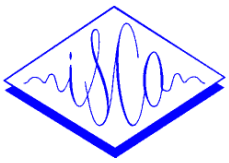

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

An Efficient Output Probability Computation for Continuous HMM using Rough and Detail Models

Yasuhiro KOMORI, Masayuki YAMADA, Hiroki YAMAMOTO and Yasunori OHORA
Media Technology Laboratory, Canon Inc.
890-12 Kashimada, Saiwai-ku, Kawasaki-shi, Kanagawa 211 JAPAN,
Phone:+81-44-549-5111 Fax:+81-44-549-5434 Email:komori@cis.canon.co.jp

Abstract

The paper presents an efficient computation algorithm of the output probability for a continuous HMM speech recognizer using a rough and detail HMM combination. In general, the more number of mixtures or the more number of contextual classes, the better accuracy with the heavier computation, and vice versa. The proposed algorithm first estimates the state output probabilities using the rough HMMs and then re-estimates those of the probable states using detail HMMs. We proposed two realizations for the algorithm and carried out experiments for each. Both results showed about 60% or 70% reduction of the output probability calculation with no reduction of recognition accuracy.

10.21437/Eurospeech.1995-211

1 Introduction

To realize a real-time continuous HMM speech recognizer, reduction of the output probability computation is indispensable. This paper proposes an efficient computation of the output probability for the continuous HMM using rough and detail HMMs.

The computation amount of the continuous HMM output probability depends upon the total number of the probabilistic mixture density functions ($N = M \cdot s \cdot m$) where M is the number of the models, s is the number of states ineach model and m is the number of mixtures in each state. In orderto reduce this computation amount, one way is to reduce the N to be calculated with no performance reduction.

There are some approaches already proposed: In the Bocchieri's approach[1], the reduction of the computation amount is realized by a selection of the probabilistic mixture density functions. The selection is obtained by comparing the input vector with the prequantized vectors which are determined by vector quantization with the mean values of the probabilistic mixture density functions. Then the probabilistic density functions which belong to the selected prequantized vector are calculated.

Watanabe's approach[2] is a sophisticated extension of the Bocchieri's approach. A probabilistic measure is adopted to the probabilistic mixture density function selection. Moreover, a tree-structure is introduced to the selection and the calculation of the multi-mixture probabilistic density functions.

The basic principle of these algorithms are that "clusters which seem to contribute to the recognition result are evaluated in detail, where clusters which do not seem to contribute are roughly evaluated."

In our algorithm, the basic principle is the same, however our algorithm is based on a simple combination of the rough HMMs and the detail HMMs.

In this paper, the proposed algorithm and the experiments on speaker-independent large vocabulary continuous speech are discussed.

2 Proposed Method

2.1 Basic Idea

The basic idea to reduce the amount of the output probability computation is to reduce the number of the probabilistic mixture density functions (hence mixtures) to be calculated by considering the contribution to the recognition result, because the computation amount is determined by the total number of mixtures of the HMMs. We introduce an "HMM state" as the "cluster" to the basic principle. The contribution to the recognition result is estimated by the state output probability of the rough HMM.

2.2 Assumption

Here, we make an assumption to realize the idea. Let's consider two HMMs (a rough HMM and a detail HMM) belonging to the same phoneme class. The structure of these HMMs is a simple left-to-right model with only several states. The assumption is that even if the two HMMs are independently trained, the related states of these two HMMs represent similar acoustic features and their output probabilities are strongly related so that they can be replaced each other. This is because the two HMMs are trained by similar parts of the training data.

2.3 Realization

In general, the more number of mixtures or the more number of contextual classes, the better accuracy with the heavier computation, and vice versa. Based on these facts and by the assumptions described above, the idea can be realized as two basic types for the rough HMM and the detail HMM combination.

1) a combination of small and large number of mixture HMMs
2) a combination of context-independent and dependent HMMs

2.4 Algorithm

TRAINING PROCESS

1) Both rough HMMs (HMM_R) and detail HMMs (HMM_D) are trained independently by the ordinary EM-algorithm using the speech database, the labels and the acoustic parameters. The structure of both HMM_R and HMM_D are exactly the same except the number of mixtures or the contextual labels.

