RESTRUCTURING GAUSSIAN MIXTURE DENSITY FUNCTIONS IN SPEAKER-INDEPENDENT ACOUSTIC MODELS

Atsushi Nakamura

ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-Cho, Soraku-Gun, Kyoto, 619-02 JAPAN

ABSTRACT

In continuous speech recognition featuring hidden Markov model (HMM), word N-gram and time-synchronous beam search, a local modeling mismatch in the HMM will often cause the recognition performance to degrade. To cope with this problem, this paper proposes a method of restructuring Gaussian mixture pdfs in a pre-trained speaker-independent HMM based on speech data. In this method, mixture components are copied and shared among multiple mixture pdfs with the tendency of local errors taken into account. The tendency is given by comparing the pre-trained HMM and speech data which was used in the pre-training. Experimental results prove that the proposed method can effectively restore local modeling mismatches and improve the recognition performance.

1. INTRODUCTION

In continuous speech recognition featuring hidden Markov model (HMM), word N-gram and time-synchronous beam search, a local acoustic modeling mismatch will often cause a likelihood score to fall locally. This may get a correct word sequence pruned away from recognition hypotheses or ranked low among all of the recognition hypotheses. Such a local acoustic modeling mismatch is frequent, especially in the recognition of speech by unrestricted speakers where a wide variety of speakers' individualities are dealt with, and in the recognition of spontaneous speech where spectral features are heavily deformed. It is crucial to overcome such modeling mismatches to achieve accurate recognition of speaker-independent and spontaneous speech.

So far, a few methods based on an operation in likelihood score during search process to avoid wrong pruning have given tentative solutions for this problem [1],[2]. These methods, however, are unable to solve the root problem, that is, some acoustic phenomena are not properly modeled. Acoustic models themselves should be improved based on acoustic phenomena to solve the root problem.

This paper proposes a method of restructuring Gaussian mixture probability density functions (pdfs) in a pre-trained speaker-independent HMM set. In this method, which aims at modeling several acoustic phenomena more properly, the number of components in each mixture pdf is inflated by copying new components from other mixture pdfs with the tendency of local errors taken into account. The tendency is given by comparing the pre-trained HMM set and speech data which was used in the pre-training. As each of the copied components is shared between source and destination mixture pdfs, the total number of Gaussian pdfs in the HMM set does not increase.

In section 2, basic ideas of the proposed method are described. Section 3 gives experimental results of the proposed method and speech recognition using a newly yielded HMM set to show the improvement in recognition performance. In section 4, the effect of restoring the local acoustic modeling mismatch is verified using examples of speech recognition results.

2. BASIC IDEAS

2.1 Overview of Restructuring Procedure

We assume that we can start with an initial HMM set yielded with conventional algorithms.

Gaussian mixture pdfs in the initial HMM set are restructured using speech data. Then, the model parameters in the restructured HMM set are re-estimated to yield an object HMM set. Since we use the same speech data in the restructuring and re-estimation as in the initial HMM set generation, we do not need new speech data. This is one of advantages of the proposed method (Fig. 1).

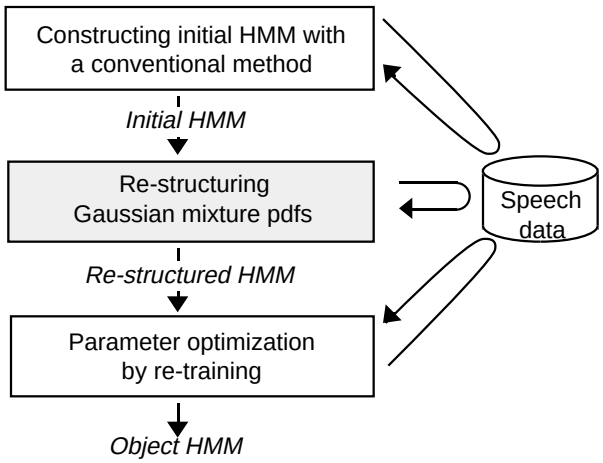

Fig. 1 HMM generation procedure using restructuring

2.2 Copying Mixture Components

The core technique of the proposed method is to inflate the size of each mixture pdf by copying new components from other mixture pdfs with the tendency of frame-level errors taken into account.

