corrected by the additional processing, permitting the time advantage to be achieved with no concomitant erosion of performance.

The amount of redundant computation that can be removed also depends explicitly upon the means for combining states as is also evident in Table II. Ultimately, an experimental procedure will likely be necessary to evaluate different means of compressing models in a given application. This is not surprising since many of the procedures applied to increase the performance and efficacy of HMM's can only be evaluated experimentally. The essence of what we have shown here is that the HMM can be manipulated into a form in which the redundancy is explicit in a certain sense and can be analytically removed up to any desired degree.

IV. Summary and Conclusions

A simple framework has been presented for evaluating the likelihood of a HMM using only $\mathcal{O}([1-\kappa]NT)$ floating point operations, where N is the number of states, T is the number of observations derived from an utterance, and κ, the compression index, is in the range $0 \leq \kappa < 1$. This is a reduction from the $\mathcal{O}(3NT)$ (Bakis form) to $\mathcal{O}(N^2T)$ (ergodic form) operations used in existing algorithms to evaluate the likelihood of an HMM. The procedure is based upon removal of redundant or nearly-redundant computations that take place in a population of HMM's. The achievable value of κ depends on the task, the original model form, the vocabulary, and the technique for consolidating similar computations. It also depends upon whether some recognition errors can be tolerated or corrected for the sake of reduced computation. Accordingly, κ depends on problem-dependent considerations, and must ordinarily be determined experimentally. In the isolated-word experiments reported here, κ ranged from 0.19 to 0.60 with no loss of performance, while κ in the range 0.43 to 0.93 could be achieved with a 10% increase in error rate. The worst case value of κ under any conditions is $1/N$ if the original HMM contains an absorbing state.

References

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE,* vol. 77, pp. 257–285, 1989.

[2] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.,* vol. 3, pp. 4–16, 1986.

[3] J. R. Deller, Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete Time Processing of Speech Signals.* New York: Macmillan, 1993.

[4] C. T. Chen, *Linear Systems Theory and Design.* New York: Holt, Rinehart and Winston, 1984.

[5] G. R. Doddington and T. B. Schalk, "Speech recognition: Turning theory into practice," *IEEE Spectrum,* pp. 26–32, Jan. 1981.

[6] S. A. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, and Signal Process.,* vol. ASSP-28, pp. 357–366, Aug. 1985.

[7] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proc. IEEE,* vol. 73, pp. 1551–1588, 1985.

[8] R. K. Snider, "Efficient discrete symbol HMM evaluation using transformation and state reduction," M.S. dissertation, Michigan State Univ., East Lansing, 1990.

[9] P. A. Divijver and J. Kittler, *Pattern Recognition: A Statistical Approach.* London: Prentice-Hall International, 1982.

A Stochastic Model of Speech Incorporating Hierarchical Nonstationarity

Li Deng

***Abstract*— The concept of two-level (global and local) hierarchical nonstationarity is introduced in this paper to describe the elastic and dynamic nature of the speech signal. A doubly stochastic process model is developed to implement this concept. In this model, the global nonstationarity is embodied through an underlying Markov chain that governs evolution of the parameters in a set of output stochastic processes. The local nonstationarity is realized by utilizing state-conditioned, time-varying first- and second-order statistics in the output data-generation process models. For potential uses in automatic uncovering relationally invariant properties from the speech signal and in speech recognition, the local nonstationarity is represented in a parametric form. Preliminary experiments on fitting the models to speech data demonstrate superior performance of the proposed model to several traditional types of hidden Markov models.**

I. Introduction

Traditional statistical time series models have been developed to deal mainly with stationary sources, or at best, with nonstationary observations that can be directly transformed into stationary observations by time differentiation [2]. Only since the advent of hidden Markov models (HMM's) [1] has it become possible to set up a natural framework in which nonstationary sources can be described in tidy terms of mathematics.

Nonstationary behaviors are exhibited in the HMM via the evolution of the underlying Markov chain that modulates the parameters of an output stochastic process. This is a powerful mechanism for representing acoustic signals in natural speech since it parallels patterns of change of the phonetic content embedded in the acoustic signal. In the standard HMM setup, however, no mechanism is provided to handle detailed variations of intrinsically dynamic speech signals and their temporal relationships given a fixed phonetic content. This has been true of all the stochastic models for speech developed so far, including the best known Baum's HMM and its extension [1], [11], and Poritz's hidden filter model and its extension [12], [9], [13]. These models all assume the state-conditioned stationarity for the observed speech data, and they rely solely on the (hidden) Markov chain to fit the overall speech nonstationarity.

