the preliminary experiment. Total speech duration is 19.0 minutes. The confusion matrix is generated by the HTK tool.

4.2. Language model for generating phoneme sequences

The trigram model of the phoneme sequence are trained with 15,571 sentences of JNAS (Japanese Newspaper Article Sentences) corpus[7]. The SLM toolkit[8] is used to build the phoneme trigram. The test set perplexity is 7.96 for 101 sentences from another set of JNAS.

4.3. Speech data for spotting

Speech data was collected from a TV news program. Recorded speech was automatically divided into 610 speech fragments, called utterances, based on pause longer than 180msec. The speech material for spotting is different from the training data for obtaining the phoneme recognition statistics. The TV news speech includes very noisy utterances, such as interview overlapped by environmental noise and narration with background music, and so on. These noisy utterances were manually removed from speech material for experiments. The total number of clean utterances are 512, and the total duration is 26.0 minutes. They includes several non-professional speakers, such as interviewees on the street, unclearly speaking politicians, and so on, as well as professional news announcers.

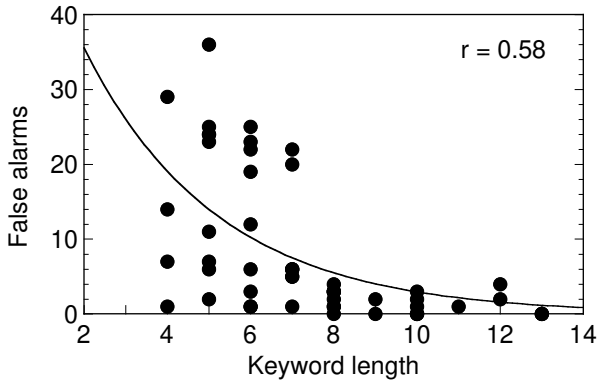
Fifty four words frequently appeared in this news speech, such as '*dairishomei* (representative signature)', '*gakko-kyushoku* (school lunch)' and so on, were manually selected as keyword. Average phoneme length of 54 keywords is 7.46. The occurrence of the keywords in the clean utterances is 573 in total and 24.5 [/hour/keyword] in average.

5. RESULTS

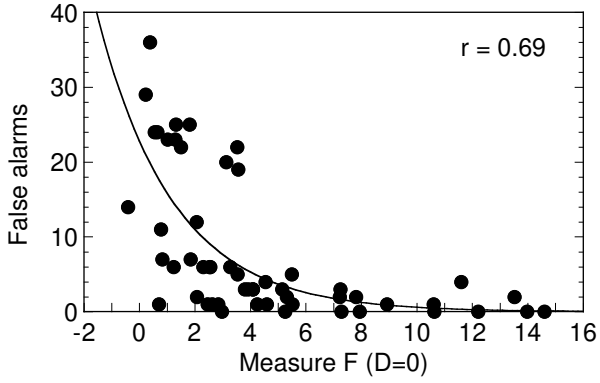
5.1. Prediction of the FA count

5.1.1. Prediction based on single measures

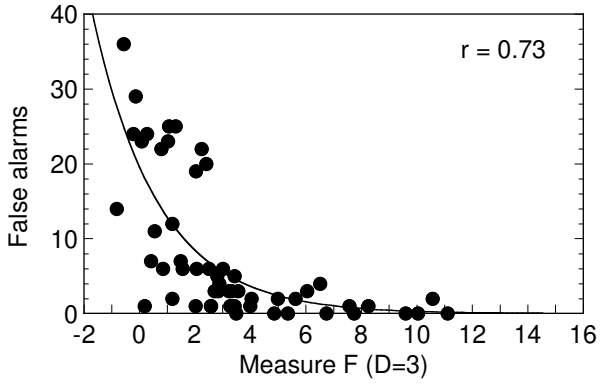
The keyword spotting was carried out with a threshold and the average detection rate was 76%. Figure 2 shows scattering plots of the FA count as function of a predictive measure. A black circle corresponds to a keyword. This figure compares the keyword length, L, and two simulation-based measures, F's with D=0 and D=3. D=0 means that neither insertions nor deletions occur in the spotting simulation. Each distribution of the FA count is approximated by the fitting curve of $a\exp(b*M+c)$. M is a predictive measure. Predictive performance for the



(a) keyword length L



(b) measure F (D=0)



(c) measure F (D=3)

Figure 2. Prediction of FA count based on single measures.

