RECOGNITION METHOD WITH PARAMETRIC TRAJECTORY GENERATED FROM MIXTURE DISTRIBUTION HMMS

Yasuhiro Minami, Erik McDermott, Atsushi Nakamura, Shigeru Katagiri
Speech Open Laboratory, NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto-fu, Japan
{minami, mcd, ats, katagiri}@cslab.kecl.ntt.co.jp

ABSTRACT
We have proposed a new speech recognition technique that generates a speech trajectory from HMMs by maximizing the likelihood of the trajectory, while accounting for the relation between the cepstrum and the dynamic cepstrum coefficients. This method has the major advantage that the relation, which is ignored in conventional speech recognition, is directly used in the speech recognition phase. This paper describes an extension of the method for dealing with HMMs whose distributions are mixture Gaussian distributions. The method chooses the sequence of Gaussian distributions by selecting the best Gaussian distribution in the state during Viterbi decoding. Speaker-independent speech recognition experiments were carried out. The proposed method obtained an 18.2% reduction in error rate for the task, proving that the proposed method is effective even for Gaussian mixture HMMs.

1. INTRODUCTION

In each state of HMMs, an acoustic parameter vector is produced by a piecewise stationary process, and the probability of a given acoustic parameter vector is independent of the sequence of acoustic parameter vectors preceding, and following, the current vector. This means that HMMs cannot treat the time-dependent characteristics of speech within the state. This is one of the drawbacks of speech recognition using HMMs.

Several attempts to introduce time-dependence concept into speech recognition have been proposed to improve recognition performance [1][2][3][4][5]. Some of these, referred to as parametric trajectory modeling methods, or segmental modeling methods, represent the speech trajectories using linear or polynomial functions to treat the time-dependence in the speech signal. Such functions act as trajectories that are used to model observed sequences of moving points of the speech signal in the acoustic parameter space [3][4][5]. Although segmental modeling methods show some improvement in limited tasks, these methods are not widely used in speech recognition. This may be due to their tendency to be overly sensitive to variation in speaker and speaking style. The main cause of this is that the modeled trajectories are often not suited to use of a time warping function.

We have previously proposed a new method that uses the cepstrum trajectory generated directly from HMM statistics [6][7]. The technique maximizes the likelihood of the generated trajectory, taking into account the relation between the cepstrum and the dynamic cepstral coefficients (delta cepstrum and delta-delta cepstrum). This method can generate a trajectory for any HMM state sequence. As a result, we can introduce time warping mechanics into our method by selecting HMM state sequences so that the generated trajectory fits the input speech feature vector series. In addition, our method has the major advantage that the relations between the cepstrum and the dynamic cepstral coefficients, which are ignored in conventional speech recognition phase, can be introduced. In our previous paper, speaker-independent word recognition results showed that the proposed method was effective when a single Gaussian distribution in each state of HMMs is used. In this paper, we extend it so that our method can use mixture Gaussian distributions in HMMs.

2. TRAJECTORY GENERATION FROM HMMS

In this section, we present the method to generate a trajectory from given HMMs. This method is based on the studies of Tokuda et al. and Masuko et al. [8][9]. Figure 1 illustrates this procedure. A trajectory is generated from an input state se-

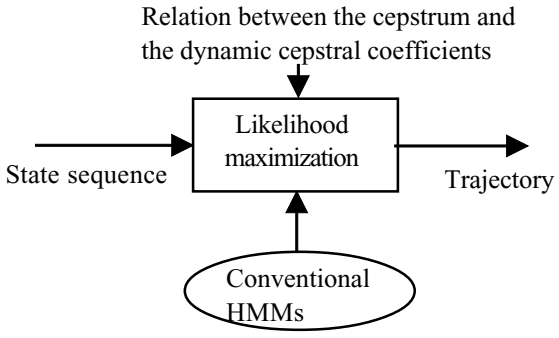

Figure 1. Diagram of generating trajectory from HMMs.