RECOGNITION PROCESS

1) Calculate all $P_j^R(O(t))$, the output probability of all states j in HMM_R, for the input parameter $O(t)$.
2) Select states $\{j\}$ of which the output probabilities are greater than the given threshold by assuming that these states will contribute to the recognition result. The threshold may be a number such as top-N or a value related to the best output probability $P_j^R(O(t))$ in the frame.
3) Find the states $\{k\}$ of the HMM_D related to the states $\{j\}$ of the HMM_R and re-evaluate the output probabilities $P_k^D(O(t))$.
4) Copy the output probabilities of the non-re-evaluated states of the HMM_D from the output probabilities of the related states of the HMM_R, which are determined in the first step ($P_k^D(O(t)) = P_j^R(O(t))$, k is related to j).
5) Perform the Viterbi search using all the determined output probabilities $P_k^D(O(t))$.

The total number of mixtures C to be calculated for each input frame in this algorithm is;

$$C_{proposed} = M_R \cdot s_R \cdot m_R + n \cdot m_D$$

while the straightforward computation is;

$$C = M_D \cdot s_D \cdot m_D$$

where M is the number of the models, s is the number of the states in each model, m is the number of the mixtures in each state, n is the number of re-evaluation states and the suffix R and D indicate the rough HMM and the detail HMM, respectively.

In the combination of small and large number of mixture HMMs, m_D is larger than m_R and in the combination of context-independent and dependent HMMs, M_D is larger than M_R. Thus, we can reduce the total number of mixtures to be calculated, as $C_{proposed} \ll C$, by selecting n considering the number of M_R, m_R, M_D, m_D, because the number of states is equal ($s_R = s_D$) in the proposed method.

Figure 1 shows the image of the algorithm.

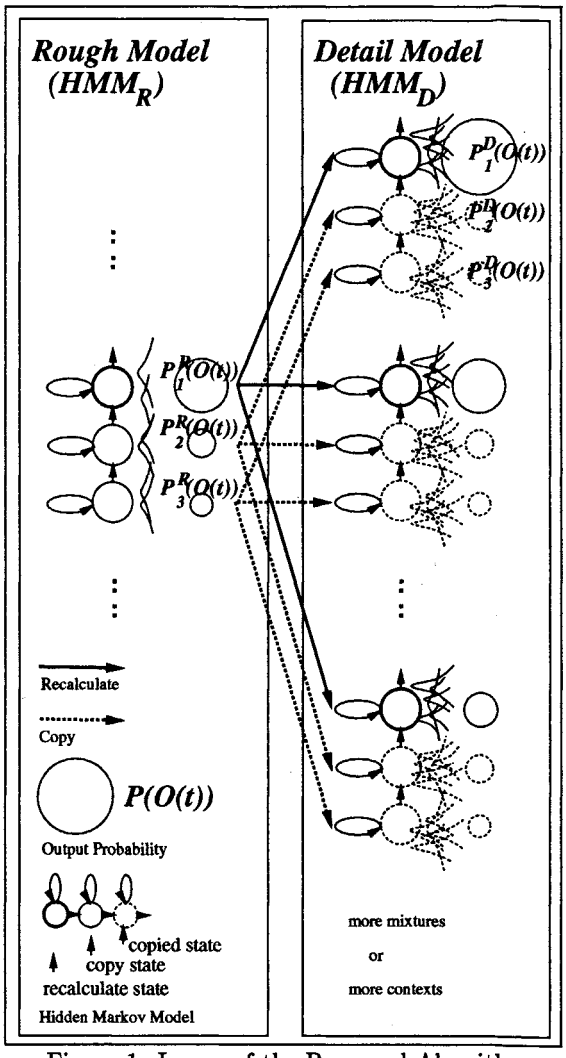

Figure 1: Image of the Proposed Algorithm

2.5 Advantages

The advantages of the proposed algorithm are:

- drastic reduction of computation
- very small computational overhead
- the same performance as the detail HMMs
- frame-by-frame time-synchronous algorithm
- facile integration into the Viterbi search
- ordinary HMM training method

3 Experiment

Two experiments are carried out. The first one, Experiment I, is carried out to evaluate the number of mixture combination of the rough and the detail HMMs. The second one, Experiment II, is carried out to evaluate the threshold types additionally to the mixture combination or context combination of the rough and the detail HMMs.

3.1 Conditions

We adopted the N-best A* Viterbi beam-search algorithm [3] in our continuous speech recognizer. The vocabulary size in the system is 1,004 and its word perplexity is 30.2.

The HMM is a continuous HMM of a diagonal covariance with multi-mixture Gaussian density functions. Two models are adopted, a context independent phone HMM (25 models) and a right context dependent HMM (243 models).

