LIKELIHOOD NORMALIZATION USING AN ERGODIC HMM FOR CONTINUOUS SPEECH RECOGNITION

Kazuhiko Ozeki

The University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo, 182 Japan
ozeki@cs.uec.ac.jp

ABSTRACT

In recent speech recognition technology, the score of a hypothesis is often defined on the basis of HMM likelihood. As is well known, however, direct use of the likelihood as a scoring function causes difficult problems especially when the length of a speech segment varies depending on the hypothesis as in word-spotting, and some kind of *normalization* is indispensable. In this paper, a new method of likelihood normalization using an ergodic HMM is presented, and its performance is compared with those of conventional ones. The comparison is made from three points of view: recognition rate, word-end detection power, and the mean hypothesis length. It is concluded that the proposed method gives the best overall performance.

1. INTRODUCTION

Evaluation of correctness of a hypothesis, which is called *scoring*, is a crucial issue in speech recognition. In many situations such as picking up the most likely word hypothesis, detecting out-of-vocabulary words, and rejecting misrecognitions, the performance depends heavily upon the way the score is defined. In recent speech technology, the score of a hypothesis is often defined on the basis of HMM likelihood. However, as many people are now aware, direct use of the likelihood as a scoring function causes difficult problems especially when the length of a speech segment varies depending on the hypothesis as in word-spotting.

According to statistical decision theory, given an observation O, we should find a hypothesis W that maximizes the *a posteriori* probability $P(W \mid O)$[1]. By Bayes' theorem, $P(W \mid O) = \{P(O \mid W)/P(O)\}P(W)$. Thus, when the observation O is fixed as in isolated word recognition, maximization of $P(O \mid W)$ is equivalent to maximization of $P(W \mid O)$: a constant $P(O)$ can be neglected. In continuous speech recognition, however, we are often faced with the problem of comparing two hypotheses such as "O_1 is the speech segment corresponding to a hypothesis W_1" and "O_2 is the speech segment corresoponding to another hypothesis W_2". In such a case, there is a possibility of obtaining a better scoring method by taking $P(O)$ into account.

The aim of this paper is to show that the likelihood *normalized* by $P(O)$, or equivalently the mutual information [2] $I(O\,;W) = \log P(O \mid W) - \log P(O)$ has in fact more desirable properties as a scoring function than $\log P(O \mid W)$ through various experiments [3]. Calculation of $I(O\,;W)$ requires the value of $\log P(O)$. For that purpose an ergodic HMM is exploited: the log-likelihood of a hypothesis is normalized by subtracting the log-probability of the speech segment estimated with an ergodic HMM.

There is another well known normalization method which is somewhat similar to the present one, though the original idea is different [4],[5],[6]. In the method, the likelihood of a hypothesis is normalized by a speech probability calculated with an all-phone model or an all-syllable model. It is reported that this method is effective in detecting misrecognitions and out-of-vocabulary words. Comparison of this normalization method with the proposed one is also carried out.

2. HYPOTHESIS GENERATION AND SCORING

2.1. Baseline Method

In order to compare various normalization methods on continuous speech, some kind of hypothesis generation technique is necessary. To that end, an HMM version of a connected word recognition algorithm[7] was employed. Let $W = \{W_1, \ldots, W_M\}$ be a set of vocabulary words, and $O = O_1 \cdots O_t \cdots O_T$ an acoustic observation, O_t being the observation vector at the t-th frame. A vocabulary word is modeled with a concatenation of phone HMMs. For each frame t, each word W_m, and each state i of W_m, the algorithm computes the accumulated log-likelihood $L(t, W_m, i)$ by Viterbi search. It also yields the word $W_{m'}$ that immediately precedes W_m and the last frame t' of $W_{m'}$ by back tracing. Thus for each frame t and each word W_m, a hypothesis "$O_{t'+1} \cdots O_t$ is the speech segment corresponding to W_m" is generated. By using the accumulated log-likelihood, a score of the hypothesis can be defined as

$$S_b(t, W_m) = L(t, W_m, s_m) - L(t', W_{m'}, s_{m'}),$$

where s_m and $s_{m'}$ are the last states of W_m and $W_{m'}$, respectively. The score $S_b(t, W_m)$ is considered to be an approximation to $\log P(O_{t'+1} \cdots O_t \mid W_m)$. This scoring method,

without normalization, is referred to as the *baseline method*.