quence by maximizing the HMM likelihood while respecting the relations between cepstrum and the dynamic cepstral coefficients. Here, it is assumed that each state of the HMM has only a single Gaussian distribution, that the speech parameters consist of cepstrum, delta-cepstrum and delta-delta cepstrum, and that all HMMs have already been trained using a sufficient amount of data. It is also assumed that the HMM state sequence is given. Let $O=\{o_1,o_2,...,o_T\}$, $\Delta O=\{\Delta o_1,\Delta o_2,...,\Delta o_T\}$, and $\Delta^2 O=\{\Delta^2 o_1,\Delta^2 o_2,...,\Delta^2 o_T\}$ be a generated speech cepstrum vector sequence of length N, a delta-cepstrum vector sequence of length N, and a delta-delta cepstrum vector sequence of length N, respectively. Let $S=\{s_1,s_2,...,s_T\}$ be the given state sequence. The joint probability of O, ΔO, and $\Delta^2 O$, given the parameters of the diagonal Gaussian distributions, is given by

$$P(O,\Delta O,\Delta^2 O,S\mid M,\Delta M,\Delta^2 M,\Sigma,\Delta\Sigma,\Delta^2\Sigma)$$
$$=\prod_{t=1}^{T}a_{t,t+1}\prod_{t=1}^{T}p(o_t\mid\mu_t,\Sigma_t)p(\Delta o_t\mid\Delta\mu_t,\Delta\Sigma_t)p(\Delta^2 o_t\mid\Delta^2\mu_t,\Delta^2\Sigma_t),\quad(1)$$

where $M=\{\mu_1,\mu_2,...,\mu_T\}$, $\Delta M=\{\Delta\mu_1,\Delta\mu_2,...,\Delta\mu_T\}$, and $\Delta^2 M=\{\Delta^2\mu_1,\Delta^2\mu_2,...,\Delta^2\mu_T\}$ are the cepstrum mean vector sequence, the delta-cepstrum mean vector sequence, and the delta-delta cepstrum mean vector sequence of the Gaussian distributions along S respectively. Here, $\Sigma=\{\Sigma_1,\Sigma_2,...,\Sigma_T\}$, $\Delta\Sigma=\{\Delta\Sigma_1,\Delta\Sigma_2,...,\Delta\Sigma_T\}$, and $\Delta^2\Sigma=\{\Delta^2\Sigma_1,\Delta^2\Sigma_2,...,\Delta^2\Sigma_T\}$ are the cepstrum variance vector sequence, the delta-cepstrum variance vector sequence, and the delta-delta cepstrum variance vector sequence of the Gaussian distributions along S, respectively. Note that here the word "vector" is used for variances because we assume that variances are diagonal. Furthermore, $a_{t,t+1}$ is the transition probability from time t to time t+1, and $p(o\mid\mu,\Sigma)$ is a Gaussian distribution whose cepstrum mean vector and variance vector are μ and Σ, respectively.

The trajectory and dynamic parameters, O, ΔO, and $\Delta^2 O$, are decided by maximizing the probability expressed in equation (1). If there were no relation between O, ΔO, and $\Delta^2 O$, this would correspond to choosing the mean values of the Gaussian distributions. However, from the definition of the dynamic parameters, there are the following explicit relations between the parameters:

$$\Delta O_t=\sum_{i=-L}^{i=L}iO_{t+i}\Big/\sum_{i=-L}^{i=L}i^2\quad\text{and}\qquad(2)$$

$$\Delta^2 o_t=\frac{\sum_{i=-L}^{i=L}\{(2L+1)i^2+(\sum_{j=-L}^{j=L}j^2)\}o_{t+i}}{2\{(\sum_{j=-L}^{j=L}j^4)(2L+1)-(\sum_{j=-L}^{j=L}j^2)^2\}},\qquad(3)$$

where L is the window size.

To maximize equation (1) under these conditions, by substituting equations (2) and (3) into equation (1) and by differentiating with respect to O, we obtain the following equation:

$$\frac{\partial\log\{P(O,\Delta O,\Delta^2 O,S\mid M,\Delta M,\Delta^2 M,\Sigma,\Delta\Sigma,\Delta^2\Sigma)}{\partial O}$$
$$=\frac{\partial\log\{P(O,S\mid M,\Delta M,\Delta^2 M,\Sigma,\Delta\Sigma,\Delta^2\Sigma)}{\partial O}=0.\qquad(4)$$

By calculating equation (4) for all O_t, we can obtain simultaneous equations and solve them for O. Finally, ΔO and $\Delta^2 O$ can be obtained by calculating first order and second order regression coefficients, respectively.