As for training, we used the ASJ and the ATR speech databases [4, 5] of about 100 speakers, about 20,000 utterances. As for testing, we used 500 continuous utterances of 10 speakers [6] fully independent from the training set. The condition of the acoustic analysis is described in Table 1.

Table 1: Experimental Conditions

acoustic analysis	
sampling rate	16kHz
frame shift	10ms
window size	25.6ms
pre-emphasis	0.97
parameter (26 dimension)	LPC-Mel-cepstrum(12), normalized log power, Δ cepstrum(12), Δ power

3.2 Baseline Experiment

As for the baseline, the 25 phone HMMs or the 243 right-context HMMs are adopted. The phone HMMs are 3 state HMMs with 2, 4, 6, 12 and 24 mixtures. The right-context HMMs are 3 state 6 mixture HMMs. Table 2 shows the results using these HMMs.

Table 2: Baseline Results

condition			results		
model type	models, states,mixs	total mixs	sentence(%) top1	sentence(%) top5	word (%)
phone	25 · 3 · 2	150	65.2	88.2	90.9
phone	25 · 3 · 4	300	73.8	91.4	93.7
phone	25 · 3 · 6	450	76.0	92.8	94.3
phone	25 · 3 · 12	900	79.0	94.4	95.3
phone	25 · 3 · 24	1800	80.4	95.8	95.9
right	243 · 3 · 6	4374	84.2	97.2	96.8

3.3 Experiment I

In the first experiment, HMMs of 25 phone models with 3 states 2, 4, 6, 12 mixtures are used for the rough models, and HMMs of 25 phone models with 3 states 24 mixtures are used for the detail models. As to determine the number of states to be recalculated top-N HMM states in a frame is adopted, where N is 5, 10, 15, 20, 25. Table 3 shows the results of the mixture combination recalculation.

Table 3: Results of Mixture Combination — top-N state selection —

condition			results		
models, states,mixs	n, mixs	total mixs	sentence(%) top1	sentence(%) top5	word (%)
25 · 3 · 2	5 · 24	270	66.8	87.0	90.7
25 · 3 · 2	10 · 24	390	71.0	89.4	92.2
25 · 3 · 2	15 · 24	510	70.2	90.4	92.0
25 · 3 · 2	20 · 24	630	71.4	90.6	92.8
25 · 3 · 2	25 · 24	750	73.8	91.2	93.4
25 · 3 · 4	5 · 24	420	75.4	92.8	93.9
25 · 3 · 4	10 · 24	540	77.6	94.0	94.7
25 · 3 · 4	15 · 24	660	77.8	94.0	94.8
25 · 3 · 4	20 · 24	780	78.0	93.8	94.8
25 · 3 · 4	25 · 24	900	78.6	94.8	95.1
25 · 3 · 6	5 · 24	570	76.2	93.0	94.3
25 · 3 · 6	10 · 24	690	81.0	94.2	95.5
25 · 3 · 6	15 · 24	810	80.6	94.0	95.4
25 · 3 · 6	20 · 24	930	80.6	94.0	95.6
25 · 3 · 6	25 · 24	1050	80.8	95.0	95.8
25 · 3 · 12	5 · 24	1020	79.4	95.6	95.8
25 · 3 · 12	10 · 24	1140	80.2	95.8	95.7
25 · 3 · 12	15 · 24	1260	79.8	95.8	95.7
25 · 3 · 12	20 · 24	1380	80.4	95.6	95.8
25 · 3 · 12	25 · 24	1500	81.0	95.6	96.0

n: number of the state recalculation

From this experiment, to achieve over 80.0% sentence accuracy, the proposed algorithm required more than 6 mixtures for the rough models and at least 10 state recalculation with the detail models. The result indicates that a certain precision is required for the rough models and a certain numbers of recalculation should be performed by the detail models.

3.4 Experiment II

In the second experiment, the following combinations of the rough HMMs and the detail HMMs are utilized:

A) the small mixture rough HMMs are 3 state 6 mixture HMMs and the large mixture detail HMMs are 3 state 24 mixture HMMs.

B) the rough HMMs are 25 phone models and the detail HMMs are 243 right-context models. Both HMMs are 3 state 6 mixture HMMs.

Also, two types of state selection thresholds for the recalculation are adopted:

a) selection of top-N HMM states in a frame: N = 5, 10, 15, 20, 25.

b) selection by threshold T ($T = \alpha \cdot P_{max}$, P_{max} is the maximum state output probability in a frame, α = 0.80, 0.75, 0.70, 0.65, 0.60).

Results of II-A with the threshold a) is shown in Table 4 and b) in Table 5.