The dynamic nature of the acoustic signals in the speech can be described in terms of statistical nonstationarity that is hierarchically organized at two distinct levels. At the global level, the nonstationarity manifests itself when phonetic contents change over time in a relatively slow fashion. A Markov chain is well equipped to describe this change. The local nonstationarity, on the other hand, is displayed generally at the lower allophonic [7] or microstructural level [6]. Both the standard HMM and the hidden filter model are a handicap in handling the local nonstationarity.

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

Manuscript received September 25, 1991; revised January 26, 1993. The associate editor coordinating the review of this paper and approving it for publication was Dr. C.-H. Lee.

The author is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ont., Canada N2L 3G1.

IEEE Log Number 9210875.

II. The Hierarchically Nonstationary Model

The global nonstationarity in the proposed model, as with the standard HMM, is implemented by a homogeneous Markov chain. The local (i.e., state-conditioned) nonstationarity is represented by a generalized autoregressive output process where both the first-order statistic (mean) and the second-order statistic (autocovariance) are made functions of time via time-varying trend functions and time-varying autoregressive coefficients, respectively. We call this model the *hierarchically nonstationary model,* or HN-model for short.

The HN-model is defined formally by the following data-generation equation:

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

where state i at a given time t is determined by the evolution of a Markov chain, and O_t, $t = 1, 2, \cdots, T$ is the vector sequence of the output observation data. The remaining variables in (1), together with the Markov chain, constitute the parameter quadraples $[A, \Theta, \Psi, \Sigma]$ of the HN-model:

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

The generality of this HN-model can be seen from four special cases: 1) If $g_t(\Theta_i) = \mu_i$ (constant) and $p = 0$, then the model is reduced to the standard Gaussian HMM [1], [11]. 2) If $g_t(\Theta_i) = \mu_i$ and $\phi_t(\Psi_i) = b_i$ (constant), the model is reduced to the hidden filter model or linear-predictive HMM [12], [9]. 3) If $g_t(\Theta_i) = \mu_i$ and $\phi_t(\Psi_i) = +1$, the model is reduced to the dynamic-parameter HMM [8]. 4) If $p = 0$, then the model is reduced to the trended HMM developed more recently [4]. On the other hand, when the time-varying function $g_t(\Theta_i)$ is set to zero and the autoregression coefficients $\phi_t(\Psi_i)$ are assumed constant given the state, then the HN-model would become a less general model than the neural predictive HMM [5], [10]. The latter allows an arbitrary nonlinear function for the autoregression while the HN-model contains only a linear form for the autoregression.

III. Parameter Estimation for the HN-Model

The EM algorithm [3] is used in this study to obtain maximum likelihood estimates for the HN-model's parameters. The estimates are obtained through an iterative procedure, where each iteration consists of two separate steps: E (expectation) step and M (maximization) step. The E step involves evaluation and simplification of the auxiliary function

$$Q(\Phi|\Phi_0) = E[\log P(O_1^T, \mathcal{S}|\Phi) | O_{1-p}^T, \Phi_0] \tag{2}$$

where the expectation is taken over the "hidden" state sequence $\mathcal{S}$, Φ stands for the model whose parameters are to be estimated in the present iteration, and Φ_0 stands for the model whose parameters were estimated in the previous iteration. (For notational convenience we denote the vector sequence $O_{1-p}, O_{2-p}, \cdots, O_1, O_2, \cdots, O_T$ as O_{1-p}^T.)

The Q function in (2) can be explicitly expressed in terms of the expectation over state sequences $\mathcal{S}$ in the form of a weighted sum

$$Q(\Phi|\Phi_0) = \sum_{\mathcal{S}} P(\mathcal{S}|O_{1-p}^T, \Phi_0) \log P(O_1^T, \mathcal{S}|\Phi). \tag{3}$$

Let D be the dimensionality of the observation vector and let $N_t(i)$ stand for log of the multivariate Gaussian likelihood

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

Then the log joint likelihood in (3) can be written as

$$\log P(O_1^T, \mathcal{S}|\Phi) = \sum_{t=1}^{T} N_t(s_t) + \sum_{t=1}^{T-1} \log a_{s_t s_{t+1}}. \tag{4}$$

The final result of the E-step is the following Q function:

$$\begin{aligned} Q = &\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1} P(s_t = i, s_{t+1} = j | O_{1-p}^T, \Phi_0) \log a_{ij} \\ &+ \sum_{i=1}^{N}\sum_{t=1}^{T} P(s_t = i | O_{1-p}^T, \Phi_0) N_t(i) \end{aligned} \tag{5}$$

with detailed steps of derivation similar to those in [1].