3. NEW LIKELIHOOD AND TIME-WARPING MECHANISM

In Section 2, we described how to generate a cepstrum sequence for a given state sequence is known. Here, we describe how to decide the state sequence for input speech. To do that, we first introduce a new likelihood based on the generated trajectory for speech recognition. The conventional Viterbi likelihood for input speech using HMMs is expressed as:

$$P(C,\Delta C,\Delta^2 C,S\mid M,\Delta M,\Delta^2 M,\Sigma',\Delta\Sigma',\Delta^2\Sigma')$$
$$=\prod_{t=1}^{T}a_{t,t+1}\prod_{t=1}^{T}p(c_t\mid\mu_t,\Sigma_t)p(\Delta c_t\mid\Delta\mu_t,\Delta\Sigma_t)p(\Delta^2 c_t\mid\Delta^2\mu_t,\Delta^2\Sigma_t),\quad(5)$$

where $C=\{c_1,c_2,...,c_T\}$ is an input speech cepstral vector sequence. The method proposed here changes the likelihood into a new likelihood function using the trajectory:

$$P(C,\Delta C,\Delta^2 C,S\mid O,\Delta O,\Delta^2 O,\Sigma',\Delta\Sigma',\Delta^2\Sigma')$$
$$=\prod_{t=1}^{T}a_{t,t+1}\prod_{t=1}^{T}p(c_t\mid o_t,\Sigma_t')p(\Delta c_t\mid\Delta o_t,\Delta\Sigma_t')p(\Delta^2 c_t\mid\Delta^2 o_t,\Delta^2\Sigma_t').\quad(6)$$

Since mean cepstrum vectors are replaced by the trajectory in equation (6), we have to reestimate new variances, $\Sigma'=\{\Sigma_1',\Sigma_2',\ldots,\Sigma_T'\}$, $\Delta\Sigma'=\{\Delta\Sigma_1',\Delta\Sigma_2',...,\Delta\Sigma_T'\}$ and $\Delta^2\Sigma'=\{\Delta^2\Sigma_1',\Delta^2\Sigma_2',...,\Delta^2\Sigma_T'\}$, along with the trajectory. It is assumed that new variances are fixed within a state, as in the conventional HMM definition. Figure 2 illustrates piecewise stationary mean sequence and the distribution spreads (related to the variances) in conventional HMMs, where the dotted and dashed line shows the input speech. Figure 3 illustrates the smooth trajectory and the new distribution spreads (related to the new variances) in the proposed method, showing that the generated trajectory has explicit time-dependence. Up to this point, we have focused on the new likelihood . From here on, we describe how to select the state sequence using the likelihood. The basic concept is that the state sequence whose trajectory is nearest to the input speech cepstrum sequence (in terms of the new likelihood) should be

selected. The following two equations can be used:

$$S = \arg\max_{S}\{P(C,\Delta C,\Delta^2 C,S \mid O,\Delta O,\Delta^2 O,\Sigma',\Delta\Sigma',\Delta^2\Sigma')\}, \quad (7)$$

and

$$O = \arg\max_{O}\{P(O,S \mid M,\Delta M,\Delta^2 M,\Sigma,\Delta\Sigma,\Delta^2\Sigma)\}, \quad (8)$$

where equation (8) is for generating the trajectory from the given states. However, it is difficult to calculate equations (8) and (8) for all S due to combinatorial explosion. We do not, therefore, calculate equation (8) for all possible state sequences; we calculate equation (8) only for the best state sequence decoded by the Viterbi algorithm using conventional HMMs, as an approximation of (7) and (8).

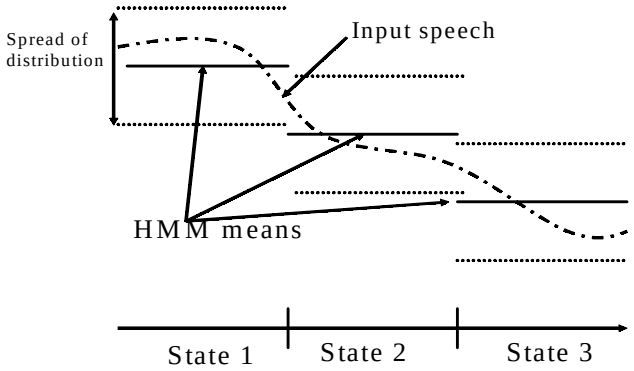

Figure 2. Illustration of piece-wise stationary mean sequence and the distribution spreads in conventional HMMs.

