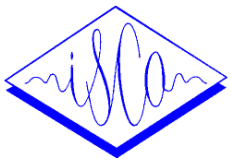

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

PERFORMANCE EVALUATION IN SPEECH RECOGNITION SYSTEM USING TRANSITION PROBABILITY BETWEEN LINGUISTIC UNITS

*Takashi Otsuki**, *Shozo Makino**, *Toshio Sone*** *and Ken'iti Kido**

*Research Center for Applied Information Sciences, Tohoku University,
2-1-1, Katahira, Aoba, Sendai, 980 Japan
**Research Institute of Electrical Communication, Tohoku University,
2-1-1, Katahira, Aoba, Sendai, 980 Japan

ABSTRACT

This paper describes performance evaluation in speech recognition system which uses transition probability between linguistic units. The lower limit of word recognition score is predicted based on phoneme recognition score and number of word pairs with short distance in a vocabulary defined by linguistic information. But it is difficult to calculate it when transition probability is used as linguistic information. We propose new algorithm to calculate it when bigram or trigram of linguistic units is used. Using this algorithm, we carry out performance prediction in speech recognition which uses bigram or trigram. Recognition score for word with 5 phonemes is more than 26% using bigram, more than 71% using trigram and more than 95% using a dictionary when phoneme recognition score is 90%, where bigram and trigram of phonemes are estimated from the 5,317 Japanese popular words. Recognition score of sentence composed of 11 words is more than 4.3% using bigram, on the other hand, more than 67% using trigram, when word recognition score is 80%, where bigram and trigram are estimated from 136 sentences represented with 18 kinds of speech.

I. INTRODUCTION

Generally, there are several errors in a sequence of linguistic unit obtained by speech recognition. It is effective to use the linguistic information such as a dictionary or transition probability between linguistic units in order to correct these errors. It is essential for current speech recognition system to use the linguistic information for error correction. However, it is meaningless to compare among these systems by their recognition scores since task complexity of each system is different. Therefore, it is important to develop a performance evaluation method of speech recognition system using linguistic information which reflects the complexity of tasks.

K.Abe et al. derived a theoretical formula to get the lower limit of word recognition score of recognition system using dictionary based on recognition score of linguistic units and the number of word pairs with short distance in a dictionary[1]. S.Makino and K.Kido carried out an investigation of the properties of phoneme pairs for distinguishing a word pair with short distance in most frequently used Japanese words[2]. S.Nakagawa made clear the relationship between phoneme recognition score and word recognition score by simulation[3]. S.Nakagawa et el. proposed an evaluation method for continuous speech recognition systems, and they made clear the relationship between task complexity and sentence recognition score[4].

In this paper, we present a method for performance evaluation in speech recognition system which uses transition probability between linguistic units based on Abe's theoretical formula[1].

II. Performance evaluation of word recognition system using a dictionary

At first we assume an input word "abcde" included in a word dictionary is recognized as a phoneme sequence of "abcfe" because of misrecognition of phoneme /d/ as /f/. In case of correct segmentation, we can determine that the uttered word consists of 5 phonemes, so we can only consider the items consisting of 5 phonemes in the word dictionary. If the word dictionary contains only one word "abcde" at the minimum distance from the "abcfe", the recognition system can define the word "abcde" as a recognition result. In consequence, the phoneme recognition error is corrected. We define the word distance between a word X and a word W by

$$d(X,W)\equiv\sum_{i=1}^{n} d(X_i,W_i)\ ,\ d(X_i,W_i)\equiv\begin{cases}1(X_i\neq W_i)\\0(X_i=W_i)\end{cases} \quad (1)$$

where $X_i$and W_i stand for phoneme and n is the length of the word. Next, let us suppose the word "abcde" is recognized as a word "gbcde". If there are two items "abcde" and "hbcde" at the minimum distance from the "gbcde", the probability of the error correction is 0.5. In general, if there are n words at the minimum distance from the recognized phoneme sequence, the probability of the error correction is $\frac{1}{n}$. If a word "abcde" is recognized as a word "hbcde" included in the word dictionary, the word recognition result is determined as the word "hbcde". In this case, it is impossible to correct the phoneme recognition error. Therefore, from the view of the word X, the whole phoneme sequences with the same length as X is divided into the following three groups (Fig.1).

(1) $\Omega_1(X)$: Even if the word X is recognized as the word in this group, the errors always can be corrected by the dictionary.

(2) $\Omega_2(X)$: If the word X is recognized as the word in this group, the error can be corrected by the dictionary, but not always.

(3) $\Omega_3(X)$: If the word X is recognized as the word in this group, the error can't be corrected by the dictionary.