Re-estimates of the model parameters are obtained in the M step by maximization of (5) with respect to all model parameters. Here we describe in detail the re-estimation result of the parameters in the time-varying trend functions and in the autoregression coefficients.

By removing optimization-independent terms and factors in (5), an equivalent objective function is obtained as

$$\begin{aligned} Q_e(\Theta_i, \Psi_i) = &\sum_{i=1}^{N}\sum_{t=1}^{T} \gamma_t(i) \Big[O_t - g_t(\Theta_i) \\ &- \sum_{k=1}^{p} \phi_t(\Psi_i) O_{t-k} \Big]^{Tr} \sum_i^{-1} \Big[O_t - g_t(\Theta_i) \\ &- \sum_{k=1}^{p} \phi_t(\Psi_i) O_{t-k} \Big] \end{aligned} \tag{6}$$

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

The re-estimation formulas are obtained by jointly solving

$$\frac{\partial Q_e}{\partial \Theta_i} = 0; \quad \frac{\partial Q_e}{\partial \Psi_i} = 0, i = 1, 2, \cdots, N. \tag{7}$$

After applications of the chain rule for vector differentiation, (7) becomes

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

and

$$\begin{aligned} &\sum_{t=1}^{T}\sum_{s=1}^{p} \gamma_t(i) \left[O_t - g_t(\hat{\Theta}_i) - \sum_{k=1}^{p} \phi_t(\hat{\Psi}_i) O_{t-k} \right] \\ &\quad \cdot O_{t-s}^{Tr} \frac{\partial \phi_t^{Tr}(\hat{\Psi}_i)}{\partial \hat{\Psi}_i} = 0 \end{aligned} \tag{9}$$

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

We now let $g_t(\Theta_i)$ and $\phi_t(\Psi_i)$ take specific forms of the time function. Polynomial functions are the simplest choices, which convert (8) and (9) to a coupled linear system of equations for solving the polynomial coefficients. That is, let

$$g_t(\Theta_i) = \sum_{m=0}^{M} B_i(m)t^m$$
$$\phi_t(\Psi_i) = \sum_{l=0}^{L} \mathbf{H}_i(l)t^l \tag{10}$$

where $B_i(k)$ 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: $b_d(m), m = 0, 1, \cdots, M$, and $h_{uv}(l), l = 0, 1, \cdots, L; d, u, v = 1, 2, \cdots, D$.

Substituting (10) into (8) and (9), we obtain the linear vector system for the re-estimate of the polynomial coefficients

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

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

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

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

Use of the matrix-calculus technique leads to straightforward solutions to the above system of equations. We show below the selected results in the data-fitting experiments based on the above solution.

IV. Experiments on Speech Data Fitting

In this section, we provide experimental evidence to show that the proposed HN-model can fit the actual speech data, both for the training data and for the test data, more closely than the traditional types of HMM's.

We recorded speech data from many tokens of word *deed/di:d/* uttered by a native English male speaker. The recorded speech data was in the form of digitally sampled signal at 16 kHz. A Hamming window of duration 25.6 msec was applied every 10 msec, which converts the data from a time domain to a frequency domain. Within each window, a seven-dimensional vector consisting of mel-frequency cepstral coefficients was computed. To save space we show here the data-fitting results only for the second-order cepstral coefficient C2 (similar results were obtained for other cepstral coefficients).

We trained the HN-model parameters using the EM algorithm described in Section III above. Eight training tokens of word *deed* were used for training a two-state left-to-right HN-model ($N = 2$) (Similar results were obtained for $N > 2$). We first show the results of model fitting to the training data. The dotted lines in Figs. 1(a), (b), (c), and (d) are the speech data (C2 sequence) from one training token to be fitted. The vertical axis represents the magnitude of C2 data and the horizontal axis is the frame number (frame size is 10 msec). Superimposed on the same graphs in Fig. 1 as solid lines are the fitting functions $[g_t(\Theta_i) + \sum_{k=1}^{p} \phi_t(\Psi_i)O_{t-k}]$ in (1) with the residual term (the variation that is regarded purely random by the model) removed. (The parameters in these fitting functions were obtained during training.) Given the model parameters, the process of fitting models to the data proceeded by first finding the optimal segmentation of the data in the HMM states and then fitting the segmented data using the fitting functions associated with the corresponding states. Optimal segmentation of the data was obtained

