A Recognition Time Reduction Algorithm for Large-Vocabulary Speech Recognition

대용량 음성인식을 위한 인식기간 감축 알고리즘

(Jun Mo Koo, Chong Kwan Un)

구 준 모*, 은 종 관*

ABSTRACT

We propose an efficient pre-classification algorithm extracting candidate words to reduce the recognition time in a large-vocabulary recognition system and also propose the use of spectral and temporal smoothing of the observation probability to improve its classification performance. The proposed algorithm computes the coarse likelihood score for each word in a lexicon using the observation probabilities of speech spectra and duration information of recognition units. With the proposed approach we could reduce the computational amount by 74% with slight degradation of recognition accuracy in 1160-word recognition system based on the phoneme-level HMM. Also, we observed that the proposed coarse likelihood score computation algorithm is a good estimator of the likelihood score computed by the Viterbi algorithm.

요 약

본 논문에서는 대용량 음성인식 시스템의 인식시간을 감축하기 위하여 후보단어를 선정하는 효과적인 방법을 제안하고 이 방법의 성능을 향상시키기 위하여 spectral smoothing과 temporal smoothing을 사용하는 것에 관하여 연구하였다. 제안된 방법은 사전내의 각 단어에 대하여 음성인식 단위의 음성 spectrum관찰확률과 길이정보를 이용하여 대강의 관찰확률을 계산하여 후보단어를 선정한다. 제안된 방법을 음소단위의 HMM을 이용하는 1160단어 인식 시스템에 적용한 결과, 전체 계산량의 74% 가량을 감축할 수 있었으며 이때 인식율의 감소는 매우 작았다. 또한 제안된 대강의 likelihood점수 계산방법은 Viterbi방법에 의하여 계산되는 likelihood 점수를 잘 추정함을 알 수 있었다.

INTRODUCTION

As the speech recognition technology progresses

*Communications Research Laboratory
Department of Electrical Engineering
Korea Advanced Institute of Science and Technology

many algorithms have been proposed for the recognition of large vocabulary. Among these algorithms, the one based on hidden Markov modeling (HMM) which requires much less recognition time than other template matching based approaches is known to be a viable one for the practical usage. Especially, the HMM-based approaches using subword unit models are widely used due to its ability of easy construction of word models from the sub-word models. Although the HMM based algorithms have an advantage in recognition time, it becomes difficult to recognize utterances in real time as the size of vocabulary grows. To alleviate this problem, several algorithms have been proposed [1][2]. One example is the two pass algorithm in which the first-stage classification is performed to select candidate words for the second-stage classification[3]. In this paper, a pre-classification algorithm based on the detection probability and duration information of the recognition units are proposed to reduce total recognition time. The computational gain is considered when the proposed algorithm is used as the first stage of a large vocabulary recognition system utilizing the phoneme level HMM. Also its classification performance is examined and compared to that of the first stage classifier based on the vowel classification.

DESCRIPTION OF ALGORITHM

The pre-classification algorithm consists of two parts : training and classification phases. In the training phase, the probability distribution of speech spectra and the durational information are examined for each recognition unit. In this work, the probability distribution of speech spectra is estimated by a nonparametric method to reduce computation. For this estimation, input speech spectra are vector quantized(VQ) and the relative frequencies of VQ codewords in recognition units are computed. Assuming that the number of the recognition units is R and the number of VQ codewords is M, respectively, than the j-th VQ codeword observation probability of i-th recognition unit, $f_i(j)$, is obtained for $1 \leqq i \leqq R$, $1 \leqq j \leqq M$. Also, the minimum and maximum durations of each recognition units, $d_{\min}(i)$ and $d_{\max}(i)$ are obtained for $1 \leqq i \leqq R$ during the training phase. Among many sub-word units, we adopted context-independent phoneme-like units as recognition units because a word model can be easily constructed by concatenating them according to the phoneme-like unit sequence given in a lexicon.

