Speech Recognition and Enhancement by A Nonstationary AR HMM with Gain Adaptation under Unknown Noise

Gunther Ruske and Ki Yong Lee*

*Inst. For Human-Machine-Communication, Munich University of Technology, Germany

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

ABSTRACT

In this paper, a gain-adapted speech recognition in unknown noise is developed in time domain. The noise is assumed to be the colored noise. The nonstationary autoregressive (NAR) hidden markov model (HMM) used to model clean speeches. The nonstationary AR is modeled by polynomial functions with a linear combination of M known basis functions. Enhancement using multiple Kalman filters is performed for the gain contour of speech and estimation of noise model when only the noisy signal is available.

I. Introduction

The autoregressive hidden markov model (ARHMM) [1,2] is useful method to represent clean speeches in speech recognition and enhancement. In the conventional ARHMM, individual states are assumed to be stationary stochastic sequences. Since speech sounds, such as fricative, glides, liquids, and transition regions between phones, reveal the most notable nonstationary nature [3-5], we can not expect to obtain better performance by the conventional methods based on the above assumption. Another basic issue of ARHMM is arising from speech recognition, when only noisy speech signals are available. It is an estimation problem of unknown noise, and a matching problem of the energy contour of signal to the model for the same signal.

To overcome these problems, the gain adapted speech recognition with an NAR-HMM and noise estimation is presented. The NAR-HMM used to model clean speeches and the parameter of nonstationary AR model were the linear combinations of M known basis functions. Then, the speech signal is blocked by samples into fixed-length frames. Our model is formally very similar to the trend HMM [3,4], but it is designed to handle speech signals at the frame level, where it is represented by signals rather than dealing with feature vectors directly. Also, for M=0, the proposed model become to the conventional ARHMM [2]. When only the noisy signal is available, the gain adapted recognition algorithm with noise estimation are developed for the NAR-HMM using the EM approach and tested in recognition of noise speech. Enhancement using multiple Kalman filters is performed for the gain contour of speech and estimation of noise model

II. NAR-HMM for clean speech

Let $y=\{y_n, n=1,..,T\}$ be the sequence of clean signal vectors, where $y_n=\{y(t),\ (n-1)N+1\le t\le nN\}$ 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 the n-th frame, speech signal conditioned on state j is expressed as a linear combination of its past values plus an excitation source with gain contour, as

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

where $B_k^j(m)$ is the state-dependent frame-varying coefficients, $e_j(\cdot)$ is the excitation source with state-dependent variance $\sigma_j^2(n)$, 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 clean speech models.

We now turn our attention to the problem of estimating 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:

$$B_k^j(n)=\sum_{m=0}^{M}B_{k,m}^j f_m(n) \quad (2)$$

where $f_m(n)$ represents the m-th basis function and $B_{k,m}^j$ the weight associated with the basis function.

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

$$f_0(n)=1,\qquad 1\le n\le T\ ,$$
$$f_1(n)=n,\qquad 1\le n\le T\ . \quad (3)$$

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

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

where $\boldsymbol{B}^j=\left[B_{1,0}^j\ B_{1,1}^j\ B_{2,0}^j\ B_{2,1}^j\cdots B_{p,0}^j\ B_{p,1}^j\right]$ and

$$Y(t-1)=\left[y((n-1)N+t-1),ny((n-1)N+t-1)\cdots,\right.$$
$$\left.y((n-1)N+t-p),\ ny((n-1)N+t-p)\right]^T\ .$$

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

$$p_\lambda(y|g) = \prod_{n=1}^{T} a_{s_{n-1}s_n} p_\lambda(y_n|s_n, g_n), \tag{5}$$

where $a_{s_{n-1}s_n}$ denotes the transition probability from state at time n-1 to state at time n, and the $p_\lambda(y_n|s_n, g_n)$ is

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

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 clean speeches is estimated from training sequences of clean speech signals. Note that λ denotes the parameter set of the ARHMM for the gain-normalized signal.

III. Training algorithm

Gain-adapted training of the nonstationary ARHMM results from maximum likelihood (ML) estimation of the parameter set λ and of the gain contour from a training sequence y. Then, λ and g can be estimated from

$$\max_{\lambda} \max_{g} p_\lambda(y|g) \tag{6}$$

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 EM approach [6,7], maximizing the following auxiliary function:

$$Q(\lambda, g) = \sum_{s} p_{\lambda'}(s|y, g) \log p_\lambda(y, s|g). \tag{7}$$

where λ' and λ are current and new estimates of model parameter, respectively.

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.

