# Hidden Markov Models with First-Order Equalization for Noisy Speech Recognition

Biing-Hwang Juang, *Fellow, IEEE*, and Kuldip K. Paliwal, *Member, IEEE*

*Abstract*—**Speech recognizers often experience serious performance degradation when deployed in an unknown acoustic (particularly, noise contaminated) environment. To combat this problem, we proposed in a previous study a family of new distortion measures that were shown to be able to withstand additive white noise without requiring 1) explicit knowledge of the noise, 2) noise reduction provisions, or 3) reference template retraining. One particularly effective distortion measure in the family is the one that takes into account the norm shrinkage bias in the noisy cepstrum. In this paper, we incorporate a first-order equalization mechanism, specifically aiming at avoiding the norm shrinkage problem, in a hidden Markov model (HMM) framework to model the speech cepstral sequence. Such a modeling technique requires special care as the formulation inevitably involves parameter estimation from a set of data with singular dispersion. We provide solutions to this HMM stochastic modeling problem and give algorithms for estimating the necessary model parameters. We experimentally show that incorporation of the first-order mean equalization model makes the HMM-based speech recognizer robust to noise. With respect to a conventional HMM recognizer, this leads to an improvement in recognition performance which is equivalent to about 15–20 dB gain in signal-to-noise ratio.**

## I. Introduction

SIGNAL observations or measurements often contain undesirable but unavoidable noisy components which make speech recognition task difficult. A speech recognizer designed or trained under clean or low noise conditions generally suffers serious performance degradation when used in an environment with different noise characteristics. One way to handle the noise problem is to include noise during training of the signal patterns in the recognizer [2]. This requires the effort of collecting the noise samples in the intended environment(s) or equipping the recognizer with a mechanism for on-line training. These are impossible to accomplish for a public telephone network service because of the high degree of variability of the talker's environment. Another way to reduce the performance degradation due to noise is to suppress the noise component in the speech signal before it is compared with the existing reference patterns in the recognizer. Well-known procedures of this type include noise subtraction and the iterative enhancement method [3]. These methods gave good results in some limited conditions. One of the drawbacks, however, is that these methods incur a significant increase in computational requirements as extra signal processing steps need to be performed.

Manuscript received March 9, 1991; revised November 25, 1991.
The authors are with AT&T Bell Laboratories, Murray Hill, NJ 07974.
IEEE Log Number 9201581.

The concept we presented in [1] as well as here is entirely different. We concentrate on the possibility of measuring and modeling features of speech that are robust to noise contamination. If this is possible, there will be no need to create noisy reference patterns or to process the signal before recognition. The fact that humans are capable of accurately recognizing spoken words in a noisy environment without requiring extensive adaptation can be viewed as an existence proof that such a process might be reasonable.

In [1], it was shown that an unconventional use of cepstral projection distortion measures led to robust recognition performance without the need for explicit noise characterization or noisy signal prototypes. This fulfills, in part, the requirement that a recognizer be designed and trained in a fixed, noiseless environment and perform properly in a noisy environment without any modifications to either the signal or the recognizer.

The results in [1] relied on some interesting characteristics of the cepstrum of unity gain autoregressive models commonly used in speech modeling. It was shown that the presence of additive white noise causes a reduction in the cepstral norm (or cepstral energy) of the noisy observation vector relative to the one derived from a clean signal. This observation, when cast in the perspective of a Euclidean vector space, explains why traditional speech recognizers inevitably suffer performance degradation under mismatched noise conditions. (By mismatch we mean the noise conditions during training and testing are different.) While it has been demonstrated in [1] that this mismatch problem can be effectively remedied with a first-order norm equalization scheme in a deterministic signal representation setup, a more general equalization mechanism based on hidden Markov models is obviously of interest. In this paper, we address the problem of generalizing the equalization scheme in a stochastic modeling framework.

This paper is organized as follows. In the next section, we review the norm shrinkage model for the LPC cepstrum of speech signals, together with the associated problems, such as centroid calculation required in clustering procedures. The section summarizes the use of first-order equalization in a nonstochastic modeling frame-

work. In Section III, we give a general multivariate formulation for the shrinkage/equalization model. We point out that such a model requires an explicit treatment of the singularity problem which had not occurred previously in related applications of the modeling technique. We then present the solution to the problem of HMM estimation in Section IV. In Section V, we report the experimental results and show the effectiveness of this model for recognizing noisy speech. Section VI summarizes the paper.