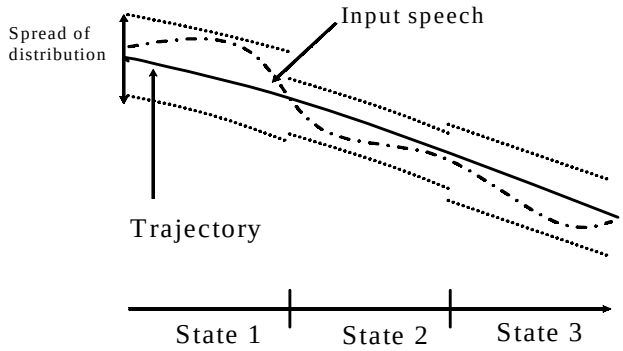

Figure 3. Illustration of the smooth trajectory and the new distribution spreads in the proposed method.

4. TRAINING VARIANCES

As described in section 3, we suppose that variances are fixed in a state. Thus, we can introduce the same training method, referred to as Viterbi training, to calculate the variances for each state along with the trajectory. The basic procedure is:

(1) Create usual HMMs for training data by the MLE based Viterbi training method.

(2) Calculate Viterbi paths using these HMMs for all training data.

(3) Generate trajectories for the training data using the method described in Section 2.

(4) Divide all the data into short data segments state by state using state paths from (2).

(5) Calculate variance values using equation (9),

$$\Sigma_s' = \frac{\sum_{k=1}^{n}\sum_{t=1}^{T_k}(c_t^k - o_t^s)(c_t^k - o_t^s)^T}{\sum_{k=1}^{n} T_k}, \quad (9)$$

where n denotes the number of data segments assigned to a state, s, by the Viterbi algorithm, c_t^k denotes the kth sample of length T_k, and o_t^k is the corresponding trajectory generated by our method. The procedure is performed only once. Although equation (8) describes only Σ', $\Delta\Sigma'$ and $\Delta^2\Sigma'$ can easily be obtained using a similar equation. Each obtained variance is stored in the corresponding HMM state in addition to the original HMM state variance.

5. EXTENSION TO MIXTURE DISTRIBUTION HMMS

Until now, this paper has described our method based on single Gaussian HMMs. We can easily extend the method to treat mixture Gaussian distributions by selecting the best Gaussian distribution in the state during Viterbi decoding.

The large number of Gaussian components in each state might

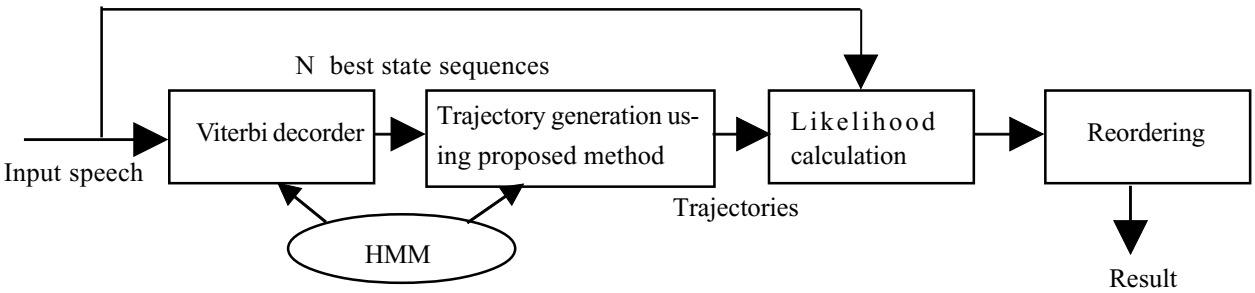

Figure 4. Recognition process based on trajectory generation and N-best reordering.

degrade the accuracy of the approximation described in section 3, as the number of possible Gaussian sequences increases exponentially with the number of mixture components. Here, however, we would like to check the applicability of our method for Mixture Gaussian HMMs. Thus, we introduce this assumption into our method.

The training method is also changed to handle Gaussian mixture HMMs. Step (2) in section 4 is changed to select the maximum Gaussian distribution during Viterbi path calculation.