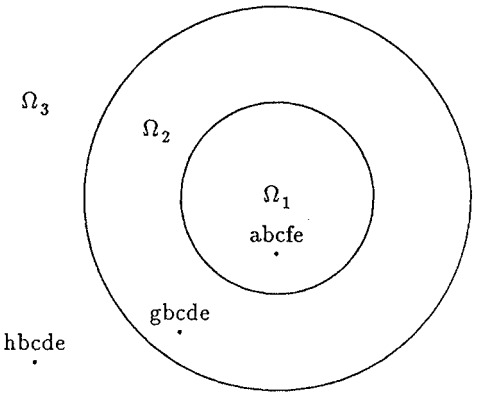

Fig.1 Three groups relating to word "abcde" in Ω

So the average error rate of word recognition P_e is given by

$$P_e=\frac{1}{N}\cdot\sum_{X\in\Omega_D}\Big(\sum_{W\in\Omega_2(X)}P_e(W|X)\cdot\Big(1-\frac{1}{n_{d(X,W)}(W)}\Big)+\sum_{W\in\Omega_3(X)}P_e(W|X)\Big) \quad (2)$$

where $P_e(W|X)$ is the probability that the word X is recognized as the word W, $n_d(Y)$ is the number of the words at a distance of d from the word Y, N is the number of words in the dictionary Ω_D.

According to the Abe's theory, when the likelihoods of the phoneme lattice is assumed to be distributed according to normal distribution such as $N(\mu,\sigma^2)$ for a correct phoneme and $N(0,\sigma^2)$ for other phonemes, the lower limit of the average word recognition score can be predicted based on phoneme recognition score γ and n_d: the average of the number of words at a distance of d. γ is given by

$$\gamma=\int_{-\infty}^{\infty}\Big[\Phi(\frac{x}{\sigma})\Big]^{M-1}\frac{1}{\sqrt{2\pi}\sigma}\cdot\exp\Big\{-\frac{(x-\mu)^2}{2\sigma}\Big\}dx \quad (3)$$

where M is the number of kinds of phoneme and $\Phi(x)$ is the standard normal distribution function as follows:

$$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \quad (4)$$

The lower limit of the average word recognition score P_c is given by

$$P_c=1-\sum_{d=1}^{n}n_d\lambda_d \quad (5)$$

where λ_d is given by

$$\lambda_d\equiv\Phi(-\frac{\sqrt{d}\mu}{\sqrt{2}\sigma}) \quad (6)$$

According to this theory, the relationship between phoneme or syllable recognition score and word recognition score of the system using a dictionary is shown in Fig.2, where the dictionary with 5,317 Japanese popular words is represented with phonemes or syllables. Recognition score for word is at least 95.9% using the dictionary when phoneme recognition score is 90%. And word recognition score is at least 96.3% using the dictionary when syllable recognition score is 90%.

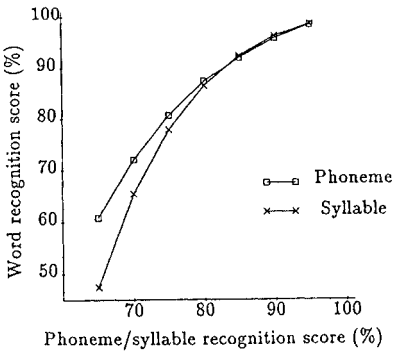

Fig.2 The relationship between recognition score of phoneme/syllable and lower limit of word recognition score with dictionary

III. Word recognition using transition probability

Let us consider that the dictionary is generated using transition probability between linguistic units. Let us consider that the transition probability is binary (1 or 0). Furthermore we call the first-order and second-order transition probabilities as "bigram" and "trigram". We show an example of the trigram of phonemes /a/ , /b/ and /c/ in Fig.3. In this case, there are 6 possible phoneme sequences with length of 3 phonemes: "abc","bcc","bca","ccc","cca" and "cab". If the word "abc" is recognized as the sequence "aba", there is only one possible phoneme sequence "abc" at the minimum distance from "aba" (Table.1). Therefore, the phoneme recognition error in the "aba" is completely corrected, the sequence "aba" belongs to the group Ω_1 in Fig.1. If the word "abc" is recognized as "bbc", there are two possible sequences, "abc" and "bcc", at the minimum distance from "bbc". So the probability of the error correction is 0.5 and the sequence "bbc" belongs to the group Ω_2 in Fig.1. If the word "abc" is recognized as "aab", there is only one possible sequence "aab" at the minimum distance from "aab" and it is impossible to correct the error. So the sequence "aab" belongs to the group Ω_3 in Fig.1.