## II. Norm Shrinkage and Equalization Model

Let $\boldsymbol{x}^t = [x_0 \;\; x_1 \;\; \cdots \;\; x_{L-1}]$ be a vector (sequence) of speech waveform samples which are said to be generated by an autoregressive source satisfying the following relationship:

$$x_n = \sum_{i=1}^{p} a_i x_{n-i} + e_n \tag{1}$$

where $e_n$ is a Gaussian i.i.d. driving sequence. The transfer function of the system is $1/A(z)$ where $A(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p}$, the well-known LPC inverse filter polynomial. The LPC cepstrum $c_i$ of the waveform $\boldsymbol{x}$ is related to the inverse filter polynomial through the expression

$$-\ln A(z) = \sum_{i=1}^{\infty} c_i z^{-i}. \tag{2}$$

In speech processing, we usually work with a truncated cepstral sequence $\boldsymbol{c}^t = [c_1 \;\; c_2 \;\; \cdots \;\; c_k]$ where $k$ is in the range of 10 to 20 due to the rapid decay of the cepstral coefficients.

Assume the observed speech waveform contains an unknown amount of white noise $\nu$; i.e.,

$$y_n = x_n + \nu_n. \tag{3}$$

Furthermore, we assume *the signal analysis/modeling mechanism in the recognizer is to remain unchanged regardless of the noise contamination.* Then, the observed signal $y_n$ is again modeled by the equation

$$y_n = \sum_{i=1}^{p} a_i' y_{n-i} + w_n \tag{4}$$

where $w_n$ is an uncorrelated sequence. A truncated noisy cepstral vector $\boldsymbol{c}'$ can similarly be defined as

$$-\ln A'(z) = \sum_{i=1}^{\infty} c_i' z^{-i}$$

$$A'(z) = 1 + a_1' z^{-1} + a_2' z^{-2} + \cdots + a_p' z^{-p} \tag{5}$$

and

$$(\boldsymbol{c}')^t = [c_1' \;\; c_2' \;\; \cdots \;\; c_k'].$$

It has been shown [1], both empirically and theoretically, that the presence of white noise $\nu$ causes reduction in the cepstral norm (or energy) defined as

$$|\boldsymbol{c}| = \left( \sum_{i=1} c_i^2 \right)^{1/2}. \tag{6}$$

That is, in general,

$$|\boldsymbol{c}'| \leq |\boldsymbol{c}| \tag{7}$$

can be observed from actual calculations of the LPC cepstrum and can be verified theoretically through the use of noisy signal models [1].

This observation of cepstral norm shrinkage led to the use of an equalization procedure in [1] in order to improve speech recognition performance. It was shown [1] that a tremendous gain in recognition performance can be obtained if the calculation of the Euclidean distance involves an equalizing scalar $\theta$. In particular, if $\boldsymbol{c}$ and $\eta$ denote the testing cepstral vector (with an unknown amount of noise) and the clean reference cepstral vector, respectively, the revised distance is defined by

$$d(\boldsymbol{c}, \eta) = (\boldsymbol{c} - \theta\eta)^t(\boldsymbol{c} - \theta\eta) \tag{8}$$

where

$$\theta = \frac{\boldsymbol{c}^t\eta}{|\eta|^2}. \tag{9}$$

Note that (8) can be rewritten as

$$d(\boldsymbol{c}, \eta) = |\boldsymbol{c}|^2(1 - \cos^2 \beta) \tag{10}$$

where

$$\cos \beta = \frac{\boldsymbol{c}^t\eta}{|\boldsymbol{c}|\,|\eta|}. \tag{11}$$

In one recognition test, this revised distance led to an improvement of 23% (from 60% to 83%) in recognition accuracy at 10-dB global signal-to-noise ratio, without knowing the actual noise level or requiring creation of noisy references.

The results in [1] indicate clearly the pragmatic advantage of the simple first-order equalization model. Although it is not derived from a theoretical point of view, its effectiveness in improving the recognizer performance under unknown noise level conditions validates its usage, at least practically. We will discuss this in more detail later in the next section.

One of the problems associated with incorporation of an equalizing factor in the distance calculation is the classical projected centroid problem. Given a set of $M$ (noisy) cepstral vectors $\{\boldsymbol{c}_i\}_{i=1}^{M}$, the centroid problem requires calculation of a vector $\eta$ that minimizes the accumulated distance defined, with the equalizing factor in the current case, as

$$D = \sum_{i=1}^{M} (\boldsymbol{c}_i - \theta_i\eta)^t(\boldsymbol{c}_i - \theta_i\eta) \tag{12}$$

where $\theta_i$ is defined in (9) for each $\boldsymbol{c}_i$. The alternative expression of (10), in which the unknown vector $\eta$ appears in a normalized form, simplifies the solution. Equation (12) can be rewritten as

