A NONSTATIONARY AUTOREGRESSIVE HMM WITH GAIN ADAPTATION FOR SPEECH RECOGNITION

*Ki Yong Lee and *Joohun Lee*

School of Electronics Engr., Soongsil University, 1-1, Sangdo 5Dong, Dongjak-Ku, Seoul 156-743, Korea

Tel: +82-2-820-0908 Fax: +82-2-814-3627 E-mail: kylee@saint.soongsil.ac.kr

*Dept of Information and Telecommunication, Dong-Ah Broadcasting College, Ansung, Korea

ABSTRACT

In this paper, a time domain approach for speech recognition is developed. The nonstationary autoregressive (AR) hidden markov model (HMM) with gain contour is proposed for modeling the statistical characteristics of the speech signal. The parameter of nonstationary AR model was modeled by the polynomial function with linear combination of *M* known basis functions. In this proposed model, speech signal is blocked by samples into fixed-length frames and modeled by nonstationary AR model controlled by markov switching sequences at each frame. Given the HMM parameter set of the speech, the gain-adapted recognition algorithm is developed for speech recognition.

1. Introduction

The autoregressive hidden markov model (ARHMM) [1,2] is useful methods to represent the statistical characteristics of the clean speech in speech recognition and speech enhancement. In the conventional ARHMM, individual states are assumed to be stationary stochastic sequences. This stationary-state assumption appears to be reasonable when a state is intended to represent piece-wise stationary segment of speech. Since speech sounds, such as fricative, glides, liquids, and transition regions between phones, reveal the most notable nonstationary nature [3,4], we can not expect to obtain the better performance by the conventional methods based on the above assumption. Another basic issue arising from ARHMM for speech recognition is matching problem of the energy contour of the signal to the energy contour of the model for that signal. The energy matching is usually attained by appropriate normalization but is not applicable when only noisy speech signals are available [5]. In this letter, for energy matching we propose a training algorithm of nonstationary ARHMM with gain adaptation and gain adapted speech recognition algorithm.

To overcome these problems, the nonstationary ARHMM with gain adaptation for gain normalized clean signals are design using maximum likelihood (ML) estimates of the gain contours of the clean training sequences. The parameter of nonstationary AR model was modelled by the polynomial function with linear combination of *M* known basis functions. Then, the speech signal is blocked by samples into fixed-length frames and modeled by nonstationary AR model with frame varying polynomial function controlled by markov switching sequences at each frame. Our model is formally very similar to the trend HMM [3,4], but it is designed to handle the speech signal at the frame level, where it is represented by the signal, rather than dealing with feature vectors directly. Also, for *M*=0, the proposed model become to conventional ARHMM [5]. The proposed method for speech recognition is combined with ML estimates of the gain contours of the clean test signals, obtained from the given clean signals, in performing recognition using the maximum *a posteriori* decision rule.

We have evaluated our nonstationary ARHMM method with *M*=1 on a base of ten isolated Korean digits with three versions of each digit pronounced by seven male speakers. In speech recognition, the proposed method was compared with a conventional ARHMM method with gain-normalized approach and with gain adaptation.

2. Nonstationary AR-HMM with gain adatation for clean speech

Let $y = y_n, n = 1,..,T$ be the sequence of clean signal vectors, where $\mathbf{y}_n = \{y(t, (n-1)N+1 \le t \le nN$ and $s_n \in \{1,...,L$, be a sequence of states corresponding to *y*. Let $g = \{g_n, n = 1,..,T$, be a sequence of gain factors, or a gain contour, for the signal *y*.

Then, at *n*-th frame speech signal conditioned on state *i* is expressed as a linear combination of its past values plus an excitation source with gain contour, as

$$y(t = \sum_{k=1}^{p} \sum_{m=0}^{M} B_k^i(m\, y(t-k + g_n \cdot e_i(t, \quad (n-1)N+1 \le t \le nN . \tag{1}$$

where $B_k^i\ m$ is the state-dependent time-varying coefficients, $\mathbf{y}\ t-1 = y\ t-1\ y\ t-2\ \mathrm{L}\ y\ t-p\ ^T$, $e_i \cdot$ is the excitation source with state-dependent variance $\sigma_i^2\ n$, N and n is the frame length and number, respectively, and $g_n > 0$ for all n is a gain term to take into account the mismatch between training data and testing data for the clean speech models.

We now turn our attention to the problem of estimating the time-varying coefficients in our model. In order to gain insight into the behavior of the coefficients and to make the estimation problem tractable. We choose to model them as a linear combination of *M* known basis functions:

$$B_k^i\ n = \sum_{m=0}^{M} B_{k,m}^i f_m\ n \tag{2}$$

where $f_m\ n$ represents the *m*th basis function and $B_{k,m}^i$ the weight associated with the basis function.