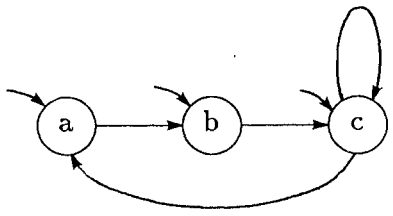

Fig.3 An example of transition probability among three phonemes

Table 1 Mutual distance of sequences which consists of three phonemes

	aba	bbc	aab
abc	1	1	2
bcc	3	1	3
bca	2	2	3
ccc	3	2	3
cca	2	3	3
cab	3	3	1

From this argument, we can evaluate the word recognition score when using the transition probability by considering the possible sequences as the items of the dictionary.

IV. The algorithm for calculating n_d

Using a dictionary Ω which contains N words, n_d is given by

$$n_d=\frac{\sum_{W\in\Omega} N_d(W,\Omega)}{N} \tag{7}$$

$$N_d(W,\Omega)=\sum_{X\in\Omega} f_d(W,X) \tag{8}$$

$$f_d(W,X)=\begin{cases}0 & \text{if } dist(W,X)\neq d\\ 1 & \text{if } dist(W,X)=d\end{cases} \tag{9}$$

In this case, the computation amount is $O(N^2)$ and there is no difficulty to calculate n_d.

But using the transition probability, the number of possible sequences, regarded as the dictionary, N_t is approximately given by

$$N_t\approx B_s{}^l \tag{10}$$

where l is length of the sequence and B_s is static blanching factor per phoneme when using transition probability. In this case, the computation amount is $O(B_s^{2l})$. Because B_s in Japanese words is about 10, the computation amount rapidly glows up with length, so it is impossible to calculate n_d. Then we propose new algorithm to calculate n_d when using bigram or trigram of phonemes.

First, we define the following symbols.

(1) $P=\{p_1,p_2,\cdots,p_M\}$: The group of M kinds of phonemes

(2) P_s: The group of phonemes allowed to be beginning of a word

(3) P_e: The group of phonemes allowed to be end of a word

(4) $\Omega^l{}_{p_i}$: The group of possible sequences which are l phonemes long and terminate with phoneme p_i

(5) $N_d(W,\Omega)$: The number of sequences in Ω at a distance of d from W

(6) $S_d{}^l(p_i,p_j)$: The number of pairs with distance of d, where one is included in $\Omega^l{}_{p_i}$ and the other is included in $\Omega^l{}_{p_j}$

(7) N^l: The number of sequences with l phonemes long

(8) $A(p_i)$: The group of phonemes which allowed to be preceded by P_i

The average of the number of sequences at the distance of d with l phonemes long is given by

$$n_d{}^l=\frac{\sum_{p_i\in P_s,p_j\in P_e}\sum_{W\in\Omega^l{}_{p_i}} N_d{}^l(W,\Omega^l{}_{p_j})}{N^l} \tag{11}$$

$$=\frac{\sum_{p_i\in P_s,p_j\in P_e} S_d{}^l(p_i,p_j)}{\sum_{p_i\in P_e} S_0{}^l(p_i,p_i)}$$

$S_d{}^l(p_i,p_j)$ is given by

$$S_d{}^l(p_i,p_j)=\sum_{p_k\in A(P_i)}\sum_{p_m\in A(p_j)} f_d(i,j,k,m) \tag{12}$$

$$f_d(i,j,k,m)=\begin{cases}S_d{}^{l-1}(p_k,p_m) & \text{if } i=j \text{ and } d<l\\ S_{d-1}{}^{l-1}(p_k,p_m) & \text{if } i\neq j \text{ and } d>0\\ 0 & else\end{cases} \tag{13}$$

The computation amount for calculation of $S_d{}^l(p_i,p_j)$ is $O(lB_s^2)$ by dynamic programming technique. Therefore the computation amount for calculation of $n_d{}^l$ is $O(l^2B_s^4)$.

V. Performance evaluation of recognition system with transition probability

5.1 Word recognition using transition probability of phonemes

The bigram and trigram of phonemes are estimated from the dictionary which consists of 5,317 Japanese popular words represented with phonemes[5]. The 24 phonemes are shown in Table.2. The relationship between the phoneme recognition score and the lower limit of the word recognition score in case of correct segmentation is shown in Fig.4. Symbol × indicates the score with no linguistic information, symbol ■ indicates the score with bigram, symbol □ indicates the score with trigram and symbol ● indicates the score with dictionary. For example, recognition score for word with 5 phonemes is more than 26% using the bigram, more than 71% using the trigram and more than 95% using the dictionary when phoneme recognition score is 90%. This results show effectiveness of the trigram comparing to the bigram.