O. Estimation of λ;

As in standard HMM, we obtain the reestimation formula

$$a_{ij} = \frac{\sum_{n=1}^{T} p_{\lambda'}(s_{n-1} = i, s_n = j|y, g)}{\sum_{j=1}^{L}\sum_{n=1}^{T} p_{\lambda'}(s_{n-1} = i, s_n = j|y, g)}, \tag{8}$$

$$\boldsymbol{B}^j = \left[\sum_{i=1}^{L}\sum_{n=1}^{T} p_{\lambda'}(s_{n-1} = i, s_n = j|y, g) \sum_{t=(n-1)N+1}^{nN} \bar{Y}(t-1)\bar{Y}^T(t-1)\right]^{-1} \cdot \left[\sum_{i=1}^{L}\sum_{n=1}^{T} p_{\lambda'}(s_{n-1} = i, s_n = j|y, g) \sum_{t=(n-1)N+1}^{nN} \bar{y}(t)\bar{Y}(t-1)\right], \tag{9}$$

$$\sigma_j^2 = \frac{\sum_{i=1}^{L}\sum_{n=1}^{T} p_{\lambda'}(s_{n-1} = i, s_n = j|y, g) \sum_{t=(n-1)N+1}^{nN} \left(\bar{y}(t) - \boldsymbol{B}^j \bar{Y}(t-1)\right)^2}{\sum_{i=1}^{L}\sum_{n=1}^{T} p_{\lambda'}(s_{n-1} = i, s_n = j|y, g)} \tag{10}$$

where the probability $p_{\lambda'}(s_{n-1} = i, s_n = j|y, g)$ can be calculated efficiently by the forward-backward algorithm, $\bar{y}(t) = \dfrac{y(t)}{g_n}$, and

$$\bar{Y}(t-1) = \left[\frac{y(t-1)}{g_n}\ n\frac{y(t-1)}{g_n}\ \frac{y(t-2)}{g_n}\ n\frac{y(t-2)}{g_n}\ \cdots\ \frac{y(t-p)}{g_n}\ n\frac{y(t-p)}{g_n}\right]^T.$$

Note g_n is a gain factor of the n-th frame.

If M is set to zero, then (8)-(10) reverts to the reestimation formula for the Gaussian mean vectors in the standard ARHMM.

O. Estimation of gain contour g;

Next, for estimating the gain contour g, assume that λ is known. 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 $Q(\lambda, g)$ with respect to g to zero. We arrive at the following gain reestimation formula

$$g_n^2 = \sum_{j=1}^{L} p_{\lambda'}(s_n = j|y, g') \sum_{t=(n-1)N+1}^{nN} \frac{\left(y(t) - \boldsymbol{B}^j Y(t-1)\right)^2}{\sigma_j^2}. \tag{11}$$

The iteration process starts with initial gain contour $g_n' = 1$ for all n, and repeats until either a fixed point $\lambda = \lambda'$, $g_n = g_n'$ is reached or the difference in likelihood in two consecutive iterations becomes sufficiently small.

IV. Noisy Speech recognition

When only the noisy speech $z=y+v$ is available, where $z = \{z_n, n = 1,..,T\}$ and $v = \{v_n, n = 1,..,T\}$, gain factor of speech would be estimated from noisy speech by matching the energy contour of the test clean signal to that of the model for the same signal. The noise model assumed here is a nonwhite stochastic AR of order q:

$$v(t) = \boldsymbol{C}^T V(t-1) + g_v w(t), \tag{12}$$

where $V(t-1) = [v(t-1), \cdots, v(t-q)]^T$, $\boldsymbol{C} = [c_1, \cdots, c_q]^T$ are the AR parameter vector of the noise process, g_v represents its power, and $w(t)$ is white Gaussian process with zero mean and unit variance. Since the noise parameters $\boldsymbol{C}$ and g_v are unknown a priori, also, they must be estimated within the speech recognition algorithm.

