A RECOGNITION METHOD USING SYNTHESIS-BASED SCORING THAT INCORPORATES DIRECT RELATIONS BETWEEN STATIC AND DYNAMIC FEATURE VECTOR TIME SERIES

Yasuhiro Minami, Erik McDermott, Atsushi Nakamura, Shigeru Katagiri
Speech Open Laboratory
NTT Cyber Space Laboratories
NTT Communication Science Laboratories
NTT Corporation
2-4, Hikaridai Seika-cho Soraku-gun Kyoto, 619-0237 Japan.

ABSTRACT

It is well known that hidden Markov models (HMMs) can only exploit the time-dependence in the speech process in a limited way. Parametric trajectory models have been proposed to exploit this time-dependency. However, parametric trajectory modeling methods are unable to take advantage of efficient HMM training and recognition methods. This paper describes a new speech recognition technique that generates a speech trajectory mean using a HMM-based speech synthesis method. This method generates an acoustic trajectory by maximizing the likelihood of the trajectory taking into account the relation between the cepstrum, delta-cepstrum, and delta-delta cepstrum. Speaker dependent and speaker independent speech recognition experiments showed that the proposed method is effective for speech recognition.

1. INTRODUCTION

It is well known that HMMs can only exploit the time-dependence in the speech process in a limited way, because in 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. The sequence of moving points of the speech signal in the acoustic parameter space is referred to as a speech trajectory. Several attempts to introduce the trajectory concept into speech recognition have been proposed to improve the 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 as linear or polynomial functions to treat the time-dependence in the speech signal. Such functions act as "trajectory means" that are used to model observed trajectories [3][4][5]. However, segmental modeling methods are unable to take advantage of efficient HMM training and recognition methods. In this paper, we propose a new speech method that generates a speech trajectory mean using conventional HMMs. Iyer et al. have demonstrated that HMMs, by and large, produce trajectories that are representative of the input data [5]. However, they pointed out that HMMs model these trajectories without the knowledge of the inherent trajectory structure in the input features [5]. We think it is possible to introduce time-dependence into the trajectory generated from the HMMs, if the relations between the cepstrum, delta-cepstrum, and delta-delta cepstrum are properly used. Recently, a technique that generates smooth speech parameter sequences using trained HMM parameters has been introduced in the field of speech synthesis [6][7][8]. The technique maximizes the likelihood of the generated speech taking into account the relation between the cepstrum and the dynamic cepstral coefficients (delta cepstrum and delta-delta cepstrum), assuming that the state sequence and Gaussian distributions are given. Given this procedure, the generated trajectory explicitly has time-dependency. The technique can obtain the most likely acoustic parameter trajectory for any HMM state sequence. This means that when the state sequence is given, a representative parameter trajectory can be generated. In our method, the cepstrum trajectory generated by the technique is then used as a trajectory mean for speech recognition. In the following sections, we present the generation of cepstrum trajectory means and then evaluate recognition performance when using the generated trajectory means for speaker dependent and independent speech recognition

2. SPEECH SYNTHESIS USING HMM

In this section, we present the synthesis method based on the studies of Tokuda et al. and Masuko et al. [6][7][8]. Here it is assumed that each state of the HMM has only a single Gaussian distribution (it is easy to extend the number of Gaussian components in the state mixture), 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 synthesized speech cepstrum vector sequence of length T, a delta-cepstrum vector sequence of length T, and a delta-delta cepstrum vector sequence of length

T respectively. Let $S = \{s_1, s_2, s_3, \ldots s_T\}$ be the given state sequence. The joint probability of O, ΔO, and $\Delta^2 O$ and S given the parameters of the Gaussian distributions is given by

