Hierarchical Non-Stationarity in a Class of Doubly Stochastic Models with Application to Automatic Speech Recognition

L. Deng
Department of Electrical and Computer Engineering
University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 deng@ccng.waterloo.edu

Abstract

In this paper we introduce the concept of two-level (global and local) hierarchical nonstationarity for describing the complex, elastic, and highly dynamic nature of speech signals. A general class of doubly stochastic process models are developed to implement this concept. In this class of models, the global nonstationarity is embodied through an underlying Markov chain (or any other scheme capable of providing nonlinear time warping mechanisms) which governs the evolution of the parameters in a set of output stochastic processes. The local nonstationarity is realized by assuming state-conditioned, time-varying first and second order statistics in the output data-generation process models. To provide practical algorithms for speech recognition which allow the model parameters to be reliably estimated, the local nonstationarity is represented in a parametric form. Simulation results demonstrated close fitting of the model to the actual speech data. Results from speech recognition experiments provided evidence for the effectiveness of the model in comparison with the standard HMM, which is a degenerated case — with single-level nonstationarity — of the proposed model.

I. Introduction

Traditional stochastic models have been developed to deal only with stationary sources, or at best, with nonstationary observations which can be directly transformed into stationary observations by simple time differentiation [2]. Only with the advent of hidden Markov models (HMMs) has it become possible to model nonstationary sources in a reasonably satisfactory manner.

Nonstationary behaviors are exhibited in the HMM via the evolution of the underlying Markov chain. This is a powerful mechanism for representing acoustic signals in natural speech since it parallels patterns of change of the phonetic content contained in the acoustic signal. However, in the HMM setup no mechanism is provided to handle detailed variations in the highly dynamic speech signal given a fixed phonetic content. This is true of all the stochastic models for speech developed so far, including the best known Baum's and Liporace's HMMs [1, 6] and Poritz's hidden filter model [7]. These models all assume the state-conditioned stationarity for the observation data and they rely solely on the (hidden) Markov chain to fit the overall speech nonstationarity.

Acoustic signals in actual speech exhibit a clear nature of hierarchical nonstationarity. At the global level, nonstationarity is exhibited when phonetic contents change over time in a relatively slow fashion. A Markov chain is well equipped to describe such changes. The local nonstationarity, on the other hand, manifests itself generally at the allophonic or at the microstructural level. The effects of such local nonstationarity are especially pronounced in transitional segments of speech whose production involves strong articulatory dynamics (e.g. glides, diphthongs, and CV/VC transitions) [5]. Both the standard HMM and the hidden filter model are a handicap in handling the local nonstationarity.

The purpose of this paper is to propose a general class of stochastic models which are capable of capturing both the global nonstationarity and the local nonstationarity in the speech signal in a parametric form.

II. The Hierarchical Nonstationary Model

The global nonstationarity in this class of models, as with the standard HMM, is implemented by a homogeneous Markov chain. The local or state-conditioned nonstationarity is implemented by an autoregressive output process where both the first-order statistic (mean) and the second-order statistic (autocorrelation function) are made a function of time via time-varying mean function and time-varying autoregression coefficients. We call this model the *hierarchical nonstationary model*, or HN-model for short.

The HN-model consists formally of the following parameter quadraples $[\mathbf{A}, \boldsymbol{\Theta}, \boldsymbol{\Psi}, \boldsymbol{\Sigma}]$:

1. Transition probabilities, $a_{ij}, i, j = 1, 2, ..., N$ of the homogeneous Markov chain with a total of N states;
2. Parameters Θ_i in the time-varying mean functions $g_t(\Theta_i)$ of the output data-generation process, as dependent on state i in the Markov chain;
3. Parameters Ψ_i in the time-varying autoregression coefficients (with a fixed regression order p) $\phi_t(\Psi_i)$ of the output data-generation process, as dependent on state i in the Markov chain;
4. Covariance matrices, Σ_i, of the zero-mean, Gaussian, IID driving noise $R_t(\Sigma_i)$, which are also state dependent.

Given the above model parameters, the observation vector sequences, $O_t, t = p+1, p+2, ..., T$ are generated from the model according to

$$O_t = g_t(\Theta_i) + \Sigma_{k=1}^{p} \phi_t(\Psi_i) O_{t-k} + R_t(\Sigma_i), \qquad (1)$$

where state i at a given time t is determined by the evolution of the Markov chain characterized by a_{ij}.

