MULTIPLE-REGRESSION HIDDEN MARKOV MODEL

Katsuhisa Fujinaga, Mitsuru Nakai, Hiroshi Shimodaira and Shigeki Sagayama

Japan Advanced Institute of Science and Technology
Tatsu-no-Kuchi, Ishikawa, 923-1292 Japan
{kfujina,mit,sim,sagayama}@jaist.ac.jp

ABSTRACT

This paper proposes a new class of hidden Markov model (HMM) called multiple-regression HMM (MR-HMM) that utilizes auxiliary features such as fundamental frequency (F_0) and speaking styles that affect spectral parameters to better model the acoustic features of phonemes. Though such auxiliary features are considered to be the factors that degrade the performance of speech recognizers, the proposed MR-HMM adapts its model parameters, i.e. mean vectors of output probability distributions, depending on these auxiliary information to improve the recognition accuracy. Formulation for parameter reestimation of MR-HMM based on the EM algorithm is given in the paper. Experiments of speaker-dependent isolated word recognition demonstrated that MR-HMMs using F_0 based auxiliary features reduced the error rates by more than 20% compared with the conventional HMMs.

1. INTRODUCTION

Spectral parameters of phonemes are influenced by number of factors, not only gender, speakers, contexts, but also speaking styles, fundamental frequency (F_0) and so on. The challenge of improving the recognition accuracy of HMM is regarded as a problem of how to neutralize the influence by those factors that degrade the recognition performance. So far, a number of efforts have been made towards speaker adaptation (MAP[1], VFS[2], MLLR[3]) and context-dependent modeling (HMNet[4]), while only a few towards speaking style and F_0 adaptation or normalization.

In spoken dialogue systems, even a single human could speak in many different styles. For example, when the system misrecognizes the speech, the user tends to speak more clearly, slowly to emphasize the misrecognized words. These sorts of speaking styles that are different from the normal utterance style cause lower accuracy of the recognizer[5]. The first step for this problem is to use separate acoustic models for the specific speaking styles [6]. Next step which is discussed in this paper will be to explore some adaptation or normalization techniques, hopefully online or frame-synchronous adaptation of HMM against the utterance variations.

Based on a knowledge that spectral features have some correlation with F_0, Singer and Sagayama [7] showed that spectrum normalization by a phoneme-wise linear regression model between F_0 and cepstral features could improve the phoneme recognition accuracy. This approach assumed that the regression coefficient did not change within a phoneme. But it would be more natural that the coefficient can vary according to the change of spectral features even within a certain phoneme. To realize such dynamic processing, spectrum normalization by F_0 and phoneme recognition should be done at the same time. This can be achieved by embedding the adaptation or normalization operation into the HMM formulation, in other words, developing a new class of HMM that adapts its model parameters depending on F_0 or other auxiliary features. Among such class of HMM, multi-regression HMM (MR-HMM), the one that employs multiple regression to modify the model parameters, i.e. mean vectors of normal distributions, is discussed here.

It should be noted that the proposed MR-HMM is completely different from the existing autoregressive HMM (AR-HMM) [8] which assumes that observation vectors are drawn from an auto-regression process.

This paper is organized as follows: the next section describes the basic formulation of MR-HMM and EM-based parameter reestimation algorithm. The third section presents experimental results of speaker-dependent isolated word recognition. Finally, the last section is devoted to conclusions.

2. MULTIPLE-REGRESSION HMM

2.1. Outline of MR-HMM

Fig. 1 shows the correlation between $\log F_0$ and the 7th mel-cepstral coefficient (MCEP) of a phoneme sample s/e/i (/e/ preceded by /s/ and followed by /i/). It can be seen from the figure that the 7th MCEP has a negative correlation with $\log F_0$. This sort of influence of F_0 has been observed on formant frequencies of vowels, and it has been explained from the biomechanical and phonological point of view [9]. This evidence implies that F_0 can be of help to recover the original spectral features from the observed spectral parameters.

Since the correlation between the spectral features and F_0 varies depending on contexts, phonemes or sub-

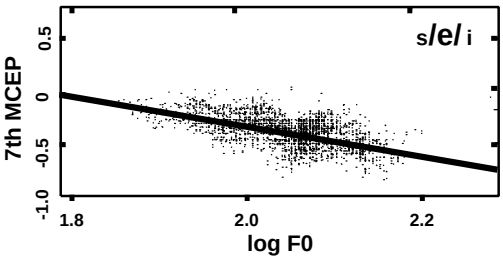

Fig. 1. Correlation between F_0 and MCEP.