$$D = \sum_{i=1}^{M} |\boldsymbol{c}_i|^2 - \sum_{i=1}^{M} (\boldsymbol{c}^t\eta_1)^2 \tag{13}$$

where $\eta_1 = \eta/|\eta|$. With the constraint that $|\eta_1| = 1$, it is straightforward to show that the solution $\eta_1$ satisfies

$$\xi\eta_1 = 2\Xi\eta_1 \tag{14}$$

where

$$\Xi = \sum_{i=1}^{M} c_i c_i^t \tag{15}$$

the sample covariance, and therefore we choose $\eta_1$ as the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix to minimize the accumulated distance of (13). This centroid calculation procedure is a very important one in the following statistical formulation of a first-order equalization model of a noisy LPC cepstrum.

The distance measure of (8) can be generalized to a multivariate probabilistic formulation with the following probability density function (pdf):

$$f(c) = K \cdot \exp\left[-\tfrac{1}{2}(c - \theta\eta)^t W(c - \theta\eta)\right] \tag{16}$$

where $K$ is the normalization factor, and $\theta$ is, similar to (9),

$$\theta = \frac{\eta^t W c}{\eta^t W \eta}. \tag{17}$$

The presence of the equalizing factor $\theta$ makes $f(c)$ in (16) an unusual density since $(I - [\eta\eta^t W/\eta^t W\eta])$ is a projection operator. For parameter estimation, this means there will not be a data support of full dimensionality $(= k)$ and the rank $\gamma(W) < k$. Note that if $W\eta = \mathbf{0}$, (16) becomes

$$f(c) = K \cdot \exp\left[-\tfrac{1}{2} c^t W c\right]. \tag{18}$$

We shall discuss this unusual formulation in the next section where solution procedures for HMM estimation are elaborated.

## III. Hidden Markov Model with Norm Equalization

A hidden Markov model (HMM) $\lambda$ is a triple $\lambda = (\pi, A, F)$, where $\pi$ is the initial state probability vector, $A$ denotes the state transition probability matrix, and $F$ is a set of observation probability densities. The probability vector $\pi$ and matrix $A$ describe an $N$-state Markov chain while the pdf set $F = \{f_i\}_{i=1}^{N}$ characterizes the distributions of the observation in each Markovian state. We summarize the modeling framework briefly in the following. For detailed descriptions of the HMM methodology, consult [5].

The density function defined by $\lambda$ for a sequence of observations, $\{c_t\}_{t=1}^{T} = (c_1, c_2, \cdots, c_T)$, is

$$P_\lambda(c_1, c_2, \cdots, c_T) = \sum_s \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} f_{s_t}(c_t) \tag{19}$$

where $a_{ij}$ are the elements of $A$, $A = [a_{ij}]_{i,j=1}^{N}$, and $\pi_i$ are the elements of $\pi$, $\pi^t = [\pi, \pi_2, \cdots, \pi_N]$. Quantity $a_{ij}$ is thus the probability of making a transition to state $j$ given that the current state is $i$ and $\pi_i$ is the probability of staying at state $i$ before the initial observation. The equalization mechanism is implemented in some specific forms in the observation density $f_i$. This will be the focus of discussion in this section.

### A. Reestimation Algorithm

Baum's reestimation algorithm [6] is an iterative maximization algorithm in which the model parameters $\lambda$, starting from an initial estimate, are iteratively improved upon in the sense of increasing likelihood. Each iteration involves the following two steps:

1) Determine the auxiliary function from the existing model; the auxiliary function is defined as a function of a new (to be found) model $\lambda'$:

$$Q(\lambda, \lambda') = \sum_s P_\lambda(\{c_t\}_{t=1}^{T}, s) \log P_{\lambda'}(\{c_t\}_{t=1}^{T}, s) \tag{20}$$

where $s$ is a state sequence, $s = (s_0, s_1, \cdots, s_T)$, and the summation is over all possible state sequences.

2) Choose a new model $\bar{\lambda}$ to maximize $Q(\lambda, \lambda')$ as a function of $\lambda'$.

During the next iteration, the new model $\bar{\lambda}$ is used in place of the old model $\lambda$ and the two steps repeat again. It can be shown [6] that each iteration guarantees an increase in likelihood, i.e.,

$$P_\lambda(\{c_t\}_{t=1}^{T}) < P_{\bar{\lambda}}(\{c_t\}_{t=1}^{T}).$$

The algorithm stops when it reaches a fixed point solution or when the increase in likelihood falls below a prescribed level.