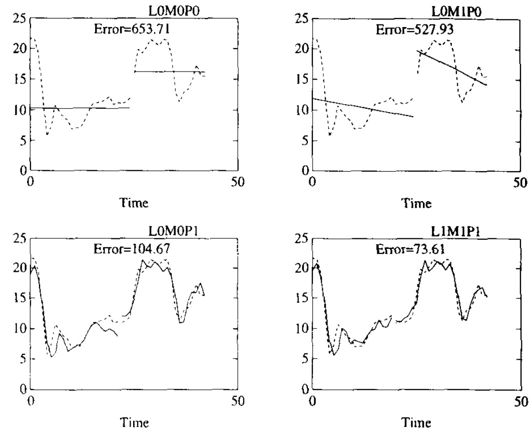

Fig. 1. The dotted lines are the speech data (C2 sequence) from a training token (word *deed*). The vertical axis represents the magnitude of C2 data and the horizontal axis is the frame number (frame size 10 msec). Superimposed on each subgraph as solid lines are the fitting functions $[g_t(\Theta_i)+\sum_{k=1}^{p} \phi_t(\Psi_i)O_{t-k}]$ in (1) with the residual term removed. Given the model parameters, the process of fitting models to the data proceeded by first finding the optimal segmentation of the data into the HMM states and then fitting the segmented data using the fitting functions associated with the corresponding states. The parameters in the fitting functions are (a) $M = 0$ and $p = 0$ (standard Gaussian HMMs); (b) $M = 1$ and $p = 0$ (the first-order trended HMM's [4]); (c) $M = 0$ and $p = 1$ (the first-order hidden filter model); and (d) $M = 1$, $p = 1$, and $L = 1$ (the first-order HN-model. The model fitting errors in (11) are shown as inserts in the graphs.

by the Viterbi algorithm. The solid lines in Figs. 1(a), (b), (c), and (d) show the fitting functions of the standard Gaussian HMM's ($M = 1$ and $p = 0$), the first-order hidden filter model ($M = 0$ and $p = 1$), and the first-order HN-model ($M = 0$ and $p = 0$), the first-order trended HMM's ($M = 1$, $p = 1$, and $L = 1$), respectively. The point in time in each graph where the otherwise continuous line is broken is the frame at which the "optimal" state transition occurs (obtained by use of the Viterbi algorithm). Comparison of closeness of the data fitting using the various types of HMMs shown in Fig. 1 demonstrates clear superiority of the HN-model (Fig. 1(d)).

To be sure that the HN-model's superior data fitting performance is not merely due to its largest number of model parameters (and hence has the highest degree of freedom), we carried out model fitting experiments on test tokens (i.e., word tokens of *deed* not used in training the HN-model). Figs. 2(a), (b), (c), and (d) show these results, again demonstrating the best fitting to the test data using the HN-model. In particular, use of state-conditioned time-varying autoregressive coefficients (Fig. 2(d) for $L = 1$) given significantly closer fit to the data than use of state-conditioned time-invariant autoregressive coefficients (Fig. 2(c) for $L = 0$).

The above qualitative description of the model fitting performance for Figs. 1 and 2 can be quantified by using a measure of fitting error, defined as the temporal summation of the state-dependent frame residual errors:

$$\text{Error} = \sum_{i=1}^{N} \sum_{t=t_{i-1}}^{t_i} \left[O_t - g_t(\Theta_i) - \sum_{k=1}^{p} \phi_t(\Psi_i)O_{t-k} \right]^2 \tag{11}$$

where $t_i, i = 0, 1, 2, \cdots, N$ are the Viterbi segmentation boundaries. These fitting error measures are shown as figure inserts in Figs. 1 and 2 and provide quantitative evidence for superior data fitting performance of the HN-model.

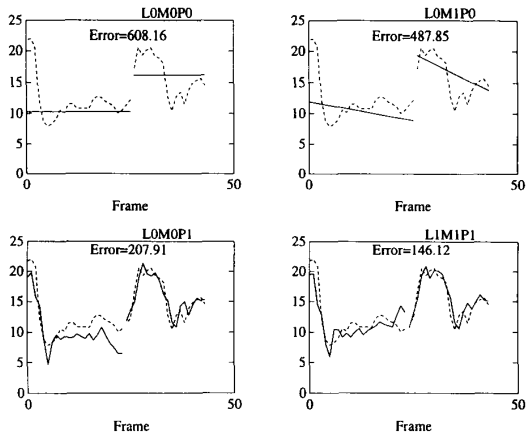

Fig. 2. Modeling fitting to speech data for a test token.