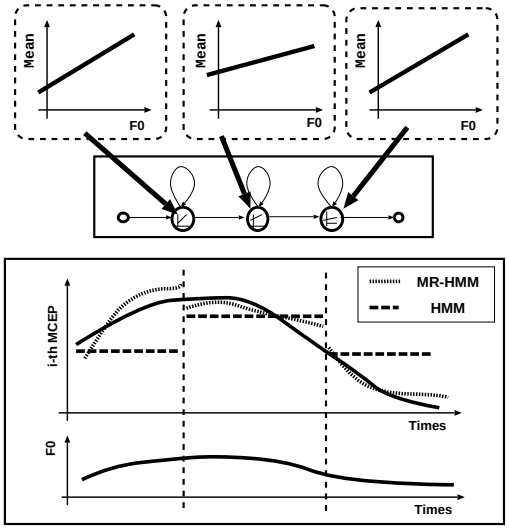

Fig. 2. Idea of Multiple-Regression HMM using F_0 as an auxiliary feature.

phonemes, normalization of spectrum parameters by using F_0 should be done simultaneously with speech recognition. MR-HMM gives such framework by changing its parameters according to F_0 and other possible features.

Fig. 2 shows a basic idea of MR-HMM. The bottom box in the figure illustrates the i th MCEP coefficient as an input feature to HMM, F_0 as an auxiliary feature, as well as the mean values of output probability distributions of each state of both MR-HMM and standard HMM. In the framework of conventional HMMs, the mean value of each state does not change, while, in the MR-HMM, the mean value of a certain time instance t changes based on the regression line (the upper 3 boxes in the figure) given as a function of $F_0(t)$, i.e. the fundamental frequency at time t. Let μ be an element of a mean vector, then μ at time t is modeled as

$$\mu = r_0 + r_1 \log F_0(t), \tag{1}$$

where r_0, r_1 are the regression coefficients. In a general case where M auxiliary features (predictor variables in terms of multiple regression) are given, the above formulation is now rewritten as

$$\mu = r_0 + r_1\xi_1 + \cdots + r_M\xi_M. \tag{2}$$

In the sense of adapting the model parameters, MR-HMM is similar to MLLR for speaker adaptation excepting to the point that MLLR uses the same feature parameters with the ones used for recognition while MR-HMM utilizes auxiliary features that are not used directly for recognition but used for adapting the model parameters on-line.

2.2. Probability evaluation in MR-HMM

Since MR-HMM differs from standard HMM only on the point that the former uses auxiliary features to calculate probability distributions but the latter not, most parts of its formulation is same with HMM. So, the notations a_{ij} (state transition probability) and π_i (initial state probability) used here have the same meanings with those in HMM.

Let $\boldsymbol{\mu}$ and $\boldsymbol{U}$ be the mean vector and the covariance matrix of a Gaussian distribution, respectively. The output probability density function $b_i(\boldsymbol{x_t}|\boldsymbol{\xi_t})$ of state i for a given N-dimensional observation vector, $\boldsymbol{x_t} = [x_{1t}, x_{2t}, \cdots, x_{Nt}]'$, and M-dimensional auxiliary vector, $\boldsymbol{\xi_t} = [\xi_{1t}, \xi_{2t}, \cdots, \xi_{Mt}]'$, is defined as

$$b_i(\boldsymbol{x_t}|\boldsymbol{\xi_t}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{U}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x_t}-\boldsymbol{\mu}(\boldsymbol{\xi_t}))'\boldsymbol{U_i}^{-1}(\boldsymbol{x_t}-\boldsymbol{\mu}(\boldsymbol{\xi_t}))}, \tag{3}$$

where $\boldsymbol{\mu}(\boldsymbol{\xi_t})$ is given by

$$\boldsymbol{\mu}(\boldsymbol{\xi_t}) = \boldsymbol{R_i}\hat{\boldsymbol{\xi}}_{\boldsymbol{t}}, \tag{4}$$

$$\hat{\boldsymbol{\xi}}_{\boldsymbol{t}} = (1, \boldsymbol{\xi_t}) = (1, \xi_{1t}, \xi_{2t}, \cdots, \xi_{Mt})',$$

where $\boldsymbol{R_i}$ is an $N \times (M+1)$-dimensional multiple regression matrix.