Here, we define that frames satisfying the following condition are in the frame-level error [3].

$$g_t \neq \underset{\gamma \in \Gamma}{\operatorname{argmax}} \{ Q(\boldsymbol{o}_t|\gamma) \} \qquad (1)$$

where,

$\boldsymbol{o}_t$:	The observation feature vector at frame t
g_t :	The Gaussian mixture pdf assigned to frame t by the Viterbi alignment
Γ:	A set of Gaussian pdfs in the initial HMM set
$Q(\boldsymbol{o}\|g)$:	The likelihood of a distribution g generating an observation $\boldsymbol{o}$.

In the following discussions, event E will denote the frame level error and E^C will denotes the complimentary event of E.

If frames at which frame-level errors occur frequently due to a Gaussian mixture pdf have similar acoustic features, the Gaussian mixture pdf will tend to fall in acoustic likelihood score when the acoustic features are examined on correct hypotheses in speech recognition.

Now, we consider a Gaussian pdf for a frame t,

$$x_t = \underset{\xi \in \Xi}{\operatorname{argmax}} \{ Q(\boldsymbol{o}_t|\xi) \} . \qquad (2)$$

Although x_t is one of the mixture components in the HMM set, we treat it here as a single Gaussian mixture pdf that gives the maximal likelihood score for the frame t.

$\{x_t\}$ is a series composed by elements ξ in the set Ξ and gives the maximal likelihood score for a feature vector series $\{\boldsymbol{o}_t\}$. We can also consider a series of Gaussian mixture pdfs $\{g_t\}$ which is composed by elements γ in the set Γ and which gives the Viterbi path for $\{\boldsymbol{o}_t\}$. We call these two series the most likely series of Gaussian pdfs and the Viterbi series of Gaussian mixture pdfs, respectively.

Then, a conditional distribution of frame level error occurrences $P(E,\xi\,|\gamma)$ ($\gamma\in\Gamma$, $\xi\in\Xi$) can be given by analyzing the frequency of each element emerging in $\{g_t\}$ and $\{x_t\}$. If the probability $P(E,\xi\,|\gamma)$ is large for a ξ, we can guess that the Gaussian mixture pdf ξ will fall in the acoustic likelihood score when frames whose acoustic features can be well modeled by γ are examined. Note that such falls can be restored by copying γ to ξ as a new component.

2.3 Gaussian Mixture Restructuring Procedure

Mixture components are copied and shared by the following procedure.

(Step-1) Align the initial HMM set and speech data by Viterbi decoding and get the most likely series of Gaussian pdfs $\{x_t\}$ and the Viterbi series of Gaussian mixture pdfs $\{g_t\}$ (Fig. 2).

(Step-2) Calculate the conditional probabilities of frame level error occurrences $P(E,\xi\,|\gamma)$ for all combinations of ξ and γ by analyzing $\{x_t\}$ and $\{g_t\}$. (Fig. 2)

(Step-3) Do Step-4 for all γ.

(Step-4) Do Step-5 for all ξ.

(Step-5) Copy ξ to γ as a new component, if $P(E,\xi\,|g)$ exceeds the pre-determined threshold value. The copied components are shared between source and destination mixture pdfs (Fig. 2, Fig. 3).

The threshold in Step-5 is expected to suppress copy-activation for accidental frame level errors caused by noise or something else in the speech data.

2.4 Parameter Re-estimation

After Gaussian mixture pdfs in the initial HMM set are restructured, the model parameters are re-estimated with a criterion such as maximum likelihood, maximum likelihood ratio, or something else.