FA count was evaluated in terms of the correlation coefficient (r) between observations and predicated values. The correlation coefficients by L, $F(D$=0), and $F(D$=3) were 0.58, 0.69, and 0.73, respectively. Table 1 shows the average prediction errors of FA count for three measures.

The simulation-based measure F's gave better predictive performance rather than L that is simple lexical information. Introduction of insertion and deletion probabilities in the spotting simulation improved the prediction of the FA count.

Table 1. Prediction errors of FA count based on single measures.

	measure	error
(a)	keyword length L	7.71
(b)	measure F (D=0)	6.89
(c)	measure F (D=3)	6.48

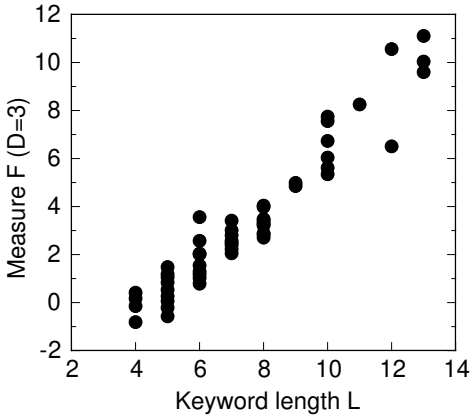


Figure 3. Relationship of keyword length and measure F.

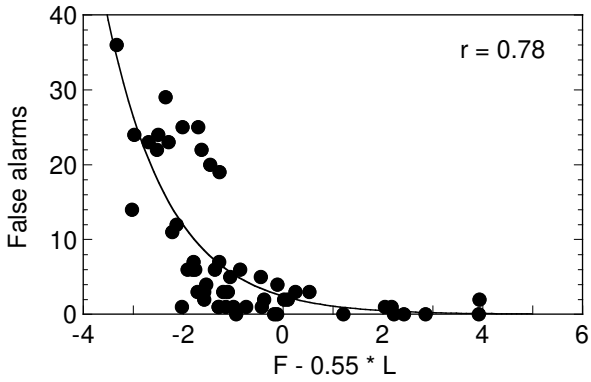
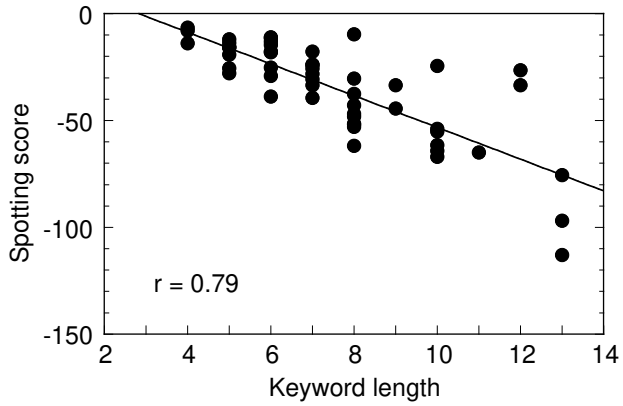


Figure 4. Prediction of FA count based on a combined measure.

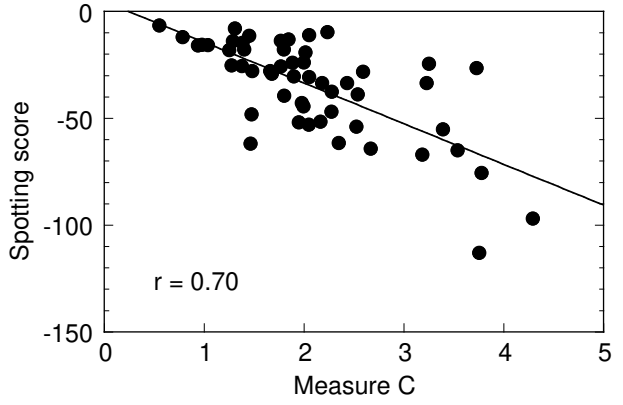
5.1.2. Prediction based on a combined measure

Figure 3 shows relationship of the keyword length L and $F(D$=3). Although the measure F is much related to L, F is not completely dependent on L. We tried to combine two measures to predict the FA count. The FA count is predicted by an equation $a\exp(b*(F+d*L)+c)$. The parameters, a, b, c, and d, were optimized by criterion of the least squared errors of the FA count. The parameter d controls the weight of L for predicting FA count. Optimized d was -0.55. Figure 4 shows prediction of the FA count based on a combined measure, $F-0.55*L$. The fitting curve, $a\exp(b*M+c)$, is also obtained by the optimization process.