The probability of observing a vector sequence $\boldsymbol{X} = (\boldsymbol{x_1} \cdots \boldsymbol{x_T})$ when given an auxiliary vector sequence $\boldsymbol{\xi} = (\boldsymbol{\xi_1} \cdots \boldsymbol{\xi_T})$ is expressed as

$$\begin{aligned} f(\boldsymbol{X}|\lambda, \boldsymbol{\xi}) &= \sum_{\theta \in \boldsymbol{\Theta}} f(\boldsymbol{X}, \theta|\lambda, \boldsymbol{\xi}) \\ &= \sum_{\theta \in \boldsymbol{\Theta}} \pi_{\theta_1} b_{\theta_1}(\boldsymbol{x_1}|\boldsymbol{\xi_1}) \prod_{t=2}^{T} a_{\theta_{t-1}\theta_t} b_{\theta_t}(\boldsymbol{x_t}|\boldsymbol{\xi_t}), \end{aligned}$$

where λ denotes a set of parameters of MR-HMM, θ means a state sequence $(\theta_1, \cdots, \theta_T)$ and $\boldsymbol{\Theta}$ expresses the set of all possible state sequences. $f(\boldsymbol{X}, \theta|\lambda)$ is the probability density of observing $\boldsymbol{X}$ with θ given $\boldsymbol{\xi}$ and λ.

It is easy to see that the forward/backward algorithm and Viterbi algorithm can be used to evaluate the above probability by just replacing the output probability density function of conventional HMM with the one given by equation (3).

2.3. Parameter estimation of MR-HMM

The parameters of MR-HMM including the multiple regression matrix $\boldsymbol{R_i}$ can be trained based on a maximum likelihood optimization criterion as well as HMM. Though both EM algorithm and Viterbi training algorithm are applicable to MR-HMM, only the EM based reestimation is described here.

Table 1. Hand-segmented phoneme recognition results by MR-HMM compared with the conventional HMM

models	features	auxiliary features	% errors				% reduction			
			Vowel	V-Cons	UV-Cons	ALL	Vowel	V-Cons	UV-Cons	ALL
HMM	C	-	17.6	28.4	38.5	18.7	-	-	-	-
	C + $\tilde{P}$	-	17.6	28.1	36.4	18.5	−0.3	2.2	2.4	0.8
	C + $\Delta\tilde{P}$	-	17.4	29.0	35.7	18.3	0.9	−1.7	7.1	2.0
	C + R	-	18.2	30.0	39.1	19.1	−3.5	−4.6	−2.1	−1.9
	C + $\tilde{P}$ + R	-	18.4	29.7	38.0	19.1	−4.7	−3.1	−1.4	−2.3
	C + $\Delta\tilde{P}$ + R	-	18.4	30.3	36.4	18.9	−4.6	−5.7	5.3	−0.7
MR-HMM	C	$\tilde{P}$	16.9	28.3	33.8	17.6	3.7	1.7	12.1	5.9
	C	$\Delta\tilde{P}$	17.4	28.2	31.5	17.8	1.2	1.1	14.9	4.9
	C	R	15.2	20.1	21.5	14.5	13.1	30.0	44.3	21.6
	C	$\tilde{P}$+R	15.2	19.7	19.9	14.3	13.3	31.3	48.4	23.1
	C	$\Delta\tilde{P}$+R	14.8	19.2	19.8	14.0	15.5	33.3	49.2	24.4

(C: MCEP(13)+ΔMCEP(13), Vowel:/a,i,u,e,o/, V-Cons:/b,d,g/, UV-Cons:/p,t,k/, ALL:26 phonemes)

The following reestimation formulas of parameters are derived by iteratively maximizing an auxiliary function given by

$$Q(\lambda, \bar{\lambda}) = \sum_{\theta \in \mathbf{\Theta}} f(\boldsymbol{X}, \theta|\lambda) \log(f(\boldsymbol{X}, \theta|\bar{\lambda})),$$

where λ and $\bar{\lambda}$ denote the current parameters and the reestimated parameters, respectively.