Here, we choose M=1 and our basis functions to be such that

$$f_0(n) = 1, \qquad 1 \le n \le T,$$
$$f_1(n) = n, \qquad 1 \le n \le T. \tag{3}$$

Therefore, (1) may be rewritten by vector form as

$$y(t) = \boldsymbol{B}^i \boldsymbol{y}(t-1) + g_n \cdot e_i(t), \quad n-1 \, N+1 \le t \le nN \tag{4}$$

where $\boldsymbol{B}^i = B^i_{1,1}\ B^i_{1,2}\ B^i_{2,1}\ B^i_{2,2}\ \mathrm{L}\ B^i_{p,1}\ B^i_{p,2}$ and

$$\boldsymbol{y}(t-1) = \left[y\big((n-1)N+t-1\big),\ my\big((n-1)N+t-1\big)\mathrm{L}\ , \right.$$
$$\left. y\big((n-1)N+t-p\big),\ my\big((n-1)N+t-p\big) \right]^T .$$

In essence, the model with time-varying coefficients has been transformed into one with time-invariant weights. The problem is now reduced to one of estimating $2P$ time-invariant parameters that completely characterize the behavior of the coefficients. It should be noted that the choice of basis functions is by no means limited to polynomials.

The likelihood of the observation sequence y under the model λ and gain contour g is calculated as

$$p_\lambda(y|g) = \sum_s p_\lambda(s, y|g)$$
$$= \sum_s p_\lambda(s|g)\, p_\lambda(y|s, g), \tag{5}$$

where $p_\lambda(s|g)$ denotes the probability of sequence of states s, and $p_\lambda(y|s,g)$ is the probability density function (pdf) of the sequence of out vectors y given g and s. For first-order HMM's, $p_\lambda(s|g)$ is given by

$$p_\lambda(s|g) = \prod_{n=1}^{T} a_{s_n - 1 s_n}, \tag{6}$$

where $a_{s_n - 1 s_n}$ denotes the transition probability from state at time t-1 to state at time t.

The $p_\lambda(y|s,g)$ is

$$p_\lambda(y|s,g) = \prod_{n=1}^{T} p_\lambda(y_n|s_n, g_n) \tag{7}$$

where the pdf $p_\lambda(y_n|s_n,g_n)$ of the vector y_n given that this vector was generated from state s_n and gain contour g_n.

Then, from (4) $p_\lambda(y_n|s_n,g_n)$ is given by

$$p_\lambda(y_n|s_n,g_n) = \prod_{t=(n-1)N}^{nN} \frac{\exp\left\{ -\dfrac{\big(y(t) - \boldsymbol{B}^{s_n} \boldsymbol{y}(t-1)\big)^2}{2 g_n^2 \sigma_{s_n}^2} \right\}}{\sqrt{2\pi g_n^2 \sigma_{s_n}^2}} \tag{8.a}$$

or in matrix form

$$p_\lambda(y_n|s_n,g_n) = \frac{\exp\left(-\dfrac{1}{2} y_n^T C_{s_n}^{-1} y_n \right)}{(2\pi)^{N/2} \det^{1/2}(C_{s_n})}. \tag{8.b}$$

where $C_{s_n} = g_n^2 \sigma_{s_n}^2 \left(A_{s_n}^T A_{s_n} \right)^{-1}$ and A_{s_n} is a $N \times N$ lower triangular Toeplitz matrix in which the first elements $2P+1$ of the first column constitute the coefficients of the AR process.

The parameter set $\lambda = \left\{ a_{ij}, \boldsymbol{B}^j, \sigma_j^2, i, j = 1, ..., L \right\}$ of the nonstationary ARHMM and gain contour g for the clean speech is estimated from training sequences of clean speech signals. Note that λ denotes the parameter set of the ARHMM for the gain-normalized signal.

3. Gain adapted training algorithm

Gain-adapted training of the nonstationary ARHMM for the word results from ML estimation of the parameter set λ from a training sequence y from word using an ML estimate of the gain contour g. Then, λ can be estimated from

$$\max_\lambda \max_g p_\lambda(y|g) \tag{9}$$

However, the gradient equations of $p_\lambda(y|g)$ with respect to $\{\lambda, g\}$ are nonlinear and therefore have no simple solution. Hence, the estimation of $\{\lambda, g\}$ is performed here iteratively using the expectation-maximization (EM) approach [6,7]. Then, each iteration constitutes one EM iteration for estimating a value of λ given g and one EM iteration for estimating g using the resulting λ.

The training algorithm for nonstationary ARHMM with gain contour can be summarized as follows.