6. RECOGNITION EXPERIMENTS

Since the generating method described in Section 2 calculates a trajectory value at a frame using past and future frame information, it is difficult to introduce an efficient search method such as the Viterbi algorithm, which requires only information from the previous frame, into our method in the recognition phase. Thus, the recognition procedure presented in Figure 4 was used. First, input speech is recognized using conventional HMMs, and the top ten candidates are generated. State-based segmentation is performed by the Viterbi algorithm for each candidate to obtain putative state durations given the input utterance. The trajectory for each candidate is then generated using the method described in Section 2. Given the generated trajectory, frame-wise likelihood between the generated trajectory and the input speech cepstrum parameters is calculated, and the original candidates are reordered according to the likelihood scores. The frame-wise modified likelihood is

$$P(C,\Delta C,\Delta^2 C,S \mid O,\Delta O,\Delta^2 O,\Sigma',\Delta\Sigma',\Delta^2\Sigma') = \prod_{t=1}^{T} a_{t-1,t} \prod_{t=1}^{T} p(c_t \mid o_t,\Sigma_t')\, p(\Delta c_t \mid \Delta o_t,\Delta\Sigma_t')^{\alpha}\, p(\Delta^2 c_t \mid \Delta^2 o_t,\Delta^2\Sigma_t')^{\beta}, \quad (9)$$

where α and β are weights for the delta-cepstrum and delta-delta cepstrum likelihood, respectively. Our preliminary experiments described in [7] showed that these weighting values are effective for our method. The mixture weights are ignored in the likelihood calculation.

To evaluate the method proposed here, we performed speaker and task independent word recognition experiments. The sampling rate was 16 kHz, the frame shift was 10 msec and the cepstrum order was 14; 503 phoneme-balanced sentences uttered by 64 speakers were used for the training data. Context-dependent HMMs were trained from the data, and the number of Gaussian distributions for each state was fixed at 3. One hundred place names uttered by 70 speakers were used for the evaluation. Table 1 shows the speaker independent recognition results. The maximum recognition rates were selected among results for α and β values of 0, 1, 2, 3, 4 and 5. While HMM word error rate was 2.2%, that of our method was 1.8%, corresponding to an error rate reduction of 18.2%. This result shows that our proposed method is effective even for mixture distributions.

7. SUMMARY

This paper extended a new speech recognition method that generates a speech trajectory using a speech synthesis method so as to enable the use of Gaussian mixture distributions in HMMs. The training of variances was also extended. Our method was evaluated with a speaker independent speech recognition experiment. Our method yielded an 18.2% reduction in error rate for the recognition task, proving that our method is effective even for Gaussian mixture HMMs.

8. REFERENCES

[1] M. Ostendorf, V. Digalakis and O. Kimball , "From HMMs to segment models: A unified view of stochastic modeling for speech recognition", IEEE Transactions on Speech and Audio Processing, vol. 4, no. 5, pp. 360-378, 1996.

[2] S. Rocous, M. Ostendorf, H. Gish and A. Derr, "Stochastic segment modeling using the estimate-maximize algorithm", Proc. ICASSP, pp. 127-130, 1988.

[3] H. Gish and K. Ng, "Parametric trajectory models for speech recognition", Proc. ICASSP, pp. 447-450, 1993.

[4] W. J. Holmes and M. J. Russell, "probabilistic-trajectory segmental HMMs", Computer Speech and Language, vol. 13, pp. 3-37, 1999.

[5] R. Iyer, H. Gish. M.-H. Siu, G. Zavaliagkos and S. Matsoukas, "Hidden Markov models for trajectory modeling", Proc ICSLP, pp. 891-894, 1998.

[6] Y. Minami, E. McDermott, A. Nakamura, S. Katagiri, "A Recognition Method Using Synthesis-Based Scoring That Incorporates Direct Relations Between Static And Dynamic Feature Vector Time Series", Workshop for Consistent & Reliable Acoustic Cues for Sound Analysis, 2001.

[7] Y. Minami, E. McDermott, A. Nakamura and S. Katagiri, "A recognition method with parametric trajectory synthesized using direct relations between static and dynamic feature vector time series", Proc. ICASSP, pp. 957-960, 2002.

[8] K. Tokuda, T. Kobayashi and S. Imai, "Speech parameter generation from HMM using dynamic features", Proc. ICASSP, pp. 660-663, 1995.

[9] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Speech synthesis from HMMs using dynamic features", Proc. ICAASP, pp. 389-392, 1996.

Table 1 Speaker-independent speech recognition results (word error rates).

Method	HMM	Proposed method
Word error rate	2.2%	1.8%
Reduction in error rate	-	18.2%