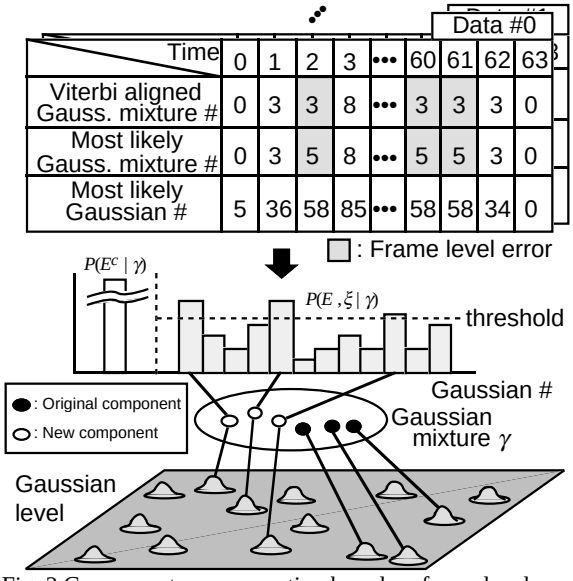

Fig. 2 Component copy operation based on frame level error

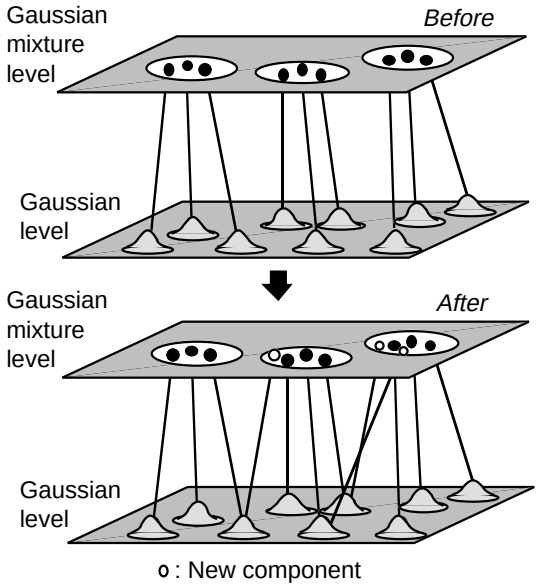

Fig. 3 Component sharing

The values of Gaussian means, Gaussian variances, and state transition probabilities in the initial HMM set are used as initial parameter values for the re-estimation. For mixture weights, initial values are determined as follows.

When Gaussian pdf ξ is copied to Gaussian mixture pdf γ as a new component, the initial mixture weight for the new component is given by,

$$\tilde{w}_{\xi}^{\gamma} = P(E,\xi\,|\,\gamma). \qquad (3)$$

The mixture weights of original components should also be adjusted to keep the summation of the weights equals to 1. They are given by,

$$\tilde{w}_{\xi}^{\gamma} = \left(P(E^c\,|\,\gamma) + P(E,\Xi_{\rho}^{\gamma}\,|\,\gamma)\right) \bullet w_{\xi}^{\gamma} \qquad (4)$$

where,

w_{ξ}^{γ} : The mixture weight for a Gaussian pdf ξ as a component of a Gaussian mixture pdf γ

Here, Ξ_{ρ}^{γ} is a set of Gaussian pdfs which are not copied to γ when the threshold is ρ, and is given by,

$$\Xi_{\rho}^{\gamma} = \{\xi\,|\,P(E,\xi\,|\,\gamma) < \rho\,;\,\xi\in\Xi\}. \qquad (5)$$

3. EXPERIMENTS

3.1 Restructuring Gaussian Mixture Pdfs

Gaussian mixture pdfs in a pre-trained (initial) HMM set were restructured by the method described in section 2.

A set of state-shared context-dependent HMMs (HMnet) was used as the initial HMM set. The HMnet was yielded by ML-SSS algorithm [4] using spontaneous speech of 175 Japanese male speakers from ATR Travel Arrangement Corpus [5],[6]. The conditions employed for the HMnet are summarized in Table 1 and Table 2.

Gaussian mixture pdfs in the HMnet were restructured at the copying threshold of 0.01 using the above mentioned spontaneous speech. The original total number of 4,000 components increased to 6,603 (copied 2,603 times) through this restructuring. In addition, the mixture sizes, which had been 10 for all of the mixture pdfs, became distributed from 10 to 36 (Fig. 4).