### B. Maximization of $Q(\lambda, \lambda')$

The auxiliary function is defined by (20). The logarithm allows breakdown of individual groups of parameters; specifically,

$$Q(\lambda, \lambda') = \sum_s P_\lambda(\{c_t\}_{t=1}^{T}, s) \cdot \left\{\log \pi'_{s_0} + \sum_{t=1}^{T} \log a'_{s_{t-1}s_t} + \sum_{t=1}^{T} \log f'_{s_t}(c_t)\right\}. \tag{21}$$

Maximization of $Q(\lambda, \lambda')$ over parameters $\pi$ and $A$ thus remain identical to the previous results that have been well studied [5]. Maximization of $Q(\lambda, \lambda')$ over $F = \{f_i\}_{i=1}^{N}$, on the other hand, remains the focus of this paper. Note the following decomposition:

$$\begin{aligned}\sum_s P_\lambda(\{c_t\}_{t=1}^{T}, s) \sum_{t=1}^{T} \log f'_{s_t}(c_t) &= \sum_{i=1}^{N} \sum_{t=1}^{T} P_\lambda(\{c_t\}_{t=1}^{T}, s_t = i) \log f'_i(c_t) \\ &= \sum_{i=1}^{N} Q_f(\lambda, f'_i)\end{aligned} \tag{22}$$

where $\boldsymbol{f}_i'$ denotes the parameter vector for $f_i'(\ )$. It allows, again, separate optimization of each individual density function $f_i(\ )$ and the solution satisfies

$$\nabla_{f_i'} Q_f(\lambda, \boldsymbol{f}_i')|_{f_i'=\bar{f}_i} = 0 \tag{23}$$

where

$$Q_f(\lambda, \boldsymbol{f}_i') = \sum_{t=1}^{T} P_\lambda(\{\boldsymbol{c}_t\}_{t=1}^T, s_t = i) \log f_i'(\boldsymbol{c}_t). \tag{24}$$

### C. Observation Density with First-Order Equalization

The proposed equalization as suggested in (8) can be incorporated in the observation density in an HMM framework in the following two ways: fixed dispersion and singularized dispersion. Fixed dispersion is a direct extension of (8), where the norm does not involve a weighting matrix. Singularized dispersion, on the other hand, aims at finding a direction in which the projection (which makes the dispersion matrix singular) of observation vectors results in a minimum average distance.

*1) Fixed Dispersion:* The particular form of observation probability densities we consider in this class is

$$f_i(\boldsymbol{c}) = (2\pi)^{-k/2} |\Sigma^{-1}|^{1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{c} - \theta\eta)^t \Sigma^{-1} (\boldsymbol{c} - \theta\eta)\} \tag{25}$$

where $\Sigma$ is positive-definite and fixed. When $\Sigma = I$, the identity matrix, this form of density function becomes identical to what (8) has implied. The fixed dispersion matrix can thus be considered a weighting matrix in the vector space. Although the following development assumes a general $\Sigma$, it is convenient to just use $\Sigma = I$ (or $\Sigma = \sigma^2 I$) unless strong *a priori* knowledge needs to be incorporated. Since

$$\log f_i'(\boldsymbol{c}_t) = -\tfrac{1}{2}[k \log 2\pi + \log |\Sigma_i| + (\boldsymbol{c}_t - \theta_{ti}\eta_i')^t \Sigma_i^{-1} (\boldsymbol{c}_t - \theta_{ti}\eta_i')] \tag{26}$$

where $\theta_{ti}$ is defined in (17). Maximization of the auxiliary function defined in (21) with respect to $\eta_i'$ becomes the typical problem outlined in (12)–(15). The factorization

$$\Sigma_i^{-1} = U_i U_i^t \tag{27}$$

facilitates the transformations

$$\boldsymbol{y}_t = U_i^t \boldsymbol{c}_t \tag{28}$$

and

$$\zeta_i = U_i^t \eta_i / |U_i^t \eta_i|. \tag{29}$$

The optimization objective becomes minimization of

$$\begin{aligned}\Omega &= \sum_{t=1}^{T} P_\lambda(\{\boldsymbol{c}_t\}_{t=1}^T, s_t = i)\{(\boldsymbol{c}_t - \theta_{ti}\eta_i')^t \Sigma_i^{-1} (\boldsymbol{c}_t - \theta_{ti}\eta_i')\} \\ &= \sum_{t=1}^{T} P_\lambda(\{\boldsymbol{c}_t\}_{t=1}^T, s_t = i)\left\{\boldsymbol{c}_t^t \Sigma_i^{-1} \boldsymbol{c}_t - \frac{[(\eta_i')^t \Sigma_i^{-1} \boldsymbol{c}_t]^2}{(\eta_i')^t \Sigma_i^{-1} \eta_i'}\right\} \\ &= \sum_{t=1}^{T} P_\lambda(\{\boldsymbol{c}_t\}_{t=1}^T, s_t = i)\{\boldsymbol{c}_t^t \Sigma_i^{-1} \boldsymbol{c}_t - [(\zeta_i')^t \boldsymbol{y}_t]^2\}\end{aligned} \tag{30}$$