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

where $\mathrm{M} = \{\mu_1, \mu_2, ..., \mu_T\}$, $\Delta\mathrm{M} = \{\Delta\mu_1, \Delta\mu_2, ..., \Delta\mu_T\}$, and $\Delta^2\mathrm{M} = \{\Delta^2\mu_1, \Delta^2\mu_2, ..., \Delta^2\mu_T\}$ 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, and $\Sigma = \{\Sigma_1, \Sigma_2, \ldots, \Sigma_T\}$, $\Delta\Sigma = \{\Delta\Sigma_1, \Delta\Sigma_2, ..., \Delta\Sigma_T\}$, and $\Delta^2\Sigma = \{\Delta^2\Sigma_1, \Delta^2\Sigma_2, ..., \Delta^2\Sigma_T\}$ 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, and $a_{t,t+1}$ is the transition probability from time t to time t+1. O, ΔO, and $\Delta^2 O$ are decided by maximizing the probability. If there was no relation between O, ΔO, and $\Delta^2 O$, these would be the mean values of the Gaussian distributions. However, from the definition of the dynamic parameters, there are the following explicit relations between those values:

$$\Delta o_t = \frac{\sum_{i=-L}^{i=L} i o_{t+i}}{\sum_{i=-L}^{i=L} i^2} \quad (2)$$

$$\Delta^2 o_t = \frac{\sum_{i=-L}^{i=L} i \Delta o_{t+i}}{\sum_{i=-L}^{i=L} i^2} \quad (3)$$

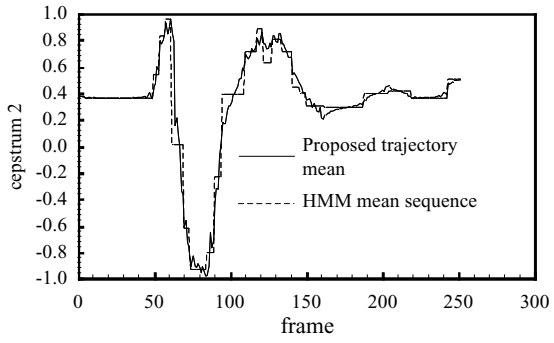
where L is the window size. To maximize equation (1) under these conditions, by substituting equations (2) and (3) in equation (1), equation

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

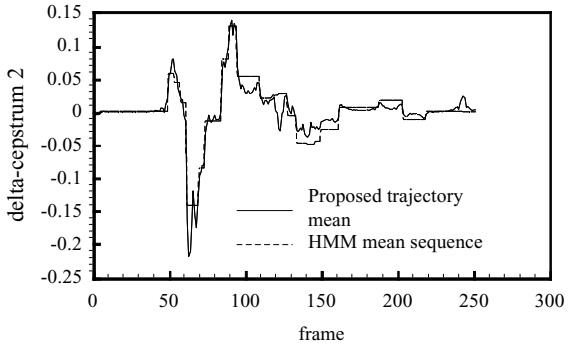
is calculated. By differentiating equation (4) for all O_t, we can obtain simultaneous equations and solve for O, ΔO, and $\Delta^2 O$ (see the studies of Tokuda et al. and Masuko et al. [4][5][6] for more details on these equations). An example of an obtained speech parameter sequence (trajectory mean) using this technique is shown in Figure 1. Figure 1 (a) shows the second cepstrum coefficient in a word utterance. Figure 1 (b) shows the corresponding coefficient of the delta-cepstrum for the same utterance. The horizontal and vertical axes indicate the frame number (5 ms shift) and cepstrum value, respectively. The dotted lines in the figure show the sequence of the mean values of the HMM states along the given state sequence. The solid lines show the trajectory mean synthesized by this technique. We can see that the trajectory means are smoother than the sequences of the HMM state mean values. At the stationary parts (where the HMM means come from the same state), the generated trajectory means are not stationary, as they are affected by the output probabilities of the neighborhood frames. On the other hand, at the state boundaries, they do not have discontinuities.