5.2 Sentence recognition using transition probability of words

The transition probability between words is often used in a text dictation system. Then we apply the

Table 2 Phonemes and their manner of articulation

Manner of articulation		Phoneme
Consonants	Voiceless stops	/p/,/t/,/k/
	Voiced stops	/b/,/d/,/g/
	Voiceless fricatives	/s/,/h/
	Voiceless affricate	/c/
	Voiced fricative	/z/
	Liquid	/r/
	Nasals	/m/,/n/,/ŋ/
Semivoewls		/j/,/w/,/y/
Vowels		/a/,/o/,/i/,/u/,/e/
Syllabic nasal		/N/
Choked sound		/Q/

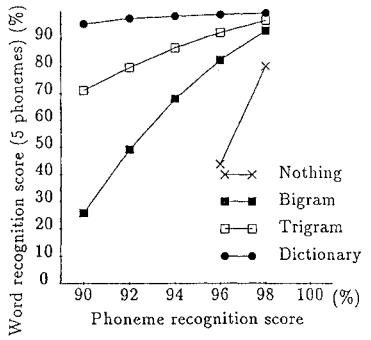

Fig.4 The relationship between phoneme recognition score and lower limit of word recognition score with transition probability

above-mentioned method to evaluate the performance of the recognition system on sentence recognition by extending the relationship between phoneme and word into the relationship between word and sentence. Sentence is represented with 18 kinds of speech such as noun, verb and etc. The bigram and trigram are estimated from 136 sentences. The relationship between word recognition score and lower limit of sentence recognition score is shown in Fig.5. Symbol ■ indicates the score with bigram, and symbol □ indicates the score with trigram. For example, recognition score of sentence composed of 11 words is more than 4.3% using the bigram, on the other hand, more than 67% using the trigram, when the word recognition score is 80%. The efficiency of the trigram is well shown.

VI. CONCLUSION

In this paper, we describe performance evaluation in speech recognition system which use transition probability between linguistic units such as phonemes or words We make clear two kinds of relationship: 1) the relationship between phoneme recognition score and lower limit of word recognition score, 2) the relationship between word recognition score and lower limit of sentence recognition score, when the transition probability is used as linguistic information. The lower limit of word recognition score is obtained based on phoneme recognition score and a mean of number of words with short distance[1].

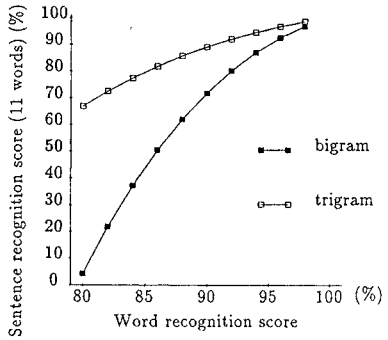

Fig.5 The relationship between word recognition score and lower limit of sentence recognition score with transition probability

But there remains difficulty in calculation of short distance words when the transition probability is used. We propose new algorithm based on dynamic programming which reduces computation amount from $O(lB_s^2)$ of conventional one to $O(l^2)$, where l and B_s are word length and static branching factor of transition probability. By means of this algorithm, we carry out the performance prediction in word recognition using bigram or trigram of phonemes, and sentence recognition when each word in a sentence is represented with 18 kinds of speech such as noun, verb and etc. Recognition score for word with 5 phonemes is more than 26% using the bigram, more than 71% using the trigram and more than 95% using a dictionary when phoneme recognition score is 90%, where the bigram and the trigram of phonemes are estimated from the 5,317 Japanese popular words. Recognition score of sentence composed of 11 words is more than 4.3% using the bigram, on the other hand, more than 67% using the trigram, when the word recognition score is 80%, where bigram and trigram are computed from 136 sentences.

References

[1] K.Abe, K.Hatano and T.Fukumura, "Performance evaluation of character recognition system with dictionary," Trans. IECE Jpn. **J52-C**, 305-312, (1969) (in Japanese)

[2] S.Makino and K.Kido, "Properties of phoneme pairs for distinguishing a word pair with a short distance," Trans. IECE Jpn. **J62-D**, 507-514, (1979) (in Japanese)

[3] S.Nakagawa, "Relationship between phoneme recognition accuracy and word recognition accuracy," Trans. Inf. Proc. Soc. Jpn. **22**, 5, 488-496, (1981) (in Japanese)

[4] S.Nakagawa, Y.Ooguro and I.Murase, "An evaluation method for continuous speech recognition systems - Relationship between task complexity and sentence recognition accuracy -," Trans. IEICE Jpn. **J73-D**, 683-693, (1990) (in Japanese)