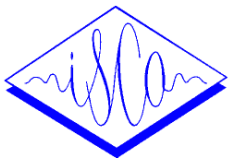


3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

EVALUATION OF UNKNOWN WORD PROCESSING IN A SPOKEN WORD RECOGNITION SYSTEM

Atsuhiko KAI and Seiichi NAKAGAWA

Department of Information and Computer Sciences,
Toyohashi University of Technology,
Tenpaku-cho, Toyohashi, 441 Japan

ABSTRACT

Detection of an unknown word or non-vocabulary word uttered by the user is necessary in realizing an effective spoken language user-interface. This paper describes the evaluation of an unknown word processing method for a subword unit based spoken word recognizer. We have assessed the relationship between the word recognition accuracy of a system and the detection rate of unknown words both by simulation and by experiment of the unknown word processing method. We have seen that the resultant detection accuracies using the unknown word processing are significantly influenced by the original word recognition accuracy.

1 INTRODUCTION

While the performance of the speech recognition systems for a large vocabulary words have been argued in previous studies, there are many difficult issues for realizing a robust system. In particular, handling the recognition error and the user utterance out of the vocabulary is essential problems that have to be considered. The user-interface functions for rejecting the misrecognized result or detecting an unknown word are desired especially for constructing a reliable robust system and will make possible the extensive application of the speech recognition system if they work effectively.

To deal with unknown words, there is the idea to model non-keyword speech as several representative acoustic models. Such a simple approach has been often used for the word spotting [1]. On the other hand, in large vocabulary oriented systems, the unknown word can be modeled by an arbitrary sequence of subword units since the word models are usually constructed by concatenating each of the subword unit models as specified in the word lexicon [1, 2, 4, 5]. While several experiments using the rejection method based on the latter approach have been reported, the relationship between the recognition accuracy and the expected detection rate of the unknown word is not clear.

We attempt to assess the relationship between the word recognition accuracy of a system and the detection rate of unknown words by the simulation of the unknown word detection method based on the sequence of subword units. We compare the simulation results with the experimental results using an actual speech recognition system.

2 DETECTION OF UNKNOWN WORD IN SPEECH

The method for detecting unknown words is formulated as follows. We assume that the unknown word is modeled by an arbitrary sequence of syllables. Let the maximum word verification score in all of vocabulary words be $\hat{S}_b$. The verification score as an unknown word $\hat{S}_u$ can be obtained by a subword unit based spoken word recognition with bigram or trigram constraint or no language model. We can simply decide whether one had uttered the word that doesn't exist in the system's lexicon by the following relation:

$$\hat{S}_u - \hat{S}_b > threshold \ , \quad \textbf{if } input \ word \notin \text{vocabulary}.$$

In general, $\hat{S}_u \geq \hat{S}_b$ is always satisfied and we should have a lower limit: $threshold > 0$. We describe the simulation method and the evaluation result for the unknown word detection method shown above in the next section.

3 EVALUATION OF UNKNOWN WORD PROCESSING BY SIMULATION

If the recognition accuracy for syllables is nearly 100%, the above method will achieve the complete detection for unknown words with a smaller *threshold*, $\simeq 0$. Since we should consider a real speech recognition system in which complete syllable recognition could not be achieved, we should decide on an appropriate threshold in terms of a trade off between the word recognition accuracy (or rejection rate of correct utterances) and the detection rate of unknown words. Thus we have assessed the relationship between both of these. Further experiments by a real system are described in section 3.3.

3.1 Simulation method

We assume that the word model scores obtained by a recognizer follow two normal distributions: $N(\mu_1, \sigma_1^2)$ for correct(uttered) words, and $N(\mu_2, \sigma_2^2)$ for incorrect words. Then, the word recognizer has a recognition accuracy as a function of both the distributions and the category size, independent of the similarity between the words or the syllables [3]. The score distributions for the syllable category models are assumed to follow two distributions: $N(\frac{\mu_1}{L}, \frac{\sigma_1^2}{L})$ for correct syllables (i.e., the word entry uttered is represented by the sequence of syllables) and $N(\frac{\mu_2}{L}, \frac{\sigma_2^2}{L})$ for the incorrect syllables, where L means the number of syllables in the target utterance.