III. Parameter Estimation for the HN-Model

As with the standard HMM, we use the EM algorithm to obtain an iterative solution to maximum likelihood estimates for the parameters in the HN-model. Each iteration in the EM algorithm consists of two steps. In the E step, the auxiliary function $Q(\Phi|\Phi_0)$ is obtained:

$$Q(\Phi|\Phi_0) = E[log P(O_1^T, \mathcal{S}|\Phi)|O_1^T, \Phi_0], \qquad (2)$$

where the expectation is taken over the "hidden" state sequence $\mathcal{S}$. For the HN-model, algebraic manipulations on (2) lead to the simplified form of Q

$$\begin{aligned} Q &= \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1} P(s_t = i, s_{t+1} = j|O_1^T, \Phi_0) \log a_{ij} \\ &+ \sum_{i=1}^{N}\sum_{t=1}^{T} P(s_t = i|O_1^T, \Phi_0) \mathcal{N}_t(i), \qquad (3) \end{aligned}$$

where $\mathcal{N}_t(i)$ stands for the log likelihood

$$\begin{aligned} -\frac{D}{2} log(2\pi) \quad -\tfrac{1}{2} log|\Sigma_i| - \tfrac{1}{2}[O_t - g_t(\Theta_i) - \Sigma_{k=1}^{p} \phi_t(\Psi_i) O_{t-k}]^{Tr} \\ \Sigma_i^{-1}[O_t - g_t(\Theta_i) - \Sigma_{k=1}^{p} \phi_t(\Psi_i) O_{t-k}]. \qquad (4) \end{aligned}$$

Estimates of the model parameters are obtained in the M step via maximization of (3). Re-estimation formulas for the transition probabilities and for the residual covariance matrices are very similar to those in the standard HMM and are thus omitted. Re-estimation of the parameters in the time-varying mean functions and in the regression coefficients requires solution of a system of equations which is derived below.

By removing optimization-independent terms and factors in (3), an equivalent objective function is obtained as

$$\begin{aligned} Q_e(\Theta_i, \Psi_i) = \sum_{i=1}^{N}\sum_{t=1}^{T} \gamma_t(i)[O_t - g_t(\Theta_i) - \Sigma_{k=1}^{p} \phi_t(\Psi_i) O_{t-k}]^{Tr} \\ \Sigma_i^{-1}[O_t - g_t(\Theta_i) - \Sigma_{k=1}^{p} \phi_t(\Psi_i) O_{t-k}], \qquad (5) \end{aligned}$$

where $\gamma_t(i) = P(s_t = i | O_1^T, \Phi_0)$, which can be computed efficiently by the use of the forward-backward algorithm [1].

The re-estimation formulas are obtained by jointly solving

$$\frac{\partial Q_e}{\partial \Theta_i} = 0; \quad \frac{\partial Q_e}{\partial \Psi_i} = 0, \quad i = 1, 2, ..., N. \tag{6}$$

Using the chain rule for differentiation, (6) becomes

$$\sum_{t=1}^{T} \gamma_t(i)[O_t - g_t(\hat{\Theta}_i) - \Sigma_{k=1}^{p} \phi_t(\hat{\Psi}_i) O_{t-k}] \frac{\partial g_t(\hat{\Theta}_i)}{\partial \hat{\Theta}_i} = 0, \tag{7}$$

and

$$\sum_{t=1}^{T} \sum_{s=1}^{p} \gamma_t(i)[O_t - g_t(\hat{\Theta}_i) - \Sigma_{k=1}^{p} \phi_t(\hat{\Psi}_i) O_{t-k}] O_{t-s}^{Tr} \frac{\partial \phi_t^{Tr}(\hat{\Psi}_i)}{\partial \hat{\Psi}_i} = 0 \tag{8}$$

for $i = 1, 2, ..., N$.

We now let $g_t(\Theta_i)$ and $\phi_t(\Psi_i)$ take specific forms of time function. Polynomial functions are the simplest choices, which convert (7) and (8) to a coupled linear system of equations for solving the polynomial coefficients. That is, let

$$g_t(\Theta_i) = \sum_{m=0}^{M} \mathbf{B}_i(m) t^m, \quad \phi_t(\Psi_i) = \sum_{l=0}^{L} \mathbf{H}_i(l) t^l, \tag{9}$$

where $\mathbf{B}_i(k)$ is a D-dimensional vector and $\mathbf{H}_i(k)$ is a $D \times D$ matrix, both associated with state i in the Markov chain and with polynomial order k. Then the model parameters Θ and Ψ are just two sets of the polynomial coefficients, $b_d(m), m = 0, 1..., M$, and $h_{uv}(l), l = 0, 1, ..., L; d, u, v = 1, 2, ..., D$.