The correlation coefficient by the combined measure is 0.78. The average prediction errors of FA count is 5.94. The combined measure reduced the prediction errors by 23% against the prediction based on only L.



(a) keyword length L



(b) measure C

Figure 5. Prediction of spotting score for correct detections.

Table 2. Prediction errors for spotting score of correct detections.

	measure	error
(a)	keyword length L	13.7
(b)	measure C	16.0

5.2. Prediction of the correct detection score

The spotting score of a keyword was also predicted by two kinds of measures, the keyword length L and the measure C. The average scores of correct detections are plotted in Figure 5 for 54 keywords, supposed that all existing keywords are detected. Predictive performance was also evaluated in terms of the correlation coefficient and the prediction errors. The correlation coefficients with linear fitting are 0.79 and 0.70 for L and C, respectively. L has stronger relationship with the spotting scores of correct detections. From the average prediction errors, shown in Table 2, L is more effective to predict the score than C. We tried a combined measure of L and C as same as 5.1.2. However, it did not improve the prediction of the scores.

6. CONCLUSION

New measures are proposed in order to predict accuracy of keyword spotting and evaluated in comparison with the phoneme length of the keyword. The simulation-based measure is effective to predict the FA count in the keyword spotting. On the other hand, a measure based on probabilities of the correct phoneme recognition is not effective to predict the spotting score of the correct detection. The spotting score, that is phoneme recognition likelihood, is much dependent on clearness of the utterance as well as phonemes composed of the keyword. Familiar words frequently appeared in TV news are sometimes uttered unclearly. It is difficult to predict the spotting score of the keyword using lexical information or statistics of the recognition system.

The prediction of the FA count is motivated by keyword selection for spotting-based topic identification (TID). It is necessary that effectiveness of the FA count prediction is evaluated in the TID task. In the keyword spotting, normalization of the spotting score is very important to balance correct detections and false alarms. Prediction of the correct detection scores is useful to normalize the score and to set up a threshold for detection.

Acknowledgment

This paper used CD *Mainichi-shinbun* '91-'94, ASJ Continuous Speech Corpus for Research, RWC text data base (RWC-DB-TEXT-95-1).

REFERENCES

[1] B. Peskin, S. Connolly, L. Gillick, S. Lowe, D. McAllaster, V. Nagesha, P. Mulbregt, and S. Wegmann : "Improvements in switchboard recognition and topic identification", Proc. of ICASSP '96, pp.303-306 (1996).

[2] Y. Itoh, J. Kiyama, and R. Oka : "Speech Understanding and Speech Retrieval for TV News by Using Connected word Spotting", Proc. of Eurospeech '95, pp.2141-2144 (1995).

[3] J.T. Foote, G.J.F. Jones, K. Sparck Jones, and S.J. Young : "Talker-Independent Keyword Spotting for Information Retrieval", Proc. of Eurospeech '95, pp.2145-2148 (1995).

[4] D.A.James : A system unrestricted topic retrieval from radio news broadcasts", Proc. of ICASSP '96, pp.279-282 (1996).

[5] Y. Yamashita, T. Tsunekawa, and R. Mizoguchi : "Topic Recognition for News Speech Based on Keyword Spotting", Proc. of ICSLP '98, pp.839-842 (1998).

[6] "Continuous Speech Corpus for Research", CD-ROM, Vol.1-3, Acoustical Society of Japan (1993).

[7] http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html

[8] P.Clarkson, R.Rosenfeld : "Statistical Language Modeling Using the CMU-Cambridge Toolkit", Proc. of Eurospeech '97, pp.2707-2710 (1997).