Since the unknown word detection method described in the previous section is based solely on the likelihood scores for both the registered word and the unknown word, we

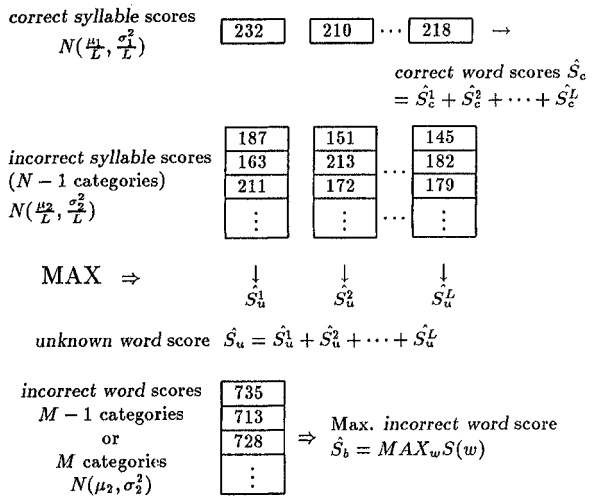


Fig. 1 Generation of word verification scores

can evaluate the performance of the method using the assumed model scores. However, deriving the equation that accurately obtains the ideal unknown word detection rate for a given parameter is not easy for the above assumption. Therefore, the evaluation is performed by generating the recognition scores based on the above score assumption, which the number of scores to be generated differs whether the input is assumed as the registered word or unknown word.

Let N and M be the number of syllable categories and that of the registered words, respectively. The scores to be generated for each trial iteration are *correct syllable* scores (L times), *incorrect syllable* scores ($(N-1) \times L$ times), and *incorrect word* scores ($M-1$ times when assuming a known word is inputted, otherwise M times). The *correct word* score $\hat{S}_c$ (which is not generated from the correct word distribution $N(\mu_1, \sigma_1^2)$) and *unknown word* score $\hat{S}_u$ are respectively the sum of L *correct syllable* scores and the sum of L best syllable scores each of which is the best score among a *correct syllable* score and $N-1$ *incorrect syllable* scores as shown in Figure 1. For convenience, let the maximum score among *incorrect word* scores be $\hat{S}_b$. The evaluation by the simulation is performed in respect with both of the following conditions.

- In case of *known word* input:

  known word false rejection *if* $\hat{S}_u - \hat{S}_b > \theta$
  correct recognition *if* $\hat{S}_c > \hat{S}_b$
  misrecognition *otherwise*

- In case of *unknown word* input:

  correct detection *if* $\hat{S}_u - \hat{S}_b > \theta$
  misrecognition *otherwise*

3.2 Word accuracy and unknown word detection rate

We preliminarily evaluated the relationship between the number of categories and word accuracies assuming several systems which have different recognition accuracies. The plot of such a relationship was obtained by the simulation in