2.2. All-Phone Method

Let $V = \{V_1, \ldots, V_N\}$ be the set of phones. The connected word recognition algorithm can also be used for connected phone recognition just by replacing the set of words with the set of phones: for each frame t, each phone V_n, and each state j of V_n, the algorithm gives the accumulated log-likelihood $L_a(t, V_n, j)$. By combining $L(t, W_m, i)$ and $L_a(t, V_n, j)$, another score can be defined for W_m terminating at frame t:

$$\begin{aligned} S_a(t, W_m) &= \{L(t, W_m, s_m) - L(t', W_{m'}, s_{m'})\} \\ &\quad - \{\max_n L_a(t, V_n, e_n) - \max_n L_a(t', V_n, e_n)\}, \end{aligned}$$

where the first term on the right-hand side is identical with $S_b(t, W_m)$, and e_n is the last state of V_n. The second term could be regarded as an approximation to $\log P(O_{t'+1} \cdots O_t)$. This scoring method is referred to as the *all-phone method*.

2.3. Ergodic Method

For each frame t and each state j of the ergodic HMM, the accumulated log-probability $L_e(t, j)$ is calculated by Viterbi search. By combining the accumulated log-likelihood $L(t, W_m, i)$ and $L_e(t, j)$, yet another score for W_m terminating at frame t is defined:

$$\begin{aligned} S_e(t, W_m) &= \{L(t, W_m, s_m) - L(t', W_{m'}, s_{m'})\} \\ &\quad - \{\max_j L_e(t, j) - \max_j L_e(t', j)\}, \end{aligned}$$

where the first term on the right-hand side is identical with $S_b(t, W_m)$. It is expected that the second term

$$\{\max_j L_e(t, j) - \max_j L_e(t', j)\}$$

is a close approximation to $\log P(O_{t'+1} \cdots O_t)$. This scoring method is called the *ergodic method*.

3. DATABASE AND HMM

ATR English speech database EM01 was used for the experiment. It contains 200 sentences read by a single male speaker of British English. From these 200 sentences, 20 were reserved for testing, and the remaining 180 were used for training phone HMMs. There are 43 distinct phones and 167 words in the 20 test sentences. Some words have pronunciation variations. If they are counted as different words, then the number of distinct words is 173. Also, some words have the same phonetic transcription in the label. For example, *a*, *or*, and a variation of *the* are all labeled as /@/. The number of tokens differs greatly from phone to phone. The minimum is 23 for /zh/, and the maximum is 867 for /@/. In training phone HMMs, the number of tokens were limited to 200. The standard LPC analysis was performed to obtain 14 mel-cepstral coefficients, 14 Δmel-cepstral coefficients, and the Δlog-power.

Each phone HMM is of left-to-right type with 4 states. Each state, except for the last one, has 4 Gaussian mixture components with diagonal covariance matrices. The ergodic HMM is also of Gaussian mixture density type with diagonal covariance matrix. The number of states and the number of mixture components of the ergodic HMM were decided to be 3 and 4, respectively, by a preliminary experiment. For training the ergodic HMM, 10 sentences were picked up from the 180 training sentences, and fed to the Baum-Welch algorithm as multiple data.

4. EXPERIMENTAL RESULTS

4.1. Recognition Rate

At each frame of a sentence, a set of word hypotheses are generated terminating at the frame, and their scores are calculated by each of the above methods. There are 223 true word-endings in the 20 test sentences, which are designated in the label.

If a scoring method is reasonable, a correct hypothesis should get a high score at the true word-endings. In order to see which of the three scoring methods is most reasonable from this point of view, we counted the number of true word-endings at which the top hypothesis was the correct one. The ratio of this number to the total number of true word-endings is referred to as *recognition rate*. We also counted the numbers of true word-endings where the correct word was contained within the top 5 hypotheses, and within the top 10 hypotheses, respectively. The results for the three methods are shown in Table 1. Although the top rank performance of the all-phone method was worse than that of the baseline method, its 5th rank and 10th rank performances were better than those of the baseline method. The ergodic method attained a significant performance gain over the baseline method and the all-phone method.

Table 1. Recognition rates (%).

	Baseline	All-phone	Ergodic
Top	30.0	28.7	57.8
5th	53.8	65.5	80.7
10th	61.4	77.1	91.9

4.2. Behavior of Maximum Scores

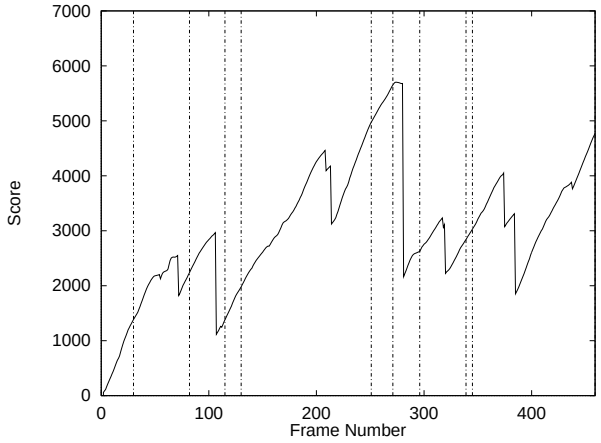
Let us consider a function of t defined as