A set of parameter reestimation formulas is described by

$$\bar{\boldsymbol{R}}_{\boldsymbol{i}} = \left(\sum_{t=1}^{T} \gamma_t(i) \boldsymbol{x_t} \hat{\boldsymbol{\xi}}_{\boldsymbol{t}}'\right) \left(\sum_{t=1}^{T} \gamma_t(i) \hat{\boldsymbol{\xi}}_{\boldsymbol{t}} \hat{\boldsymbol{\xi}}_{\boldsymbol{t}}'\right)^{-1},$$

$$\bar{\boldsymbol{U}}_{\boldsymbol{i}} = \frac{\sum_{t=1}^{T} \gamma_t(i) (\boldsymbol{x_t} - \boldsymbol{R_i} \hat{\boldsymbol{\xi}}_{\boldsymbol{t}}) (\boldsymbol{x_t} - \boldsymbol{R_i} \hat{\boldsymbol{\xi}}_{\boldsymbol{t}})'}{\sum_{t=1}^{T} \gamma_t(i)},$$

$$\bar{\pi}_i = \gamma_1(i), \qquad \bar{a}_{ij} = \frac{\sum_{t=2}^{T} \gamma_{t-1}(i,j)}{\sum_{t=2}^{T} \gamma_{t-1}(i)},$$

where γ_t is the probability of being in state i at t, $\gamma_{t-1}(i,j)$ is the probability of being in state i at $t-1$, and state j at t. Both γ_t and $\gamma_{t-1}(i,j)$ are calculated by the same manner with the conventional HMM.

In case that no auxiliary features are given, the above equations are equivalent to those for the conventional HMM.

3. EXPERIMENTS

3.1. Experimental setup

The formulation of MR-HMM given in the previous section does not restrict the sorts of auxiliary features that are used as the explanatory variables of multiple-regression. Since this is the first attempt to evaluate the MR-HMM for speech recognition, F_0 was chosen as a basic auxiliary feature. To exclude the influence of other features such as speaking rate, and to compare the recognition accuracy between the MR-HMM and conventional HMMs, read-speech database uttered in normal speaking style was used.

The proposed MR-HMM was evaluated in speaker-dependent hand-segmented phoneme recognition and isolated word recognition experiments.

Speech data of 4 people (2 male and 2 female) were collected from the ATR A-set at a sampling frequency of 16 kHz. 13 mel-cepstral coefficients (MCEPs) and 13 delta mel-cepstral coefficients (ΔMCEPs) were calculated with a frame length of 25ms and a frame shift of 5ms. Both MCEPs and ΔMCEPs include the 0th coefficients. F_0 were calculated with a frame length of 40ms and a frame shift of 5ms. To extract F_0, the cepstrum method was employed. In the experiments, we tried three auxiliary features, that were $\log F_0$ with linear interpolation for unvoiced sounds ($\tilde{P}$), $\Delta\tilde{P}$, and the power of low-frequency-band spectrum (R) as a simple indicator of voiced-sound existence.

In training, the odd numbered words out of the 5240 Japanese common words, and the 516 phonetically balanced words were used. In testing, the even numbered words out of the 5240 words were used. The phoneme categories for recognition were / N, a, b, ʧ, d, e, f, g, h, i, ʒ, k, m, n, o, p, q, r, s, ʃ, t, ʦ, u, w, j, z / .

3.2. Phoneme Recognition Experiments

Table. 1 shows the experimental results, in which context-independent, 3-state, single-mixture left-to-right HMM is employed with a diagonal covariance matrix for each output probability distribution.

In all the cases, MR-HMM reduced the error rates successfully compared with the baseline HMM that does not use any auxiliary features. Surprisingly, the feature R contributes to increase the recognition accuracy than $\tilde{P}$ does. Since the value of R has a connection with pitch existence, this results indicate that HMM parameters should be adapted depending on pitch existence and such adaptation

Table 2. Isolated word recognition results by MR-HMM compared with the conventional HMM

models	features	auxiliary features	% errors	% reduction
HMM	C	-	4.7	-
	C+$\tilde{P}$	-	5.2	-12.3
	C+$\Delta\tilde{P}$	-	4.7	-1.6
	C+R	-	4.8	-2.9
	C+$\tilde{P}$+R	-	5.3	-15.3
	C+$\Delta\tilde{P}$+R	-	5.1	-9.3
MR-HMM	C	$\tilde{P}$	4.2	8.6
	C	$\Delta\tilde{P}$	4.4	4.4
	C	R	**3.4**	**23.0**
	C	$\tilde{P}$+R	**3.4**	**23.0**
	C	$\Delta\tilde{P}$+R	**3.5**	**21.1**

(C: MCEPs(13)+ΔMCEP(13))

is automatically taking place in the MR-HMM when given the feature R. Though those auxiliary features are effective for MR-HMM, they are not for the conventional HMM in which they are incorporated into the observation vectors.