subject to $|\zeta_i'| = 1$. The objective now is identical to (13)–(15) and the solution $\bar{\zeta}_i$ is thus the eigenvector satisfying

$$\xi_{\max} \bar{\zeta}_i = 2\Xi \bar{\zeta}_i \tag{31}$$

where

$$\Xi = \sum_{t=1}^{T} P_\lambda(\{\boldsymbol{c}_t\}_{t=1}^T, s_t = i)\, \boldsymbol{y}_t \boldsymbol{y}_t^t \tag{32}$$

and $\xi_{\max}$ is the maximal eigenvalue of $\Xi$. The difference between (32) and (15) is the weighting factor due to $P_\lambda(\{\boldsymbol{c}_t\}_{t=1}^T, s_t = i)$.

*2) Singularized Dispersion:* There are two parameter categories involved in the estimation of distributions with the form of (25), $\eta_i'$ and $\Sigma_i'$. In the above fixed dispersion approach, $\Sigma_i' = \Sigma_i$ is fixed. These two categories are related not only because they appear simultaneously in the density function, but because the equalizing factor $\theta_{ti}$ is chosen to maximize the exponent in (25). Estimation of these two parameter categories thus has to take the equalizing factor into account.

The equalizing factor results in the following:

$$\boldsymbol{c}_t = \theta_{ti}\eta_i = \left(I - \frac{\eta_i \eta_i^t \Sigma_i^{-1}}{\eta_i^t \Sigma_i^{-1} \eta_i}\right)\boldsymbol{c}_t \tag{33}$$

where the projector $(I - [\eta_i \eta_i^t \Sigma_i^{-1} / \eta_i^t \Sigma_i^{-1} \eta_i])$ projects a vector onto a hyperplane perpendicular to $\eta_i$. The projection operator results in singularized dispersion and thus necessitates a particular treatment based on the theory of singular distribution. A theory of singular Gaussian distributions has been well developed by Khatri [4]. We summarize his results relevant to our discussion in the Appendix for clarity.

If $\Sigma^{-1}$ in (33) is singularized along a particular direction and $\eta_i$ is chosen to be in that direction, the density function collapses to (18). Therefore, norm equalization can be embedded in the singularization process. This naturally leads to the use of the following singular multivariate density:

$$f_i(\boldsymbol{c}_t) = (2\pi)^{-k/2} |\Sigma_i^-|^{1/2} \exp\{-\tfrac{1}{2} \boldsymbol{c}_t^t \Sigma_i^- \boldsymbol{c}_t\} \tag{34}$$

where $\Sigma_i^-$ is a general inverse of the singularized covariance matrix (with dimensionality reduced to $k - 1$). Let $\rho_1, \rho_2, \cdots, \rho_k$ be the eigenvectors of $\Sigma_i^{-1}$ with corresponding eigenvalues $\xi_1, \xi_2, \cdots, \xi_k$. Also assume that $\xi_k = \min_i \xi_i$. The singularized inverse covariance matrix is then chosen as

$$\Sigma_i^- = V \Lambda V^t \tag{35}$$

where

$$V = [\rho_1, \rho_2, \cdots, \rho_{k-1}] \tag{36}$$

and

$$\Lambda = \begin{bmatrix} \xi_1 & & 0 \\ & \ddots & \\ 0 & & \xi_{k-1} \end{bmatrix}. \tag{37}$$

Note that Khatri's results of (A6)–(A8) become directly applicable in this case. The reestimation transformation for the covariance matrix, therefore, involves two steps:

1) Compute $\bar{\Sigma}_i$ as

$$\bar{\Sigma}_i = \frac{\sum_{t=1}^{T} P_\lambda(\{c_t\}_{t=1}^{T}, s_t = i) c_t c_t^t}{\sum_{t=1}^{T} P_\lambda(\{c_t\}_{t=1}^{T}, s_t = i)}. \quad (38)$$

2) Singularize $\bar{\Sigma}_i^{-1}$ to $\bar{\Sigma}_i^-$ according to (35)–(37).