$$F_b(t) = \max_m S_b(t, W_m).$$

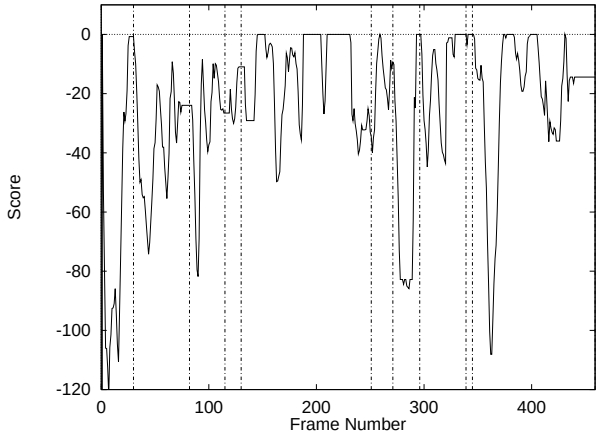
Similarly we define

$$F_a(t) = \max_m S_a(t, W_m), \quad F_e(t) = \max_m S_e(t, W_m).$$

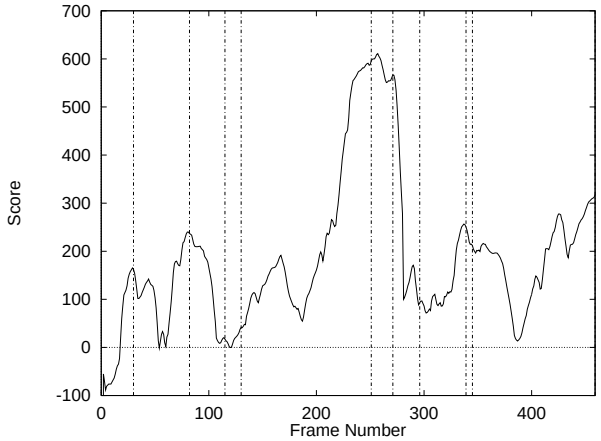
These functions should have a local peak at a true word-ending because there must be a correct hypothesis terminating there. Fig.1 (i), (ii), and (iii) are graphs of $F_b(t)$, $F_a(t)$, and $F_e(t)$, respectively for a sentence: "*She flicks through a magazine when she gets a chance.*"



(i) Baseline method: $F_b(t)$.



(ii) All-phone method: $F_a(t)$.



(iii) Ergodic method: $F_e(t)$.

Fig.1. Maximum scores by each of the three methods.

The vertical broken lines indicate the positions of true word-endings. As can be seen in these graphs, $F_b(t)$ keeps on growing even after t passes through a true word-ending, and no peak can be observed at a true word-ending. $F_a(t)$ varies wildly, and it is difficult to tell if a true word-ending coincides with a local peak or not. In contrast to these cases, $F_e(t)$ shows a local peak at most of the true word-endings. The respective behaviors of these functions are similar on other sentences as well. This is another evidence that the ergodic method is superior to the other two.

4.3. The Mean Hypothesis Length

Table 2 shows the mean length of word hypotheses which had chances to be the top in the example sentence. The length was measured by the number of phones. The table also shows the mean length of the words which actually appear in the sentence. From this table it is concluded that long word hypotheses get abnormally high scores in the baseline method, whereas short word hypotheses are given excessive scores in the all-phone method. In the ergodic methods, the mean length of hypothesized words is close to that of true words. This suggests that the ergodic method gives reasonable scores to word hypotheses regardless of their lengths.

Table 2. The mean length of hypotheses and true words in the example sentence.

Baseline	All-phone	Ergodic	True
5.8	1.8	3.3	3.1

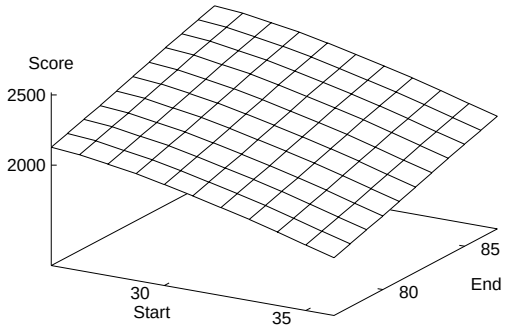
4.4. The Score for a Speech Segment with Given Starting and Ending Frames

In the above experiments, the starting frame of a hypothesis was decided by the connencted word recognition algorithm automatically. Another experiment was conducted in which the scores were calculated by Veterbi search as functions of specified starting and ending frames without using the connected word recognition algorithm. Fig.2(i)-(iii) show scores of a hypothesized word *flicks* on the example sentence for various starting and ending frames. The true starting and ending frames for the word *flicks* are 31 and 82, respectively.