3. GENERATING TRAJECTORY MEANS

In Section 2, we described how to generate the trajectory mean only when the state sequence is known. In this section, we describe how to decide the state sequence. The basic concept is that the trajectory which is nearest to the input speech cepstrum sequence should be selected. Let $C = \{c_1, c_2, \ldots, c_T\}$ be



(a) 2nd order cepstrum coefficients



(b) 2nd order delta-cepstrum coefficients

Figure 1. An example of the cepstral coefficients generated by proposed method.

an input speech cepstrum sequence. The following two equations were used:

$$S = \arg\min_{S} \{dist(C - O(S))\} \quad (5)$$

$$O(S) = \arg\max_{O} \{P(O,S|,\mathrm{M},\Delta\mathrm{M},\Delta\Delta\mathrm{M},\Sigma,\Delta\Sigma,\Delta\Sigma)\}, \quad (6)$$

where dist() is a function that calculates the distance between the input speech parameters and the generated speech parameters. However, it is hard to calculate equation (6) for all S due to combinatorial explosion. We therefore do not calculate equation (6) for all possible state sequences, but select the best state sequence decoded by the Viterbi algorithm with the HMMs, and calculate equation (6) only for the state sequence, as an approximation of (5) and (6). We can easily extend this method to mixture Gaussian distributions by selecting the best Gaussian distribution in the state during Viterbi decoding. However, if the number of Gaussian components in each state mixture is large, the accuracy of this approximation is degraded, because the number of possible Gaussian sequences increases exponentially with the number of mixture components. Thus, we fix the number of the mixture components at 1 in this paper.

4. RECOGNITION EXPERIMENTS

To evaluate the method here, speaker independent and speaker dependent word recognition experiments were performed. The experiments aimed to examine the effectiveness for recognition of the trajectory mean generated by our method. In this paper, we only describe how to generate the trajectory mean, not how to generate the trajectory variance. Thus, the evaluation procedure diagrammed in Figure 2 was used to evaluate the performance only for the trajectory mean. First, input speech is recognized using the HMMs, and the top three candidates are generated. State based segmentation is carried out for each candidate to obtain putative state durations given the input utterance. The trajectory mean for each candidate is then generated using the method described in Sections 2 and 3. Given the generated trajectory mean, frame-wise distances between the generated trajectory mean and the input speech cepstrum parameters are calculated, and the original candidates are reordered according to the distance scores. These results are compared with the results obtained from reordering the candidates according to the distances between the HMM mean vector sequences along the Viterbi alignment of the candidates and the input speech cepstrum parameters. The frame-wise distance calculation is

$$dist = \sum_{t=1}^{T} |c_t - o_t|^n + \lambda_1 |\Delta c_t - \Delta o_t|^n + \lambda_2 |\Delta\Delta c_t - \Delta\Delta o_t|^n, \quad (7)$$

where λ_1 and λ_2 are weights for the delta-cepstrum and delta-delta cepstrum distances. The maximum recognition rates were obtained from n values of 1 and 2, and λ_1 and λ_2 values of 2, 5, and 20.

4.1 Speaker Dependent Speech Recognition Experiment

The sampling rate was 10 kHz, the frame shift was 5 msec, and the order of the cepstrum coefficients was 15. The numeric utterances, syllabic utterances, 503 phoneme balanced sentences, and 216 isolated words from the ATR database were used for training data. Speaker MHT was used. Context dependent phoneme HMMs were trained using this data. The number of Gaussian components was fixed at 1 for each state. 5240 words uttered by the same speaker were used for the evaluation.

4.2 Speaker Independent Speech Recognition Experiment

This evaluation was done under more usual recognition condition. The sampling rate was 16 kHz, the frame shift was 10 msec and the cepstrum order was 14; 503 phoneme balanced sentences