V. Discussion

The statistical properties of the traditional types of HMM's and the proposed HN-model can be compared to examine an increasing level of generality for the development of stochastic models of speech. The standard HMM is a locally IID model, which has a very simple statistical structure. It contains locally constant (degenerated) first-order statistics. Given a state, the second-order statistics (autocovariance functions) of the HMM are zero for any non-zero time lag. The hidden filter model generalizes the standard HMM in providing time-origin independent, exponentially decaying functions in its state-conditioned second-order statistics. Because of time independence both in the trend functions and in the regression coefficients for each state, the hidden filter model remains a locally stationary model. The trended HMM as developed in [4] is a locally nonstationary model, but only for the first-order statistics. The HN-model described in this paper generalizes all the above models in providing state-conditioned time-varying statistics of both the first order and of the second order.

The generality and the associated advantages of the HN-model over the traditional types of HMMs are illustrated in a set of preliminary experiments where several types of models are used to fit speech data. Residual errors for the fitting are shown to be significantly smaller when the HN-model is used than any member in the traditional class of the HMMs. In view of the closeness between the nonstationary properties of the actual speech signal and those provided by the HN-model and of its generality, the HN-model is a potentially more powerful mathematical tool for automatic speech recognition than the types of HMMs currently in use. Furthermore, because the temporal variation of the speech signal is compactly parameterized in the HN-model, a convenient means is provided, via the parameter characterization of the HN-model, to automatically determine the acoustic properties in the speech signal which is relationally invariant in the temporal dimension.

Acknowledgment

The author wishes to thank Mr. C. Rathinavelu for some programming assistance, and Natural Sciences and Engineering Research Council of Canada, Bell-Northern Research, Canada, and Ontario government through the URIF fund for their support of this work.

VII. 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.* San Francisco, CA: Holden-Day, pp. 67–72, 1976.
[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.,* vol. 39, pp. 1–38, 1977.
[4] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Processing,* vol. 27, Apr. 1992, pp. 65–78.
[5] L. Deng, K. Hassanein, and M. Elmasry, "Neural-network architecture for linear and nonlinear predictive hidden Markov models: Application to speech recognition," in B. H. Juang, S. Y. Kung, and C. A. Kamm, eds., *Neural Networks for Signal Processing.* Princeton, NJ: IEEE Catalog No. 91TH0385, 1991, pp. 411–421.
[6] L. Deng and K. Erler, "Microstructural speech units and their HMM representation for discrete utterance speech recognition," in *Proc. ICASSP,* May 1991, pp. 193–196.
[7] L. Deng, M. Lennig, F. Seitz and P. Mermelstein, "Large vocabulary word recognition using context-dependent allophonic hidden Markov models," *Comp. Speech and Lang.,* vol. 4, 1990, pp. 345–357.
[8] V. Gupta, M. Lennig, and P. Mermelstein, "Integration of acoustic information in a large vocabulary word recognizer," in *Proc. ICASSP,* Apr. 1987, pp. 697–700.
[9] P. Kenny, M. Lennig, and P. Mermelstein, "A linear predictive HMM for vector-valued observations with applications to speech recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing,* vol. 38, Feb. 1990, pp. 220–225.
[10] K. Iso and T. Watanabe, "Speaker-independent word recognition using a neural prediction model," in *Proc. ICASSP* (Albuquerque, NM), Apr. 1990, pp. 441–444.
[11] L. A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Information Theory,* vol. 28, pp. 729–734, 1982.
[12] A. B. Poritz, "Hidden Markov models: A guided tour," in *Proc. ICASSP* (New York, NY, Apr. 1988, pp. 7–13.
[13] H. Sheikhzadeh and L. Deng, "Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization," *IEEE Trans. Signal Processing,* accepted.

Correction to "Transform Representation of the Spectra of Acoustic Speech Segments with Applications—II: Speech Analysis, Synthesis, and Coding"

V. Ralph Algazi, David H. Irvine, Kathy L. Brown, Christie L. Cadwell, Michael Ready, and Sang Chung

On page 277 of the above paper,[1] and in the issue's Table of Contents, the authors' names were not shown in the proper order. The ordering shown above is correct.

Manuscript received September 27, 1993.

[1] V. R. Algazi *et al., IEEE Trans. Speech and Audio Processing,* vol. 1, no. 3, July. 1993.