Table 1 Conditions of acoustic analysis

Sampling freq. :	12 kHz
Quantization :	16 bit linear
Pre-emphasis :	$1 - 0.97z^{-1}$
Window :	20 ms Hamming
Frame shift :	10 ms
Feature vector :	log-power + 16-order LPC-Cep. + Δlog-power + 16-order ΔCep.

Table 2 Topological conditions of initial HMM set

401 state male-speaker independent HMnet	
400 states for state-shared allophone HMMs (Triphone-context-dependent HMMs)	
1 state for a silence HMM	
Acoustical units :	Japanese 25 phonemes + silence
Mixture size:	10 mixture/state
Covariance type:	Diagonal

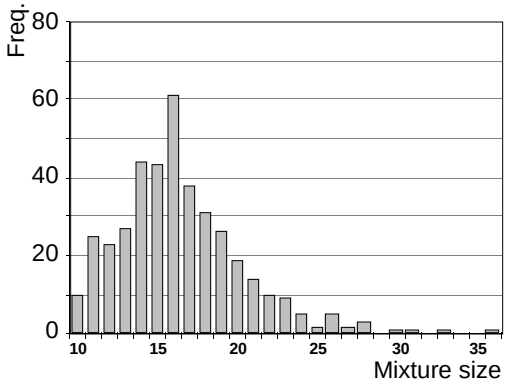

Fig. 4 Frequencies of mixture sizes after restructuring

The relationships between source and destination mixture pdfs were also examined. It was found that more than half of the cases of component copy operations were activated between mixture pdfs representing the same center phoneme. The HMnet used in this work was yielded by splitting states of context-independent HMMs successively in the contextual or temporal domain [4]. We suppose that a component copy operation between two identical center phonemes mainly restores mismatches caused by sub-effects in the process of state splitting.

The most frequent source and destination combinations in the cases of different center phonemes are shown in Fig. 5. We can see that component copy operations were activated frequently between phonemes acoustically similar or tending to be co-articulated with each other.

3.2 Parameter Re-estimation

The model parameters in the restructured HMM set were re-estimated. Although it is possible to enhance the effectiveness of restructuring by choosing an appropriate estimating criterion, the maximum likelihood criterion was employed in this work, as in the initial HMM set generation, in order to evaluate the pure effect given by the restructuring itself.

The following model parameters were re-estimated by the Baum-Welch algorithm using the above mentioned spontaneous speech of 175 male speakers.

- Gaussian means
- Gaussian variances
- Mixture weights
- State transition probabilities

3.3 Continuous Speech Recognition Tests

Continuous speech recognition tests were carried out using the HMM set given by the restructuring and the re-estimation (Restructured HMM set) in order to evaluate how the recognition performance was improved in comparison with the tests using the initial HMM. The experimental conditions were as follows.

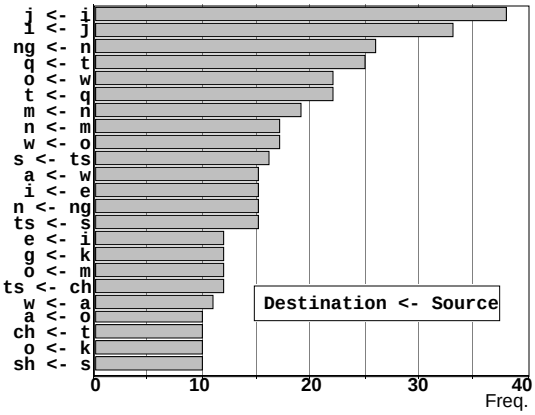

Fig. 5 Frequencies of component copy operations between different center phonemes

- Continuous speech recognizer:
  A recognizer featuring multi-pass search and word-graph outputs [7]
- Language model :
  Variable-length word class N-gram [8], 500 classes in total
- Lexicon :
  6,922 words
- Test data :
  Speaker-open 7 males' spontaneous speech data
  The Travel Arrangement Corpus
  81 utterances, 937 words (accumulative)
- Evaluation :
  The word accuracy and the word %correct for the most likely path in the word graph

The beam width and the language model scale were set experimentally so that the initial HMM set could perform at its best accuracy.