Given the model λ from clean speech, the likelihood of the noisy speech z under the noise model $\lambda_v = \{C, g_v\}$ and gain

contour g is calculated as

$$p_\lambda(z|g,\lambda_v)=\sum_{s}\int p_\lambda(s,y,z|g,\lambda_v)dy, \quad (13)$$

where

$$p_\lambda(s,y,z|g,\lambda_v)=\prod_{n=1}^{T}a_{s_{n-1}s_n}p_\lambda(y_n|s_n,g_n,\lambda_v)p_{\lambda_v}(z_n-y_n).$$

Then, g and λ_v can be estimated from

$$\max_{g,\lambda_v} p_\lambda(z|g,\lambda_v). \quad (14)$$

The gradient equations with respect to g and λ_v are nonlinear. Hence, ML estimation of g and λ_v is performed iteratively using an auxiliary function,

$$Q(\lambda_v,g)=\sum_{n=1}^{T}\sum_{s_n}p_\lambda(s_n|z,g_n)\int p(y_n|g_n,z,s_n)\cdot\log p(z_n,y_n,s_n|g_n,\lambda_v)\,dy_n. \quad (15)$$

Using (14), (15) can be rewritten as

$$\begin{aligned}Q(\lambda_v,g)&=\sum_{n=1}^{T}\sum_{s_n}p_\lambda(s_n|z,g_n)\int p_\lambda(y_n|z,s_n,g_n')\big[\log a_{s_{n-1}s_n}\\&\quad+\log p_\lambda(y_n|s_n,g_n,\lambda_v)+\log p_{\lambda_v}(z_n-y_n)\big]dy_n\\&=\sum_{n=1}^{T}\sum_{s_n}p_\lambda(s_n|z_n,g_n)\Big[\log a_{s_{n-1}s_n}\\&\quad+\mathrm{E}\{\log p(y_n|s_n,g_n)|z,s_n,g_n'\}\\&\quad+\mathrm{E}\{\log p_{\lambda_v}(z_n-y_n)|z,s_n,g_n'\}\Big].\end{aligned} \quad (16)$$

At state $s_n=j$, we will use the notation

$$(\cdot)_j=\mathrm{E}\{\cdot|z,s_n=j,g'\}. \quad (17)$$

From the gradient equations of $Q(\lambda_v,g)$ with respect to g and λ_v, we get the following reestimation formula:

$$g_n^2=\sum_{j=1}^{L}p_\lambda(s_n=j|z,g')\sum_{t=(n-1)N+1}^{nN}\frac{\left(\hat{y}_j(t)-\boldsymbol{B}^j\hat{Y}_j(t-1)\right)^2}{\sigma_j^2}, \quad (18)$$

$$C_n=\left[\sum_{j=1}^{L}p_\lambda(s_n=j|z,g')\sum_{t=(n-1)N+1}^{nN}\hat{Y}_j(t-1)\hat{Y}_j^{\,T}(t-1)\right]^{-1}\cdot\left[\sum_{j=1}^{L}p_\lambda(s_n=j|z,g')\sum_{t=(n-1)N+1}^{nN}\hat{y}_j(t)\hat{Y}_j(t-1)\right], \quad (19)$$

$$g_{v,n}^2=\sum_{j=1}^{L}p_\lambda(s_n=j|z,g')\sum_{t=(n-1)N+1}^{nN}\left(\hat{v}_j(t)-C_n^T\hat{V}_j(t-1)\right)^2, \quad (20)$$

where g′ is the gain contour estimate obtained at the priori iteration and $p_\lambda(s_n=j|z,g')$ can be efficiently calculated using the forward-backward procedure [2]. For initial condition of noise model λ_v at frame n=1, we can estimate λ_v from segments of the signal where no speech is detected and at n>1, we use λ_v obtained from the (n-1)-th frame.

In (18)-(20), $\mathrm{E}\{\cdot|z,s_n=j,g'\}$ can be obtained from the Kalman filter with state $s_n=j$ as [8], following the state-space form:

$$X(t)=\Phi(j)X(t-1)+Gr_j(t), \quad (21)$$

$$z(t)=H^TX(t) \quad (22)$$

where

$X(t)=[y(t)\ldots y(t-p)\,v(t)\ldots v(t-q)]^T$, $r_j(t)=[e_j(t)\ \ w(t)]^T$,

$\Phi(j)=\begin{bmatrix}\Phi_y(j) & \boldsymbol{0}\\ \boldsymbol{0} & \Phi_v\end{bmatrix}$, $G=\begin{bmatrix}g & \boldsymbol{0}\\ \boldsymbol{0} & g_v\end{bmatrix}$, $H^T=\begin{bmatrix}H_y{}^T & H_v{}^T\end{bmatrix}$,

$G_y=H_y{}^T=[10\ldots0]$, $G_v=H_v{}^T=[10\ldots0]$,

$\Phi_y(j)=\begin{bmatrix}B_{1,0}^J+B_{1,1}^Jn & \ldots & B_{p,0}^J+Bp_{1,1}^Jn & 0\\ 0 & & \boldsymbol{I} & 0\end{bmatrix}$,

$\Phi_v=\begin{bmatrix}c_1 & \ldots & c_q & 0\\ 0 & & I & 0\end{bmatrix}$.