Note that the direction of $\bar{\eta}_i$ is embedded in the singularization process of (35)–(37). This is similar to the fixed dispersion case (31), (32) where the direction of the mean vector is chosen to coincide with the eigenvector corresponding to the largest eigenvalue of the weighted covariance matrix. Here, we express it in terms of the covariance inverse, therefore it coincides with the eigenvector that corresponds to the smallest eigenvalue. The singularization process, after the direction of mean vector is found, produces the transformation matrix $V$ that leads to the density of (34). The key here is $\bar{\Sigma}_i^- \bar{\eta}_i = 0$.

## IV. Speech Recognition Experiments and Results

In the preceding section, we have presented a methodology and an extension of the LPC cepstral norm shrinkage model to a stochastical modeling framework. To cope with the observation norm shrinkage bias problem, a first-order equalization mechanism is introduced in the stochastic framework to fully take advantage of the consistency that a hidden Markov model is able to offer. In order to see whether the current extension of hidden Markov models is able to achieve better results due to the implied consistency in the parameter estimate, we conduct here speech recognition experiments where we study HMM framework with and without the cepstral norm shrinkage model for the recognition of noisy speech. Results of these experiments are described in this section.

In these experiments, an HMM-based speech recognizer is used for the recognition of isolated words. Here, the HMM for each word has five states. Transitions between states are allowed only in left-to-right direction with no skipping of states. Single multivariate Gaussian functions are used to characterize the probability density functions of cepstral vectors in different states. The Viterbi algorithm is used for training as well as for testing the recognizer.

We use here the HMM-based speech recognizer in multispeaker mode and study it for the following two vocabularies: 1) a vocabulary of 10 English digits (0–9), and 2) a vocabulary of 39 English alpha-digits (26 alphabets (A–Z) + 10 digits (0–9) + 3 command words "stop," "error," and "repeat"). The data base consists of speech from 4 talkers (2 males and 2 females). Twenty-four utterances of each word from these 4 talkers were used for training and an additional 40 utterances for testing. The training and testing utterances were recorded over the local dialed-up telephone lines, and digitized at a sampling rate of 6.67 kHz. An eighth-order LPC analysis was performed every 15 ms with a frame width of 45 ms using the autocorrelation method (with Hamming window and no preemphasis), and each frame was represented in terms of 12 cepstral coefficients [7]. Endpoints of each utterance were manually determined.

In order to show the effect of cepstral norm shrinkage for the recognition of noisy speech, we study the following four configurations of the HMM-based speech recognizer:

*Configuration 1:* The recognizer does not incorporate the cepstral norm shrinkage model; i.e., it is a conventional HMM-based speech recognizer.

*Configuration 2:* The HMM-based speech recognizer uses the first-order norm equalization model with fixed dispersion, where the identity matrix is used for the fixed dispersion matrix; i.e., $\Sigma = I$.

*Configuration 3:* The HMM-based speech recognizer uses the first-order norm equalization model with fixed dispersion, where the covariance matrix obtained from the training process in configuration 1 is used for the fixed dispersion matrix.

*Configuration 4:* The HMM-based speech recognizer uses the first-order norm equalization model with singularized dispersion as described in Section IV-C2.

Speech recognition experiments were performed with each of the four configurations for noisy speech at different signal-to-noise ratios (SNR's). Machine-generated, zero-mean, white Gaussian noise was added to each test utterance to get the desired SNR. Recognition results for noisy speech at eight different SNR's ($\infty$, 35, 30, 25, 20, 15, 10, and 5 dB) are shown in Table I for the 10-word English digit vocabulary. Here, SNR = $\infty$ means that no noise is added to the test utterance.

The recognition results pertaining to configuration 1 clearly demonstrate the degree of degradation in recognition performance caused by the additive noise. The recognizer performed perfectly with 100% recognition accuracy for "clean" speech, but could achieve only about 42% recognition accuracy for noisy speech at 15-dB SNR. In configuration 2, a simple norm equalization model is incorporated without sophisticated dispersion modeling. The recognition results show an increased resistance to noise when this simple equalization model is employed, but the recognizer suffers considerable degradation in "clean" condition. With more elaborate norm equalization modeling as in configurations 3 and 4, the recognizer was able to maintain a satisfactory recognition performance for noisy speech with wide range of SNR values. For example, at 15-dB SNR, the recognizer still could achieve about 90% recognition accuracy which corresponds to the performance of the conventional recognizer of configuration 1 at SNR of 30–35 dB. An equivalent SNR improvement of 15–20 dB is thus achieved. Similar results are obtained for the 39-word English alpha-digit vocabulary as shown in Table II.