In the classification phase, the coarse likelihood score is computed for every word $w_k (1 \leqq k \leqq V)$ in a lexicon of size V, and a part of them are chosen as candidate words for the second-stage recognition according to their likelihood scores. When an input speech is uttered, the feature of the input speech is extracted and vector quantized by the VQ codebook. Then, the input speech is represented by a series of VQ indexes $O=\{o_1, o_2, \cdots, o_T\}$, where T is the number of feature frames in the input speech. From this series of indexes, the detection probability of every recognition unit, $P_i=\{p_i(1), p_i(2), \cdots, p_i(T)\}$, is obtained for $1 \leqq i \leqq R$ where $p_i(t) = f_i(o_t)$. From these values, the coarse likelihood score L_k for k-th word w_k which is composed of n recognition units $\{r_1, r_2, \cdots, r_n\}$ is computed as follows. First, a possible starting point s_i and an ending point e_i of the recognition unit $r_i (1 \leqq i \leqq n)$ are computed as

$$s_i = min\left(\sum_{j=1}^{i-1} d_{\min}(r_j), T\right), \tag{1}$$

$$e_i = min\left(\sum_{j=1}^{i} d_{\max}(r_j), T\right), \tag{2}$$

where $min(x, y)$ is equal to x if $x \leqq y$, or y otherwise. The coarse likelihood score L_k for word w_k is computed as

$$L_k = \prod_{t=1}^{T} \max_{r_i, s_i \le t \le e_i} (p_{r_i}(t)), \quad (3)$$

where $\max_{c}(x)$ choose the maximum value of x under the condition c. According to these values $L_k (1 \le k \le V)$, a part of vocabulary are chosen as candidate words and the second-stage classification is performed for those words. An example of the above procedure is depicted in Fig. 1, where a

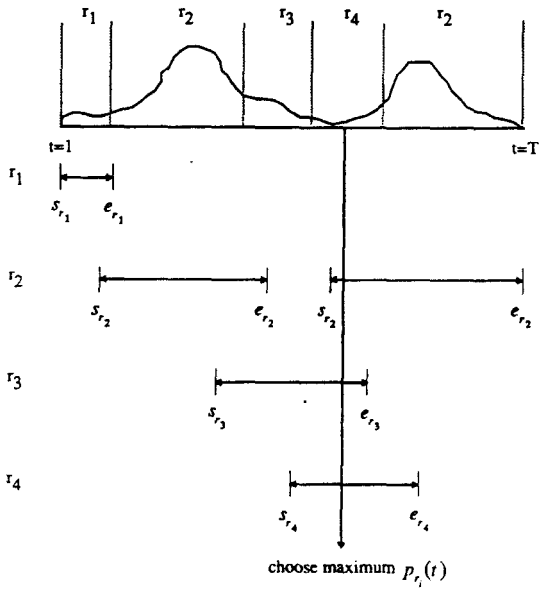

Fig 1. Description of coarse likelihood score computation procedure

word consists of four recognition units.

As we explained above, the time reduction algorithm requires the vector quantization process. If the proposed algorithm is used as a preprocessor of HMM based speech recognition system which requires vector quantization, the computation can be reduced further by sharing the VQ process. Like the recognition system based on the discrete HMM, the performance of the time reduction algorithm largely depends on the parameter estimation procedure of training phase. That is, if the observation probability of speech spectra is estimated from a small amount of training data, the performance of the pre-classification become degraded significantly. To alleviate this problem in the training phase, we adopt a spectral smoothing method which smoothes the probability distribution by the fuzzy mapping concept[4]. If a VQ codeword, o_l which shows a high observation probability for a recognition unit r_i is observed, the codewords observed near the observation time t will also show a high observation probability for the same unit. To accommodate this tendency in the classification phase, we propose a temporal smoothing method which smoothes the recognition unit observation probability $p_i(t)$ as

$$p_i(t) = (c_1 \cdot p_i(t-1) + c_2 \cdot p_i(t) + c_3 \cdot p_i(t+1)) / (c_1 + c_2 + c_3), \quad (4)$$

where c_1, c_2, c_3 are constants. The temporal smoothing compensates for the fact that each output codeword is treated independently in (3). In this work, the temporal smoothing method is applied after fuzzy smoothing to improve the classification performance further, and we observed their contributions to the improvement of classification performance by computer simulation.