Fig. 6 shows the recognition performance achieved with the language model scale varied around the best setting for the initial HMM set. We can see that the Restructured HMM set given by the proposed method (Restructured) outperformed the initial HMM set (Baseline) in both accuracy and %correct.

4. DISCUSSION

In order to verify that the improvement in the recognition rate described above was certainly given by restoring local acoustic modeling mismatches, temporal behaviors of acoustic likelihood scores in the correct recognition hypotheses were compared between the initial HMM set and the Restructured HMM set.

Since the speech recognizer used in this work employed a beam search based on the difference in likelihood from the most likely hypothesis, the differences were also compared to see how correct hypotheses could possibly be pruned.

We found that falls in likelihood were effectively restored in most of cases of the performance improvement as shown in the following two examples.

(Example 1)

soo de su ka dewa hjakudoru no heja o onegaishimasu

- Recognized with the initial HMM set :
  soo de su ka zhja shita kodomo na ija onegaishimasu
- Recognized with the Restructured HMM set :
  soo de su ka zhja shita kodomo no heja o onegaishimasu

(Example 2)

zhjuugatsu tooka ni juuzhin to iqpaku sa se teitadaki ta i ng de su ga

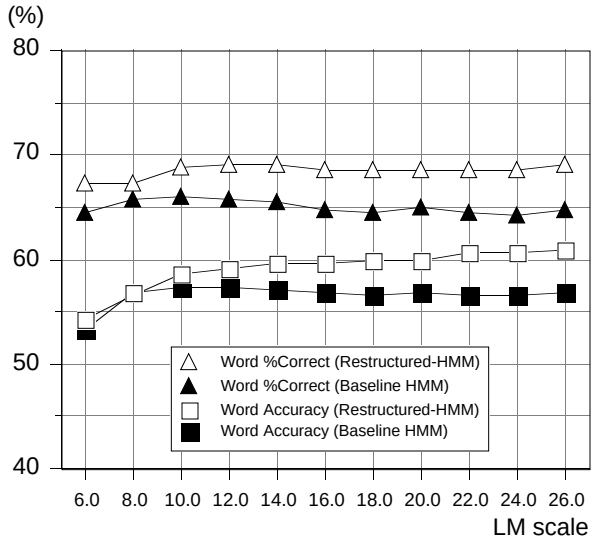

Fig. 6 Comparison in the recognition performance

- Recognized with the initial HMM set :
  zhjuugatsu tooka ni ire te ru to i i ta ku to de teitadaki ta i ng de su ga
- Recognized with the Restructured HMM set :
  zhjuugatsu tooka ni ire te ru to iqpaku sa se teitadaki ta i ng de su ga

For Example 1, the temporal behavior of the acoustic likelihood score near "*no heja o*", which was recognized correctly with only the Restructured HMM set, is shown in Fig. 7 We can see the fall around the 100th frame was restored. The hypotheses including "*no heja o*" by the initial HMM set were ranked below in the word graph. The proposed method properly restored local acoustic modeling mismatches and this pushed up the more accurate hypothesis to the first order.

For Example 2, the temporal behavior of the acoustic likelihood score near "*iqpaku sa se*", which was recognized correctly with only the Restructured HMM set, is shown in Fig. 8. Falls were restored at several points. The speech portion "*iqpaku sa se*" was pronounced rather dis-fluently and we could not find any hypotheses that included "*iqpaku sa se*" in the word graph given by the initial HMM set. The proposed method successfully kept the more accurate hypothesis in the word graph and ranked it to be of the first order.

5. CONCLUSION

A method of restructuring Gaussian mixture pdfs in a speaker-independent HMM based on speech data was proposed. Experimental results have proven that the proposed method effectively restores local acoustic modeling mismatches and improves the recognition performance. As future work, a study is being planned on a more appropriate model parameter re-estimation to enhance the effectiveness of the proposed method.