The estimator $\hat{X}_j(t)$ on the state $s_n=j$ can be obtained from the conventional Kalman filtering as

$$\hat{X}_j(t)=\Phi(j)\hat{X}_j(t-1)+K_j\cdot\left\{z(t)-H^T\Phi(j)\hat{X}_j(t-1)\right\}, \quad (23)$$

$$M_j(t)=\Phi(j)P_j(t-1)\Phi^T(j)+GQ(j)G^T, \quad (24)$$

$$K_j(t)=M_j(t)H\left[H^TM_j(t)H\right]^{-1}, \quad (25)$$

$$P_j(t)=M_j(t)-K_j(t)HM_j(t). \quad (26)$$

where $Q(j)=E\{r_j(t)r_j{}^T(t)\}$.

We note that $\hat{y}_j{}^2(t)$, $\hat{Y}_j(t-1)\hat{Y}_j{}^T(t-1)$, $\hat{y}_j(t)\hat{Y}_j(t-1)$, $\hat{v}_j{}^2(t)$, $\hat{V}_j(t-1)\hat{V}_j{}^T(t-1)$ and $v_j(t)\hat{V}_j{}^T(t-1)$ of (18)-(20) may similarly be extracted from $\hat{X}_j(t)\hat{X}_j{}^T(t)$.

The algorithm for local maximization of $p_\lambda(z|g,\lambda_v)$ over g and λ_v can be summarized as follows;

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

Step-1: Calculate the posterior probabilities $p_\lambda(s_n|z,g_l)$ for $s_n=1,..,L$ and $n=1,...,T$,

Step-2: Speech enhancement: Calculate the $\hat{X}_j(t)\hat{X}_j{}^T(t)$ by (23)-(26),

Step-3: Estimation of gain factor and noise parameter: estimate g_{l+1} and λ_v using (18)-(20).

Step-4: If $p_\lambda(z|g_{l+1},\lambda_{v,l+1})-p_\lambda(z|g_l,\lambda_{v,l})\le\varepsilon$, assign $\underset{g,\lambda_v}{\mathrm{Max}}\,p_\lambda(z|g,\lambda_v)=p_\lambda(z|g_{l+1},,\lambda_{v,l+1})$ and stop

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

Finally, the decision rule for noisy spoken word is

$$\max_{1\le i\le W} p_{\lambda}\left(z|W_i, g, \lambda_v\right). \tag{27}$$

where W is number of the total word for speech recognition.

V. 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 the AR order of 15. Training and recognition were performed on nonoverlapping vectors of the speech word whose dimension was N=256. In these experiments, the effects of adding Gaussian colored noise were studied. For the noise, we used the car noise. Then, the model is assumed to be AR with the 8-th order.

The gain adapted recognition method with noise estimation was compared to an approach without noise estimation. Recognition results in noise are given in Table 1. From this result, noise estimation of the proposed method is sufficiently good to improve recognition in noise. Up to four iterations of the noise estimation algorithm were used. Table 2 shows a result of comparison on the conventional ARHMM with noise estimation and the proposed method with gain adaptation (in this case, conventional method is equal to proposed method with M=0).

VI. Conclusions

The gain-adapted speech recognition when the noisy signal is available is developed in time domain. The noise is assumed by the colored noise. It uses the NAR-HMM to model the clean speech. The nonstationary AR is modeled by the polynomial function with linear combination of M known basis functions. The gain adapted recognition algorithm with noise estimation are developed for the NAR-HMM using the EM approach and tested in recognition of noise speech. Enhancement is performed by the application of multiple Kalman filters formed from speech and noise estimates to each frame. Simulation results presented for the additive stationary colored noise show the proposed method to be effective.

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

References

[1] 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.

[2] 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.

[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] K.Y. Lee and J. Lee, "A nonstationary autoregressive HMM with gain adaptation for speech recognition," *Proc. ICSLP'98*, Dec. 1998, Australia, *to be published.*

[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.

[8] K.Y. Lee and J. Rheem, "A nonstationary autoregressive HMM and its application to speech enhancement," *Proc. Eurospeech'97*, vol.4, pp. 1407-1411, Sep. 1997, Greece.

Table 1. Recognition results in noise

SNR (dB)	without noise estimation(%)	with noise estimation (%)
0	15	87
5	45	91
10	70	93
15	75	95
20	85	97

Table 2. Recognition results in noise

SNR (dB)	M=0 (%)	M=1 (%)
0	85	87
5	87	89
10	90	93
15	93	95
20	95	97