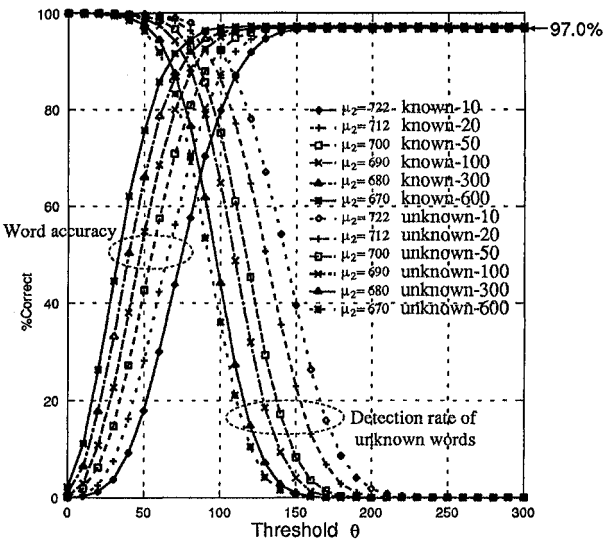


Fig. 2 Word accuracy and unknown word detection rate ($\mu_1 = 850, \sigma_1 = 25, \sigma_2 = 40, W.Acc. = 97.0\%$)

the case of known word input with *threshold* $= \infty$. As the result of preliminary evaluation, the parameters and conditions for the simulation was chosen as follows (Japanese consists of about 110 syllables):

Number of word categories: $10 \sim 600$
Number of syllable categories: 100
Number of syllables in a word: 4
The system parameters for word recognizer:
$\mu_1 = 850, \sigma_1 = 25, \sigma_2 = 40$ (fixed)
$\mu_2 =$ variable

Then, the curves of the relationship among the threshold θ, word accuracy and detection rate of unknown words were plotted, assuming that the system has original word accuracies of 97% or 95% with vocabulary size of 10~600. The result for the assumption of 97% word accuracy is shown in Figure 2. The figure shows the relation that the word accuracy decreases as the threshold changes to the way the detection rate of unknown words increases. Table 1 shows the detection rate of unknown words in case of the time that the word accuracy degrades by 1~10% using the unknown word processing method. This table shows that the different systems, which originally have word accuracies of 97% and 95% respectively, causes a significant difference of 15~20% in the detection rate. However, the rate is not sensitive to the vocabulary size. Roughly speaking, we can say that the unknown words are detected at the rate of 50% or more when the original word accuracy is about 97% and the false rejection rate of the registered word is 2% or less. On the other hand, if a desired detection rate is 99% or 97%, we can say that the registered words will be correctly recognized at the rate of 50% or more when the original word accuracy is about 97%, and almost all others (i.e., 97 – 50%) may be rejected.

Table 1 Relationship among the word accuracy, vocabulary size and unknown word detection rate

(a) Original word accuracy = 97%

	After unknown word processing					
	96%	94%	92%	87%	60%	50%
M	unknown word detection rate					
10	54	70	78	87	98	98
20	51	67	77	87	98	99
50	45	64	76	87	98	99
100	49	71	80	89	98	99
300	44	66	76	87	98	99
600	54	73	81	90	98	99

(b) Original word accuracy = 95%

	After unknown word processing					
	94%	92%	90%	85%	60%	50%
M	unknown word detection rate					
10	34	55	65	78	94	97
20	26	49	63	79	95	97
40	40	60	69	82	96	98
100	32	53	65	79	96	98
200	32	57	67	81	96	98
500	33	58	69	83	97	98

3.3 Compensation of parameters for simulation

The simulation method described in the previous section has the assumption that the incorrect word score distribution is independent of the correct syllable score distribution. This assumption may cause a different result between simulation and experiment by an actual system because an incorrect word may include syllables which are also included in the correct word. Therefore, in particular, if all the syllables do not occur equivalently, we should consider the fact that the incorrect word score involves the correct syllable score. The compensation for the different assumption by modifying the simulation parameter can be considered as follows:

- Method-1
  The test set syllable perplexity or the number of equivalent syllable categories in the system's lexicon is taken into account by simulation instead of the true number of syllable categories.

- Method-2
  The mean value of the incorrect syllable score distribution becomes, in practice, of lower value than that of the previous assumption if the incorrect word score distribution is assumed to be adequate (in other words, the mean of the incorrect word distribution is slightly larger than μ_2). Thus the approximative compensation can be done by directly adjusting the parameter for the incorrect syllable as:

$$\mu_2^{(s)} = \frac{\mu_2}{L} \implies \mu_2^{(s)'} = \frac{\mu_2}{L} - C_0$$

  where C_0 is the adjustment value.