REFERENCES

[1] I. Zeljkovic: "Decoding optimal state sequence with smooth state likelihoods," *Proc. of ICASSP*-96, pp.129-132 (1996).

[2] T. Shimizu, H. Yamamoto and Y. Sagisaka: "Delayed decision beam search for continuous speech recognition," *Proc. of Acoust. Soc. Jap., Autumn* (1996) (in Japanese).

[3] A. Nakamura: "Predicting speech recognition performance," *Proc. of Eurospeech*-97, pp.1567-1570 (1997).

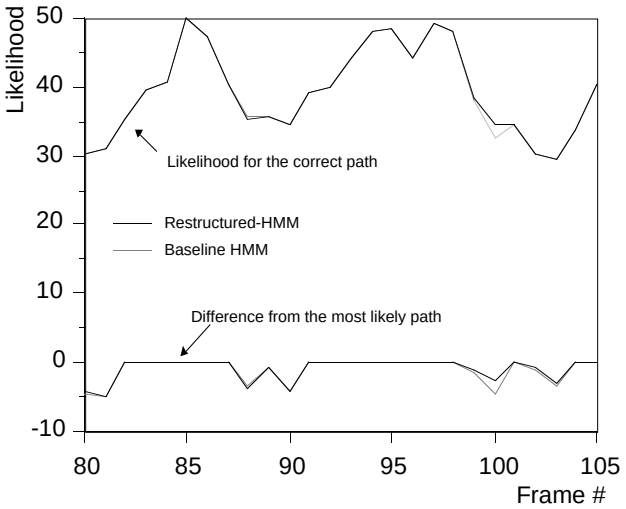

Fig. 7 Temporal behavior of the acoustic likelihood score (Example 1: "*no heja o*")

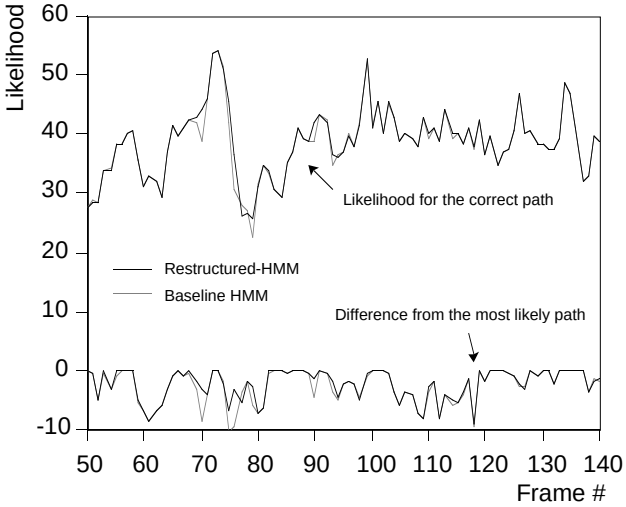

Fig. 8 Temporal behavior of the acoustic likelihood score (Example 2: "*iqpaku sa se*")

[4] M. Ostendorf and H. Singer: "HMM topology design using maximum likelihood successive state splitting," Computer Speech and Language,11, pp. 17-41 (1997).

[5] T. Morimoto, N. Uratani, T. Takezawa, O. Furuse, Y. Sobashima, H. Iida, A. Nakamura, Y. Sagisaka, N. Higuchi and Y. Yamazaki: "A speech and language database for speech translation research," *Proc. of ICSLP*-94, pp.1791-1794 (1994).

[6] A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura and Y. Sagisaka: "A Japanese speech databases for robust speech recognition," *Proc. of ICSLP*-96, pp. 2199-2202 (1996).

[7] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga and Y. Sagisaka: "Spontaneous dialogue speech recognition using cross-word context constrained word graphs," *Proc. of ICASSP*-96, pp.145-148 (1996).

[8] H. Masataki and Y. Sagisaka: "Variable order N-gram generation by word class splitting and consecutive word grouping," *Proc. of ICASSP*-96, pp. 188-191 (1996).