In the experiment II-A, to achieve over 80.0% sentence accuracy, the proposed algorithm required only about 38.3%, 690 mixtures (= 25 × 3 × 6 + 10 × 24: 3 state 6 mixture HMM for the first estimation and the top 10 state re-estimation using 3 state 24 mixture HMM), while the baseline required 1,800 mixtures (25 phone HMMs × 3 states × 24 mixtures).

Results of II-B with the threshold a) is shown in Table 6 and b) in Table 7.

In the experiment II-B, to achieve over 84.0% sentence accuracy, the proposed algorithm required about 32.1%, 1,404 mixtures on the average (= 25 × 3 × 6 + 159 × 6: 25 phone HMMs for the first estimation and the top 159 state re-estimation on the average using right-context 3 state 6 mixture HMMs), while the baseline required 4,374 mixtures (243 right contextual HMMs × 3 states × 6 mixtures).

4 Conclusion

The paper proposed a new effective algorithm which drastically reduced the output probability computation for the continuous HMM. The algorithm was based on the rough and detail HMM combination. Two realizations were proposed and experiments for each realization were carried out. Both results showed about 60% or 70% reduction of the output probability calculation with no reduction of recognition accuracy.

Acknowledgment

The authors wish to thank Dr. Hideyuki Tamura, Head of the Media Technology Laboratory at Canon Inc., for giving us the opportunity of this study.

References

[1] Bocchieri E.: Vector quantization for the efficient computation of continuous density likehoods, ICASSP'93, II, pp. 692-69, 1993.

[2] Watanabe T., et al.: Speech recognition using tree-structured probability density function, ICSLP'94, S07-10, pp. 223-226, 1993-10.

[3] Soong F. et al.: Tree-trellis based fast search for finding the N best sentence hypotheses in continuous speech recognition, ICASSP'91, pp.705-708, 1991.

[4] Kobayashi T., et al.: ASJ continuous speech corpus for research, JASJ, Vol.48 (No.12), pp.888-893.

[5] Sagisaka Y., et al.: ATR spoken language database, JASJ, Vol.48 (No.12), pp.878-882.

[6] Yamada M., et al.: Information retrieval from CD-ROMs using speech conversation, Proc. ISSD'93 Tokyo, 10-12 Nov., pp.117-120.

Table 4: Mixture Base Recalculation
— top-N state selection —

condition			results		
T	n, mixs	total mixs	sentence(%) top1	sentence(%) top5	word (%)
—	5 · 24	570	76.2	93.0	94.3
—	10 · 24	690	81.0	94.2	95.5
—	15 · 24	810	80.6	94.0	95.4
—	20 · 24	930	80.6	94.0	95.6
—	25 · 24	1050	80.8	95.0	95.8

T: threshold for state selection
n: number of the state recalculation
total mixs = 450(base) + n· mixs

Table 5: Mixture Base Recalculation
— state selection by threshold —

condition			results		
T	n mixs	total mixs	sentence(%) top1	sentence(%) top5	word (%)
0.80	7.9 · 24	638	77.8	93.2	94.9
0.75	11.3 · 24	720	79.4	91.6	95.0
0.70	15.2 · 24	814	79.0	93.2	95.0
0.65	19.5 · 24	917	79.6	93.6	95.3
0.60	23.9 · 24	1024	80.0	95.0	95.6

T: threshold for state selection
n: number of the state recalculation
total mixs = 450(base) + n· mixs

Table 6: Context Base Recalculation
— top-N state selection —

condition			results		
T	n, mixs	total mixs	sentence(%) top1	sentence(%) top5	word (%)
—	125 · 6	1200	80.8	94.0	95.6
—	250 · 6	1950	84.2	96.8	96.8
—	375 · 6	2700	84.8	96.8	97.0
—	500 · 6	3450	84.8	96.6	96.8

T: threshold for state selection
n: number of the state recalculation
total mixs = 450(base) + n· mixs

Table 7: Context Base Recalculation
— state selection by threshold —

condition			results		
T	n, mixs	total mixs	sentence(%) top1	sentence(%) top5	word (%)
0.80	81.4 · 6	938	82.4	95.6	95.9
0.75	117.2 · 6	1153	83.2	96.4	96.3
0.70	159.0 · 6	1404	84.0	96.2	96.5
0.65	205.2 · 6	1681	84.0	96.6	96.5
0.60	252.1 · 6	1956	84.4	96.6	96.6

T: threshold for state selection
n: number of the state recalculation
total mixs = 450(base) + n· mixs