- Method-3
  If we assume that all words tend to be misrecognized as the word that has the least difference in syllable transcription, the incorrect syllable score distribution is approximately compensated as follows:

  1. Estimate the average Hamming distance between the confusable words defined as:

$$D_h = \frac{1}{M}\sum_{w_i} \min_{j,i\neq j} d(w_i, w_j)$$

     where $d(w_i, w_j)$ is Hamming distance in terms of syllable sequences between the word w_i and w_j.

  2. Adjust the parameters for the incorrect syllable score assuming that the incorrect word has included same $(L - D_h)$ syllables as the correct word on average.

$$\mu_2^{(s)} = \frac{\mu_2 - (L - D_h)\mu_1/L}{D_h}$$
$$\sigma_2^{2(s)} = \frac{\sigma_2^2 - (L - D_h)\sigma_1^2/L}{D_h}$$

The simulation results which employ the compensation method-2 is shown in the next section, while the preliminary tests showed that both compensation method-2 and -3 could obtain almost the same effect.

4 EXPERIMENTS OF UNKNOWN WORD DETECTION

To evaluate the relationship between the speech recognition performance and the unknown word detection rate, the experiment using an actual word recognizer was performed. We employed the Viterbi based continuous speech recognition algorithm using the syllable unit based HMMs. The syllable unit HMMs as the subword acoustic model, which have 5 states, 4 Gaussian densities and 4 discrete duration distributions each, were trained using the 503 sentence utterances of 6 males and 216 phonetically balanced word utterances of 10 males included in the *ATR speech database*. Furthermore, a total of 4500 sentence utterances of 30 males are used for additional training by the MAP adaptation method without segmentation[6].

We used a part of the *Tohoku-Matsushita speech database* as the test set, which consists of isolated spoken words of a 212 word vocabulary and is uttered by 15 male speakers. The test data set was down-sampled to 12kHz and 14 LPC cepstral coefficients were obtained by analyzing speech for every 8msec. These coefficients were transformed to 10 LPC mel-scaled cepstral coefficients and their regressive coefficients.

For testing the performance of the unknown word detection, 50 or 100 words out of a 212 word vocabulary were selected at random and registered to the system's lexicon. The experiments were carried out using different sets of word lexicons with the same vocabulary size. The score as the unknown word model is obtained by the syllable unit based continuous speech recognizer with no language model. With common threshold values for every speakers, the relationship between the false detection rate of unknown words for

utterances of the registered word and the correct detection rate of unknown words were evaluated.

Figure 3 shows the experimental results, where solid lines show the simulation results for the systems which are assumed to have the original word accuracy of 95%, 97% and 99%, respectively. Note that the simulation results were slightly different from the results shown in Table 1 because a parameter of the syllable score distribution was compensated as described in section 3.3 (using Method-2, $C_0 = 20$). For example, in this simulation test condition ($C_0 = 20$), the syllable recognition accuracy is about 64% without language model or lexicon when the word accuracy is about 97% for a 100 word vocabulary. This condition approximates the real syllable recognition with no language model well. This figure shows a similar relation for both simulation and experiments, while the difference in the content of system's lexicon slightly affects the performance of the unknown word detection. We can also say, as mentioned in the previous section, that the performance of the unknown word detection method is significantly affected by the original word recognition accuracy.

5 CONCLUSION

The performance of the detection of unknown word based on the subword unit was evaluated by simulation and experiments using an actual system. The simulation and experimental results showed almost the same relationship between the known word false rejection rate and the unknown word detection rate. As the result of the simulation and experiment, we have seen that the detection rate is mostly affected by the word accuracy of a recognizer and typically more than 50% of the unknown words are correctly detected when the system has original word accuracy of about 97% and the known word false rejection rate is 2% or less. On the other hand, if a desired detection rate is 99% or 97%, we can say that the registered words will be correctly recognized at the rate of 50% or more when the original word accuracy is about 97%, and almost all others(97 − 50%) may be rejected.