3.1 Estimation of λ;

Assuming that the gain contour g_n for all n is known, the estimation formulas for λ can be derived from maximizing Baum's auxiliary function [3,4]. Each iteration of the Baum algorithm starts with an old set of parameters, say λ_{l-1}, and estimates a new set of parameters, say λ_l, by maximizing the following auxiliary function:

$$Q_1(\lambda, g) = \sum_s p_{\lambda_{l-1}}(s|y,g) \log p_{\lambda_l}(y, s|g). \tag{10}$$

The objective function can be simplified to

$$Q_1(\lambda, g) = \sum_{ij=1}^{L} \sum_{n=1}^{T} p_{\lambda_{l-1}}(s_{n-1} = i, s_n = j|y,g), \times \left[\log a_{ij} + \log p_{\lambda_l}(y_n|s_n,g_n) \right] \tag{11}$$

where

$$\log p_{\lambda_l}(y_n|s_n,g_n) = -\frac{N}{2} \log\left(2\pi g_n^2 \sigma_j^2\right) - \sum_{t=(n-1)N+1}^{nN} \frac{\big(y(t) - \boldsymbol{B}^j \boldsymbol{y}(t-1)\big)^2}{g_n^2 \sigma_j^2}$$

and $p_{\lambda_{l-1}}(s_{n-1} = i, s_n = j|y,g)$ is the *posteriori probability* of

the transition from state *i* to state *j* given the signal y and the gain contour g_n at the (*l*-1)th iteration.

The auxiliary function $Q_1(\lambda, g)$ is optimized by differentiation with respect to each of $a_{ij}, \boldsymbol{B}^j, \sigma_j^2, (i, j = 1,..., L)$, respectively. As in standard HMM, the reestimation formulas of the Markov chain and nonstationary AR parameter are established by setting

$$\frac{\partial Q_1}{\partial a_{ij}} = 0,$$
$$\frac{\partial Q_1}{\partial \boldsymbol{B}^j} = 0,$$
$$\frac{\partial Q_1}{\partial \sigma_j^2} = 0, \qquad (12)$$

for *i,j*=1,…, *L*, subject to the constraint $\sum_{j=1}^{L} a_{ij} = 1$. We obtain the reestimation formula

$$a_{ij} = \frac{\sum_{n=1}^{T} p_{\lambda_{l-1}}(s_{n-1} = i, s_n = j | y, g)}{\sum_{j=1}^{L}\sum_{n=1}^{T} p_{\lambda_{l-1}}(s_{n-1} = i, s_n = j | y, g)} \qquad (13)$$

where the probability $p_{\lambda_l}(s_{n-1} = i, s_n = j | y, g)$ can be calculated efficiently by the forward-backward algorithm.
Also, the reestimation formulas for nonstationary AR parameter are obtained by solving

$$B^j = \left[\sum_{i=1}^{L}\sum_{n=1}^{T} p_{\lambda_{l-1}}(s_{n-1} = i, s_n = j | y, g) \sum_{t=(n-1)N+1}^{nN} \hat{\boldsymbol{y}}(t-1)\right] \left[\hat{\boldsymbol{y}}^T(t-1)\right]^{-1} \left[\sum_{i=1}^{L}\sum_{n=1}^{T} p_{\lambda_{l-1}}(s_{n-1} = i, s_n = j | y, g) \times \sum_{t=(n-1)N+1}^{nN} \hat{y}(t)\hat{\boldsymbol{y}}(t-1)\right] \qquad (14)$$

and

$$\sigma_j^2 = \frac{1}{\sum_{i=1}^{L}\sum_{n=1}^{T} p_{\lambda_{l-1}}(s_{n-1} = i, s_n = j | y, g)} \times \left[\sum_{i=1}^{L}\sum_{n=1}^{T} p_{\lambda_{l-1}}(s_{n-1} = i, s_n = j | y, g) \times \sum_{t=(n-1)N+1}^{nN} \left(\hat{y}(t) - \boldsymbol{B}^j \hat{\boldsymbol{y}}(t-1)\right)^2\right] \qquad (15)$$

where $\hat{y}(t) = \frac{y(t)}{g_n}$ and $\hat{\boldsymbol{y}}(t-1) = \left[\frac{y(t-1)}{g_n} n \frac{y(t-1)}{g_n} \frac{y(t-2)}{g_n} n \frac{y(t-2)}{g_n} \text{L} \frac{y(t-p)}{g_n} n \frac{y(t-p)}{g_n}\right]^T$ are normalized observation sequence by gain contour.

If *M* is set to zero, then (14)-(15) reverts to the reestimation formula for the Gaussian mean vectors in the standard ARHMM. This is expected since, as the time-varying component in the trend function is removed, the resulting degenerated nonstationary AR HMM is no different from the standard ARHMM.