COMPUTATIONAL GAIN

We compared the amount of computation of the recognition system based on the phoneme-level HMM with the proposed time reduction algorithm to that of the recognition system without the proposed algorithm. Since the feature extraction and vector quantization procedure can be shared by each other, and the time consumed is negligible to that of the classification procedure in a large vocabulary recognition system, we will consider only the time required in the classification. We used a three-state phonemethel HMM with eight

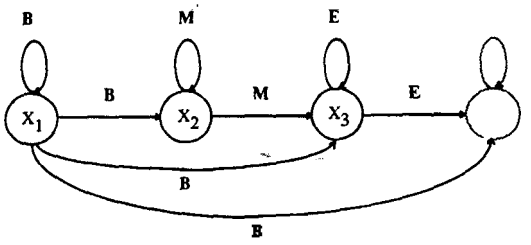

Fig 2. HMM for a phoneme-like unit(the same capital letters represent the same output symbol observation probability vectors)

transitions as shown in Fig. 2. Consider a word which consists of n phonemes and whose length is T. To compute the Viterbi score, $[8(T-n)-1$ $2]\cdot n$ multiplications, the same number of additions and $3(T-n-1)\cdot n$ comparisons are required. For simplicity, if we assume that the time for all operations are equal to time τ and every word in the lexicon has the same number of phonemes and the same length, then the time required to recognize a word by full search of the lexicon of size V is $(19nT-19n^2-27n)\cdot\tau\cdot V$. To choose the candidate words by the proposed time reduction algorithm, we should compute $\mathbf{P}_i(1\leqq i\leqq R)$ once and L_k for every word $w_k(1\leqq k\leqq V)$. Since $V\rangle\rangle R$, the computation required for calculating $\mathbf{P}_i$ can be neglected, and the number of operations required to compute L_k are less than $n\cdot T$ comparisons and T multiplications. But, for simplicity, let the computation time required to perform the first stage classification be $(n+1)\cdot T\cdot\tau\cdot V$. Then the ratio of the computation time of the full search method to that of the first stage classification is

$$G=19\frac{n}{n+1}-19\frac{n^2}{(n+1)T}-27\frac{n}{(n+1)T}. \qquad (5)$$

We applied the proposed algorithm to a recognition system with $V=1160$, $R=44$. The average length of input words was 84.4 frames and the average number of phonemes in a word was 9. 32. Substituting these values in (5), we get the computational overhead $G^{-1}=6\%$ approximately. That is, the time for the first-stage classification is only 6% of the time required for performing the Viterbi score computation for every word in a lexicon. If we choose v words as candidate words from the lexicon by the pre-classification algorithm, the total time required to recognize an input speech is approximately equal to $(v\cdot 100/V+6)\%$ of the time required when the first-stage classification is removed. Through our computer simulation, we confirmed that the real amount of recognition time reduction was almost the same as our estimation.

SIMULATION RESULTS

The performance of the proposed time reduction algorithm was tested in a speaker-independent isolated word recognition system. The recognition unit was 44 Korean context-independent phoneme-level HMM and the size of the lexicon was 1160 which are used for an automatic telephone number information query system. The recognition units used in this work are based on the definition of Korean phonemes except some consonants and vowels. For training, we used a speech data base consisting of 75 Korean phonetically balanced words uttered by 5 male speakers. The test data base of size 1160 words was constructed from the utterances of another male speaker. All input utterances were low-pass filtered with cut-off frequency of 4.5 KHz and digitized with the sampling frequency of 10KHz. End points were detected manually, and phonemes were hand-segmented for the training data. But, the end points were detected automatically in the test procedure. Twelve-order LPC cepstral coefficients and differenced LPC cepstral coefficients were obtained as the features in every 10ms. We used two separate codebooks of size 256 for each feature, and

treated them independently[5]. The Viterbi algorithm was used to compute the likelihood of a model and a test utterance in the second stage.

We compared the classification performances of the first-stage classification algorithms of different smoothing methods. Table 1 shows their inclusion rate according to its candidate word selection rang where the selection rang is the ratio of the number of candidute words for the second stage classification to the number of vocabulary.

Table 1. Average inclusion rate(%) for different selection ranges and average rank of spoken words in a candidate word list.

Algorithm	Selection range						avg.
	0.1	0.2	0.3	0.4	0.5	0.6	rank
Basic	83.3	92.1	95.3	97.3	98.4	99.2	67.3
Spectral	89.7	95.8	97.9	98.3	99.2	99.7	44.6
Temporal	90.6	95.9	97.9	98.3	98.7	99.7	40.5

One can see from Table 1 that the effect the fuzzy smoothing and temporal smoothing improves the classification performance substantially, and that if we select 20% of the vocabulary as candidate words, about 4% of input utterances are not included in the candidate word list. In this case, we can save approximately 74% of computation as compared to the system without the pre-classification stage. We also observed that the everage rank of an input speech in the candidate word list was very small, that is about 4% of the vocabulary size. The recognition accuracy of the large vocabulary recognition system according to the different selection range are shown in Table 2.