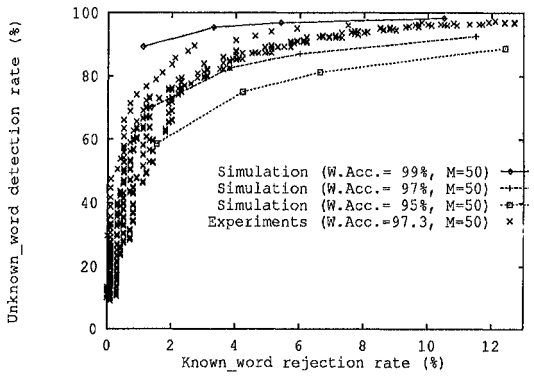
References

[1] R. A. Sukkar and J. G. Wilpon : A two pass classifier for utterance rejection in keyword spotting, Proc. ICASSP, pp.II-451–454 (1993).

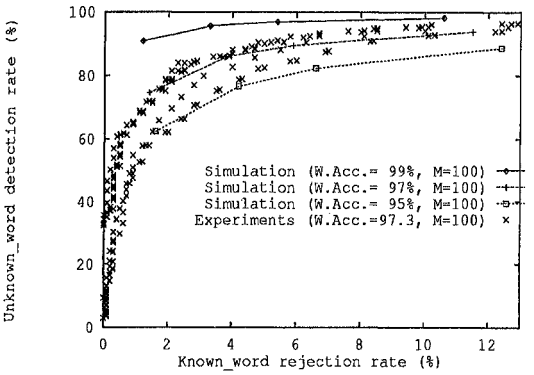
[2] A. Kai and S. Nakagawa : Improvements of the Japanese continuous speech recognition system - SPOJUS-SYNO - and its evaluation, IEICE Technical Rep., SP93-20, pp.49–56 (1993.6)(in Japanese).

[3] S. Nakagawa and I. Murase : Relationship among phoneme/word recognition rate, perplexity and sentence recognition and comparison of language models, Proc. ICASSP, pp.I-589–592 (1992).

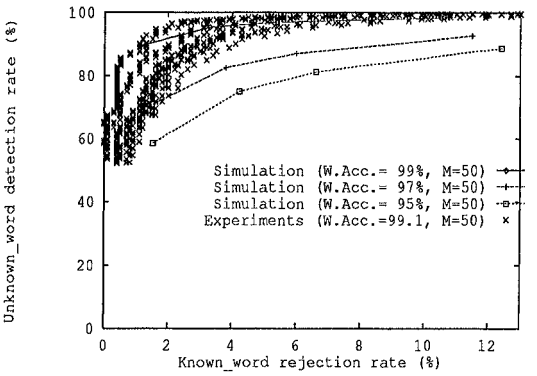
[4] K. Kita, T. Ehara and T. Morimoto : Processing unknown words in continuous speech recognition, Trans. IEICE, Vol.E74, No.7, pp.1811–1815 (1991).



(a) Vocabulary size = 50, Original word accuracy = 97.3%



(b) Vocabulary size = 100, Original word accuracy = 97.3%



(c) Vocabulary size = 50, Originial word accuracy = 99.1%

Fig. 3 Performance of the unknown word detection in isolated word recognition
(a),(b) : 10 mel-cepstral coefficients
(c) : mel-cepstrum + regressive coefficients

[5] K. Itou, S. Hayamizu and H. Tanaka : Detection of unknown words and automatic estimation of their transcriptions in continuous speech recognition, Proc. ICSLP, Banff (1992).

[6] Y. Tsurumi and S. Nakagawa : An unsupervised speaker adaptation method for continuous parameter HMM by maximum a posteriori probability estimation, (to appear in this conference).