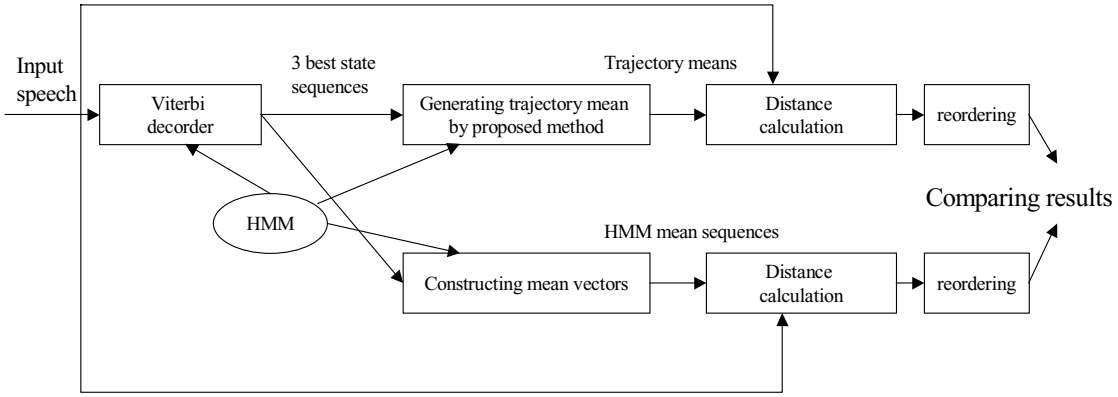


Figure 2. Diagram of evaluation for proposed method.

uttered by 64 speakers were used for the training data. Context dependent HMMs were trained from the data. The number of Gaussian distributions for each state was fixed at 1. One hundred place names uttered by 10 speakers were used for the evaluation.

5. EXPERIMENTAL RESULTS

5.1 Speaker Dependent Speech Recognition Results

Table 1 shows the speaker dependent speech recognition results. Column 1 of the table shows the recognition rate of the HMM mean vector sequence using the distance computation from equation (7). The recognition rate of the proposed method is shown in column 2. Comparing column 1 and column 2, the recognition rate of the proposed method is 2% higher than that of the mean value sequence of the HMMs along the Viterbi alignment. This result is very positive. In this paper, we have only described how to generate the trajectory mean, not how to generate the variance. Although it is not easy to obtain accurate trajectory variance [9], if we have a good method for this, the performance of the proposed trajectory-based method will be significantly improved.

5.2 Speaker Independent Speech Recognition Results

Table 2 shows the speaker independent recognition results. It is clear that our method improved the recognition rate. This shows that the proposed trajectory mean method was also effective for speaker independent speech recognition. In our most recent work with this approach, we performed an experiment using HMMs that have six Gaussian mixture distributions per state to check the approximation described in Section 3. Table 3 shows the results of this experiment. From these results, the recognition accuracy of the proposed method did not indicate any improvement. Looking into the Viterbi alignments of the input speeches, we found that the ID numbers of the mixture distributions were changing frequently within the same state during a short frame length. Because of this, a stable trajectory could not be obtained. To obtain a stable trajectory, we must calculate equations (5) and (6) with a more accurate method.

6. SUMMARY

This paper describes a new speech recognition method that generates a speech trajectory mean using a speech synthesis method. The speech synthesis method used here generates a speech cepstrum trajectory by maximizing the likelihood taking into account the definitional relation between the cepstrum, delta-cepstrum, and delta-delta. The generated speech cepstrum trajectory is used as the trajectory mean for recognition. We evaluated our method with speaker dependent and speaker independent speech recognition experiments. The results showed that our method was effective for speech recognition. We will extend our method in a more general manner in the future so that it can be used to train the trajectory variance and treat mixtures of Gaussian distributions.

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

[7] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi and S. Imai, “An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features”, Proc. Eurospeech, pp. 757-760, 1995.

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

[9] T. Fukada, Y. Sagisaka and K K. Palliwal, “Model parameter estimation for mixture density polynomial segment models”, Proc. ICASSP, pp. 1403-1406, 1997.

Table 1. Speaker dependent word recogniton rate.

HMM mean vector sequence	proposed trajectory mean
86.4%	88.7%

Table 2. Speaker independent word recogniton rate (1 mixture Gaussian components).

HMM mean vector sequence	proposed trajectory mean
93.4%	94.4%

Table 3. Speaker independent word recogniton rate (6 mixture Gaussian components).

HMM mean vector sequence	proposed trajectory mean
96.2%	96.2%