As mentioned earlier, the cepstral norm shrinkage

TABLE I
SPEECH RECOGNITION PERFORMANCE FOR 10-WORD ENGLISH DIGIT VOCABULARY AS A FUNCTION OF SNR (WITH WHITE NOISE)

| SNR (dB) | Recognition Accuracy (%) | | | |
|---|---|---|---|---|
| | Configuration 1 | Configuration 2 | Configuration 3 | Configuration 4 |
| ∞ | 100.00 | 85.00 | 98.75 | 98.25 |
| 35 | 93.75 | 83.50 | 98.75 | 98.25 |
| 30 | 88.50 | 81.50 | 98.50 | 98.00 |
| 25 | 79.25 | 78.50 | 97.75 | 97.50 |
| 20 | 60.00 | 73.25 | 95.50 | 96.00 |
| 15 | 41.75 | 66.75 | 89.50 | 90.50 |
| 10 | 30.50 | 57.50 | 73.50 | 77.25 |
| 5 | 17.75 | 38.50 | 58.75 | 62.25 |

TABLE II
SPEECH RECOGNITION PERFORMANCE FOR 39-WORD ENGLISH ALPHA-DIGIT VOCABULARY AS A FUNCTION OF SNR (WITH WHITE NOISE)

| SNR (dB) | Recognition Accuracy (%) | | | |
|---|---|---|---|---|
| | Configuration 1 | Configuration 2 | Configuration 3 | Configuration 4 |
| ∞ | 88.78 | 50.13 | 86.28 | 86.06 |
| 35 | 72.76 | 46.41 | 84.94 | 84.55 |
| 30 | 63.53 | 45.71 | 83.33 | 82.18 |
| 25 | 56.15 | 42.18 | 80.60 | 78.97 |
| 20 | 40.90 | 37.95 | 70.08 | 72.31 |
| 15 | 26.54 | 32.37 | 62.50 | 70.71 |
| 10 | 15.83 | 25.51 | 48.08 | 45.19 |
| 5 | 10.32 | 17.18 | 32.24 | 30.26 |

TABLE III
SPEECH RECOGNITION PERFORMANCE FOR 10-WORD ENGLISH DIGIT VOCABULARY AS A FUNCTION OF SNR (WITH COLORED NOISE)

| SNR (dB) | Recognition Accuracy (%) | | | |
|---|---|---|---|---|
| | Configuration 1 | Configuration 2 | Configuration 3 | Configuration 4 |
| ∞ | 100.00 | 85.00 | 98.75 | 98.25 |
| 35 | 98.25 | 82.50 | 98.75 | 98.25 |
| 30 | 98.25 | 80.00 | 98.75 | 98.25 |
| 25 | 90.50 | 76.25 | 98.50 | 98.25 |
| 20 | 83.25 | 69.75 | 98.00 | 98.00 |
| 15 | 70.00 | 60.50 | 94.75 | 93.00 |
| 10 | 47.00 | 48.50 | 85.00 | 83.25 |
| 5 | 28.75 | 37.75 | 69.25 | 66.50 |

model used in the present paper assumes the additive noise to be white. However, it will be of interest to see how this method works for colored noise. For this, we performed the recognition experiments at different SNR's with additive colored Gaussian noise. Colored Gaussian noise was generated by filtering the white Gaussian noise by a second-order AR filter (with coefficients $-0.8018$ and 0.3995) [8]. Results for the 10-word English digit vocabulary are shown in Table III. The recognizer with norm equalization (configurations 3 and 4) improves the recognition performance for noisy speech, but its advantage over the conventional recognizer (configuration 1) is only about 10–15 dB in SNR, which is less than that obtained for the white noise case (as expected).

It might be noted that we have used here only the cepstral coefficients as parameters for speech recognition. However, it might be advantageous to use, in addition, other parameters (such as the first- and second-order temporal derivatives of cepstral coefficients, energy and its derivatives, etc.) for speech recognition. It is not clear, at present, how these parameters behave in the presence of noise. Therefore, these parameters have not been used in the present study.

## V. SUMMARY

We have presented a methodology and an extension of the LPC cepstral norm equalization model to a statistical framework. To cope with the problem of norm shrinkage, a first-order quantization mechanism is incorporated in the

hidden Markov model for speech recognition. Particular care in dealing with the dispersion problem was extensively addressed. Isolated word recognition experiments were conducted and the results indicate that the norm equalization model is an effective measure to resist noise. When compared to a conventional recognizer without noise compensation, the norm equalization model leads to an improvement in recognition performance which is equivalent to about 15–20 dB gain in SNR.

## Appendix
## Singular Multivariate Normal (SMN) Distribution

The theory of singular Gaussian distributions has been well developed by Khatri [4]. Here, we summarize the results that are relevant to our discussion.