Selection range	0.1	0.2	0.3	1.0
Recognition accuracy	(71.81)	(72.59)	(72.84)	(72.67)

Table 2. Recognition accuracy(%) of the large vocabluary recognition system for different selection ranges.

Since the recognition accuracy is not different from that of the full search case when the selection range is greater than 0.3, only three selection ranges were considered. Even when we set the selection range to be 0.1, the recognition accuracy degrades very slightly. From this result, we can assert that the proposed coarse likelihod score computation coincides with the likelihood score computed by the Viterbi algorithm very well. Note that the perplexity of the recognition system is 1160 and the mode of the recognition is speaker-independent.

We also compared the classification performance to that of the first stage classification algorithm based on the classification of vowel class[6]. The algorithm detects the position of vowels and recognizes the vowel class by their formants. According to the sequence of vowel classes, they choose the candidate words for the second stage classification. The algorithm choose 20% of the vocabulary as the candidate words with the inclusion rate of 97% in a speaker dependent mode. Although the classification performance is comparable to the proposed algorithm, the total computation time required is much longer than the proposed algorithm, since the vowel classification procedure should extract additional features like formants.

REFERENCES

1. L. Bahl et al. "Matrix fast match : A fast method for identifying a short list of candidate words for decoding", in *Proc of Int. Conf Acoust., Speech, Signal Processing*. Paper S6.24, 1989.
2. L. Fissore, G. Micca and R.Pieraccini, "Very large vocabulary isolated utterance recognition : A comparison between one pass and two pass strategies", in *Proc of Int Conf Acoust Speech, Signal Processing*" Paper S5.6, 1988.
3. T. Kaneko and N. R. Dixon, "A hierarchical decision approach to large vocabulary discrete utterance recognition", *IEEE Trans. Acoust., Speech, Signal processing*. Vol. ASSP-31, pp.1061-1066, Oct.1983.
4. J. M. Koo and C. K. Un, "Fuzzy smoothing of HMM

parameters in speech recognition", *Electron. Lett.*, Vol. 26, No.11, pp.743-744, May 1990.

5. V. N. Gupta, M. Lennig and P. Mermellstein, "Integration of acoustic information in a large vocabulary recognizer", in *Proc. of Int. Conf. Acoust., Speech, Signal Processing*, pp.697-700, Apr., 1987.
6. Y. J. Chung and C. K. Un, "A time reduction algorithm using vowel classification for large vocabulary speech recognition", in *Proc. of the Seoul Int. Conf. on Natural Language Processing*, pp.182-185, 1990.

Koo Jun MO 正會員

Education : 85.2 B.S, Electronic Engineering, Seoul National University

87.2 M.S, Electrical and Electronics Engineering, Korea Advanced Institute of Science and Technology

91.8 Ph. D, Elctrical and Electronics Engineering, Korea Advanced Institute of Science and Technology

Experience : 87.7~Research engineer, DigiCom Institute of Telematics

Fields of Interest : Speech recognition and synthesis, Digital signal processing, Application of fuzzy theory, Neural networks

▲ **Dr. Chong Kwan Un** received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Delaware, Newyork, Delaware, in 1964, 1968 and 1969, respectively. From 1969 to 1973 he was Assistant Professor of Electrical Engineering at the University of Maine, Portland, where he taugth communications and did research on synchronization problems. In May 1973, he joined the Staff of the Telecommunication Sciences Center, SRI International, Menlo Park, CA, where he did research on voice digitization and bandwidth compression systems. Since June 1977 he has been with KAIST, where he is Professor of Electrical Engineering and Head of the Communications Research Laboratory, teaching and doing research in the areas of digital communications and digital signal processing. He has authored or coauthored over 200 papers and 70 reports in speech coding and processing, packet voice / data transmission, synchronization, and digital filtering. Also, he holds 7 patents granted or pending. From February 1982 to June 1983 he served as Dean of Engineering at KAIST.

Dr. Un has received a number of awards including the 1976 Best Paper Award from the IEEE Communications Society, a number of awards including the 1976 Best Paper Award from the IEEE Communications Society, the National Order of Merits (Dong Baik Jang) from the Government of Korea, and Achievement Awards from KITE, KICS and ASK. He is a Fellow of IEEE and a member of Tau Beta Pi and Eta Kappa Nu Honor Societies.