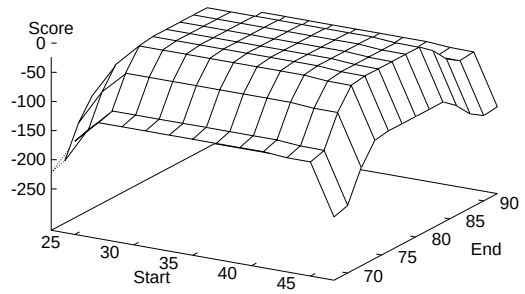
Fig.2(i) shows the result for the unnormalized case. The score changes almost linearly with respect to the length of a hypothesis regardless of its position; it is very insensitive to a shift of the hypothesis against the speech, and therefore has little power to detect the true starting and ending frames. Fig.2(ii) is the result for the case in which the log-likelihood was normalized by using the all-phone model. It has a wide plateau near the true starting and ending co-ordinate (31, 82). Fig.2(iii) illustrates the log-likelihood normalized by using the ergodic HMM. It has a local peak at (29, 81), which is very close to the true starting and ending co-ordinate.

The role of the log-probability estimated with an ergodic HMM can be explained in the following way. The log-likelihood of a hypothesis contains a component that grows almost linearly at a certain rate as the number of frames increases regardless of whether the hypothesized word fits the speech segment or not. When it fits, the log-likelihood grows faster than the average rate, and otherwise it grows less fast. The linearly growing component can be estimated by using an ergodic HMM. By subtracting the component from the log-likelihood, a reasonable measure of goodness of

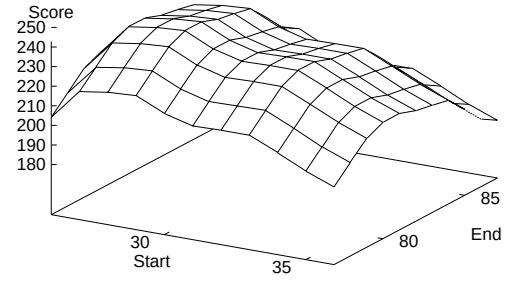
fit between a hypothesized word and a speech segment can be obtained.



(i) Baseline method.



(ii) All-phone method.



(iii) Ergodic method.

Fig.2. Scores of *flicks* in the example sentence calculated for various starting and ending frames.

5. CONCLUSION

Three scoring methods have been compared from various points view. Overall, the ergodic method shows the best performance. In the baseline method, long word hypotheses get unduly high scores, whereas in the all-phone method short ones are favored too much. In the ergodic method, long word hypotheses as well as short ones seem to be given reasonable scores.

In the all-phone method and the ergodic method, the scores are considered to be approximations to the mutual information. The results here lead to a conclusion that the mutual information is a better scoring function than the log-likelihood, and that the log-probability of speech estimated with an ergodic HMM gives a better approximation to the mutual information than that estimated with an all-phone model.

A wide range of potential applications of the ergodic method are conceivable: word-spotting, rejection of out-of-vocabulary words, rejection of misrecognitions, enhancement of robustness in speech recognition, and etc. where the score plays the decisive role. Extension of the ergodic method to speaker independent case is another important future work.

ACKNOWLEDGEMENT

A preliminary part of the work presented in this paper was done while the author was staying at CSTR, the University of Edinburgh, during July and August 1995 as a visiting researcher. The author wishes to thank Dr. Stephen Isard for helpful discussions. He would also like to thank Dr. Paul Taylor for his technical support.

REFERENCES

[1] X. D. Huang, Y. Ariki, and M. A. Jack, "Hidden Markov Models for Speech Recognition", Edinburgh University Press, p.17, 1990.

[2] R. G. Gallager, "Information Theory and Reliable Communication", John Wiley & Sons, Inc., p.16, 1986.

[3] K. Ozeki, "The mutual information as a scoring function for speech recognition", Technical Report of IEICE, SP95-101, 1995.

[4] T. Watanabe and S. Tsukada, "Unknown utterance rejection using likelihood normalization based on syllable recognition", Trans. IEICE, Vol.75-D-II, No.12, pp.2002-2009, 1992.

[5] S. R. Young, "Detecting misrecognitions and out-of-vocabulary words", Proc. ICASSP, II-21-24, 1994.

[6] S. Hayamizu, K. Itou, and K. Tanaka, "Detection of unknown words in large vacabulary speech recognition", J. Acoust. Soc. Japan (E) Vol.16, No.3, pp.165-171, 1995.

[7] H. Ney, "The use of one-stage dynamic programming algorithm for connected word recognition", IEEE Trans. Vol.ASSP-32, No.2, 1984.