3.2 Estimation of gain contour *g*;

Next, for estimating the gain contour g, assume that λ is known. Let g_{l-1} and g_l be a current and a new estimate of the gain contour of the signal *y*, respectively. Then, the objective function for g_l is given by

$$Q_2(\lambda, g) = \sum_s p_\lambda(s | y, g_{l-1}) \log p_\lambda(s | y, g_{l-1}) = \sum_{i,j=1}^{L}\sum_{n=1}^{T} p_\lambda(s_{n-1} = i, s_n = j | y, g_{l-1}) \times \left[\log a_{ij} + \log p_\lambda(s | y, g_{l-1})\right]. \qquad (16)$$

Similarly to what we have seen before, maximization of the auxiliary function over *g* results in an estimate of the gain contour *g* by setting the gradient of $Q_2(\lambda, g)$ with respect to *g* to zero; $\frac{\partial Q_2}{\partial g^2} = 0$.
We arrive at the following gain reestimation formula

$$g_{l,n}^2 = \sum_{j=1}^{L} p_\lambda(s_n = j | y, g_{l-1}) \sum_{t=(n-1)N+1}^{nN} \frac{\left(y(t) - \boldsymbol{B}^j \boldsymbol{y}(t-1)\right)^2}{\sigma_j^2}. \qquad (17)$$

The iteration process start with initial gain contour $g_{l=0,n} = 1$ for all *n*, and repeated until either a fixed point $\lambda_{l+1} = \lambda_l, g_{l+1,n} = g_{l,n}$ is reached or the difference in likelihood in two consecutive iterations is sufficiently small.

4. Gain-adapted Speech recognition

In speech recognition, given the word speech sequence *y* the decision rule for spoken word *y* is

$$\max_{1 \le i \le Z} \max_g p_\lambda(y | W_i, g). \qquad (18)$$

where *Z* is number of the total word for speech recognition.
The algorithm for local maximization of $p_\lambda(y|g)$ over g can be summarized as follows;

Step-0: Initialization: For given the parameter $\lambda = \{a_{ij}, \boldsymbol{B}^j, \sigma_j^2, i, j = 1,..., L\}$, $g_0 = 1$, and ε, evaluate $p_\lambda(y|g_0)$ and *l*=0.

Step-1: Gain estimation: Calculate the posterior probabilities $p_\lambda(s_n | y, g_l)$ for $s_n = 1,..., L$ and $n = 1,..., T$, and estimate g_{l+1} using (17).

Step-2: If $p_\lambda(y|g_{l+1}) - p_\lambda(y|g_l) \le \varepsilon$,

assign $\underset{g}{\text{Max}}\, p_\lambda(y|g) = p_\lambda(y|g_{l+1})$ and stop

Otherwise, set $l \rightarrow l+1$ and goto **Step-1**.

5. Experimental Results

We have evaluated our new method on a base of ten isolated Korean digits with three versions of each digit pronounced by seven male speakers. Only 50 speech data of five speakers have participated in training and other 160 speech data have been used for test. This speech data were sampled at 10kHz and modeled by state L=5 and AR order of 15. Training and recognition were performed on nonoverlapping vectors of the speech word whose dimension was N=256. The proposed method with gain-adapted recognition was compared with a conventional gain-normalized approach [2]. For the data (A) of speakers who participated in training and those (B) of the speakers who didn't participate in training, the recognition rates of the conventional training algorithm has scored 96 % and 90%, respectively while those of the proposed algorithm has scored 98.8% and 93.2 %, respectively. Table 1 shows a results of comparison of the recognition accuracies. Table 2 shows a results of comparison on conventional ARHMM with gain adaptation [5], and proposed method with gain adaptation (in this case, conventional method is equal to proposed method with M=0) and M=1. From this result, the proposed method had a good performance.

Table 1. Recognition results

Data	Gain-normailzed	Gain-adapted
A	96%	98.8%
B	90%	93.2%

Table 2. Recognition results

Data	ARHMM with M=0	ARHMM with M=1
A	98%	98.8%
B	92.5%	93.2%

ACKNOWELEDGEMENT: **This work was supported in part by KOSEF (971-0917-105-1)**

References

[1]A.B. Poritz, "Hidden markov models: A guided tour," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1988, pp. 7-13.

[2] B. Juang and L.R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 1404-1413, Dec. 1986.

[3] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for speech signal," *Signal Processing*, 27, pp. 65-72, 1992.

[4] L.Deng, M.Aksmanovic, X. Sun, and C.F. JeffWu," Speech recognition using HMM with polynomial regression functions as nonstationary states," *IEEE Trans. Speech and Audio Processing*, vol.2, no.4, pp.507-520, Oct. 1994.

[5] Y. Ephraim, "Gain adapted hidden Markov models for recognition of clean and noisy speech," IEEE Trans. Signal Processing, vol. 40, no.6, pp.1303-1316, June 1992.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc. B*, vol. 39, no.1, pp. 1-38, 1977

[7] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41, pp.164-171, 1970.