Substituting (9) into (7) and (8), we obtain the linear vector system for the re-estimate of the polynomial coefficients

$$\sum_{t=1}^{T} \gamma_t(i)[\mathbf{O}_t - \sum_{m=0}^{M} \hat{\mathbf{B}}_i(m) t^m - \sum_{k=1}^{p} \sum_{l=0}^{L} \hat{\mathbf{H}}_i(l) t^l O_{t--k}] t^u = 0,$$

for $u = 0, 1, ..., M$, coupled with the linear matrix system

$$\sum_{t=1}^{T} \sum_{s=1}^{p} \gamma_t(i)[\mathbf{O}_t - \sum_{m=0}^{M} \hat{\mathbf{B}}_i(m) t^m - \sum_{k=1}^{p} \sum_{l=0}^{L} \hat{\mathbf{H}}_i(l) t^l O_{t-k}] t^v O_{t-s}^{Tr} = 0,$$

for $v = 0, 1, ..., L$.

We have so far implemented the model which contains only the component of time-varying means, or the above model with $\mathbf{H} = \mathbf{0}$.

IV. Application to Speech Recognition

Preliminary speech recognition experiments have been carried out to evaluate the HN-model described above. In an E-set recognition experiment using words as the speech units [5], we found that with a large amount of training data, the HN-model performs the nearly same as the standard HMM containing a large number of states. However, with only two tokens to train each word model, the standard HMM suffers from the undertraining problem but the HN-model does not. In the second set of experiments, we intend to improve the previously developed locus model [4] by using the quadratically time-varying mean functions in the HN-model to fit formant transition data. The model in [4] requires the linearity assumption on the acoustic transition data but with the HN-model, polynomial fitting to the data of any order can be implemented with mathematical tidiness to reliably estimate the "locus" parameters.

V. Discussion

The statistical properties of the standard HMM, the hidden filter model, and the currently developed HN-model can be compared to appreciate the increasing level of generality in the model development. The standard HMM is a local IID model, which has a very simple statistical structure. It contains locally constant (degenerated) mean functions. The hidden filter model generalizes the standard HMM in just providing time-origin independent (and hence remains a locally stationary model), exponentially decaying functions in the second-order statistics. The first-order statistics remain the same as those in the standard HMM. The HN-model developed in this paper generalizes the above models in providing locally time-varying first-order and second-order statistics, and hence a locally nonstationary model.

The HN-model provides a mechanism for dealing with two levels of nonstationarity in the signal to be modeled: the global nonstationarity as controlled by the Markov chain, and the local nonstationarity as accommodated by the state-conditioned time-varying statistics up to the second order. The first order statistic is important since it catches the general trend for the dynamic movement of speech data over time. In the standard HMM [3], states are intended and can only be used to represent piece-wise stationary segments of speech although the acoustic realization of many types of speech sounds exhibits highly dynamical trends and varies in a truly continuous manner. Such trends can be much more efficiently and accurately described by time-varying mean functions, rather than using many HMM states to approximate the trends piece-wise constantly.

The theoretical significance of the correlation function lies in the fact that any random process, in a second-order theory with which we assume state-conditioned speech data are in conformity, identifies itself with this function. Intuitively, we argue for close relationships between the correlation function $\rho(\tau)$ and speech frame (Y_t) dependence as follows: If $\rho(\tau)$ has large values at τ, then the acoustic data in a speech frame would have strong influences on those in another speech frame which is τ frames away. For instance, suppose $\rho(\tau)$ has a large positive value; then if Y_t is greater than the mean value, $Y_{t+\tau}$ would tend to be forced to move above the mean value so as to keep $\rho(\tau)$ positive.

In view of the close relationship between the properties of the actual speech signal and those provided by the HN-model, and of essentially the same computational complexities between the standard HMM and the HN-model, we predict strong utilities of the HN-model in speech recognition. Our preliminary experiments are consistent with this prediction.

References

[1] L.E. Baum. "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes", *Inequalities*, Vol. 3, pp. 1–8, 1972.

[2] G.E.P. Box and G.M. Jenkins. *Time Series Analysis—Forecasting and Control*, Holden-Day, San Francisco, CA, pp. 67–72, 1976.

[3] L. Deng, P. Kenny, M. Lennig, and P. Mermelstein. "Phonemic hidden Markov models with continuous mixture output densities for large vocabulary word recognition," *IEEE Trans. Signal Processing*, Vol. 39, No. 7, July, 1991, pp. 1677–1681.

[4] L. Deng, P. Kenny, M. Lennig, and P. Mermelstein. "Modeling acoustic transitions in speech by state-interpolation hidden Markov models," *IEEE Trans. Signal Processing*, scheduled to appear in February, 1992.

[5] L. Deng and K. Erler. "Structural Design of HMM Speech Recognizer Using Multi-Valued Phonetic Features: Comparison with Segmental Speech Units," submitted to *J. Acoust. Soc. Am.*.

[6] L.A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Information Theory*, Vol.28, pp. 729-734, 1982.

[7] A.B. Poritz, "Hidden Markov models: A guided tour," *Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing*, New York, New York, April 11–14, 1988, pp. 7–13.