3.3. Isolated Word Recognition Experiments

Table. 2 shows the experimental results, in which context-dependent, single-mixture, left-to-right model with a diagonal covariance matrix for each output probability distribution was used. The ML-SSS algorithm [10] was employed to train the context-dependent HMMs with 406 states and MR-HMMs having the same topologies with the conventional HMMs. In testing, the even numbered words out of the 5240 words were used excepting the 225 words that contain phonemes not appearing in the training data. The lexicon was comprised of all testing words. Half of testing words were used for the evaluation.

It can be seen from the table that MR-HMM reduced the error rate by 8.6% ($\tilde{P}$), 23.0% (R) compared with the baseline HMM. On the other hand, conventional HMM failed to reduce the errors even they were fed any of those auxiliary features. This might be caused by the "curse of dimensionality" problem, i.e. adding any of auxiliary features and increasing the dimension of feature vector of HMM can lead the recognition system to poorer results.

Fig. 3 illustrates the average variances of the output probability densities of MR-HMM in comparison with those of conventional HMM. We can see that MR-HMM has smaller variances than conventional HMM, especially in the lower order MCEPs. This result indicates that MR-HMM represents the information of the training data more efficiently than conventional HMM.

4. CONCLUSION

The proposed MR-HMM is a general framework for incorporating extra features into HMM not like the way of just

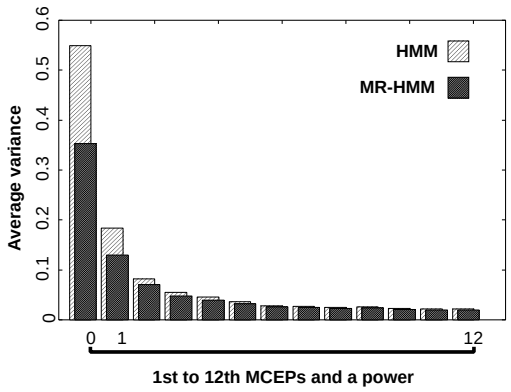

Fig. 3. Average variances of models.

combining the new features with the existing features. Although F_0 and R were considered in this paper, other features that have some correlations with the existing features can be employed. The authors are extending its formulation to adapt not only the mean vectors but also the covariance matrices of the distributions. Furthermore, the proposed model is applicable to speech synthesis to control the speaking style.

5. REFERENCES

[1] C.H. Lee, C.-H. Lin, and B.-H. Juang. A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models. *IEEE Trans. SP*, 39(4):806–814, April 1991.

[2] K. Ohkura, M. Sugiyama and S. Sagayama. Speaker Adaptation based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs. In *Proc. ICSLP-92*, pp.369–372, 1992.

[3] C. J. Leggetter and P. C. Woodland. Speaker Adaptation of HMM's Using Linear Regression. Technical Report TR.182, Cambridge University, 1994.

[4] Jun Takami and Shigeki Sagayama. A Successive State Splitting Algorithm for Efficient Allophone Modeling. In *Proc. ICASSP-92*, volume I, pp.573–576, 1992.

[5] H. Soltau and A. Waibel. On the influence of hyperarticulated speech on the recognition performance. In *Proc. ICSLP98*, pp.229–232, 1998.

[6] H. Soltau and A. Waibel. Specialized acoustic models for Hyperarticulated Speech. In *Proc. ICASSP-2000*, pp.1779–1782, 2000.

[7] H. Singer and S. Sagayama. Pitch Dependent Phone Modeling for HMM Based Speech Recognition. In *Proc. ICASSP-92*, volume I, pp.273–276, 1992.

[8] B.-H. Juang. Mixture autoregressive hidden Markov models for speech signals. *IEEE ASSP*, 33(6):1404–1413, 1985.

[9] K. Honda. Relationship Between Pitch Control and Vowel Articulation. In *Vocal Fold Physiology*, pp.286–289, College-Hill Press, 1983.

[10] M. Ostendorf and H. Singer. HMM Topology Design Using Maximum Likelihood Successive State Splitting. *Computer Speech and Language*, 11(1):17–41, 1997.