Let $\mathcal{R}_k$ be the usual $k$-dimensional vector space and $\boldsymbol{c} \in \mathcal{R}_k$. Vector $\boldsymbol{b}$ relates to $\boldsymbol{c}$ through the projection operation

$$\boldsymbol{b} = \left(I - \frac{\eta\eta^t W}{\eta^t W \eta}\right)\boldsymbol{c}. \tag{A1}$$

Random vector $\boldsymbol{b}$ is said to have Gaussian distribution, i.e., $f(\boldsymbol{b}) = \mathcal{N}_k(0, \Sigma)$ where the rank of $\Sigma$, $\gamma(\Sigma)$, is $\gamma(\Sigma) = q < k$. (In the current case, $q = k - 1$.) The zero mean assumption is arbitrary.

Further let $B$ be a $k \times q$ matrix of orthonormal column vectors belonging to the linear space spanned by the columns of $\Sigma$. In addition, $H$ is a $k \times (k - q)$ matrix, of rank $k - q$, that satisfies $H^t\Sigma = 0$. Column vector $\boldsymbol{b}$ can be transformed into a concatenation of two vectors, $\boldsymbol{r}$ and $\boldsymbol{d}$

$$\boldsymbol{b}^t = (\boldsymbol{r}^t : \boldsymbol{d}^t) \tag{A2}$$

where $\boldsymbol{r} = B^t\boldsymbol{b}$ and $\boldsymbol{d} = H^t\boldsymbol{b}$. The dispersion matrix of $\boldsymbol{r}$ is $B^t\Sigma B$, resulting in $\boldsymbol{r} = \mathcal{N}_q(0, B^t\Sigma B)$. Specifically, $\boldsymbol{r}$ has pdf

$$f(\boldsymbol{r}) = \frac{(2\pi)^{-q/2}}{|B^t\Sigma B|^{1/2}} \exp\left\{-\tfrac{1}{2}\,\boldsymbol{r}^t(B^t\Sigma B)^{-1}\boldsymbol{r}\right\}. \tag{A3}$$

Furthermore, we use the following expression:

$$\begin{aligned}\boldsymbol{r}^t(B^t\Sigma B)^{-1}\boldsymbol{r} &= \boldsymbol{b}^t B(B^t\Sigma B)^{-1}B^t\boldsymbol{b} \\ &= \boldsymbol{b}^t\Sigma^-\boldsymbol{b}\end{aligned} \tag{A4}$$

for a certain choice of general inverse $\Sigma^-$.

Khatri's results [4] included maximum likelihood estimation of the parameters for singular multivariate normal distributions. Suppose the matrix $G = (\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_M)$ results from the projection operation on $(\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_M)$ according to (A1). It is natural to choose the likelihood function to be

$$\mathcal{L}(G|\Sigma) = (2\pi)^{-qM/2}[\nu(\Sigma)]^{-M} \exp\left\{-\tfrac{1}{2}\,\mathrm{tr}\,\Sigma^- GG^t\right\} \tag{A5}$$

where $\nu(\Sigma) = (\xi_1\ \xi_2 \cdots \xi_q)^{1/2}$, and "tr" denotes "trace of."

Since the transformation $H$ leads to null dispersion of $\boldsymbol{d}$, the subspace of $H$ is thus of no concern to the estimation objective. The ML estimate of $\Sigma$, $\hat{\Sigma}$, can be found [4] to satisfy

$$B^t\hat{\Sigma}B = \frac{1}{M}B^tSB \tag{A6}$$

where

$$S = \sum_{i=1}^{M} \boldsymbol{b}_i\boldsymbol{b}_i^t \tag{A7}$$

the sample covariance. That is

$$\hat{\Sigma} = \frac{1}{M}S. \tag{A8}$$

This shows that the conventional results of sample covariance can be straightforwardly applied to the ML estimation problem even though the data may appear to be of insufficient rank support.

## References


[1] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1659–1671, Nov. 1989.

[2] B. A. Dautrich, L. R. Rabiner, and T. B. Martin, "On the effect of varying filter bank parameters on isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 793–806, Aug. 1983.

[3] Y. Ephraim, J. G. Wilpon, and L. Rabiner, "A linear predictive front-end processor for speech recognition in noisy environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (Dallas, TX), Apr. 1987, pp. 1324–1327.

[4] C. G. Khatri, "Some results for the singular normal multivariate regression models," *Sankya*, vol. A30, pp. 267–280, 1968.

[5] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *ASSP Mag.*, vol. 3, no. 1, pp. 1, pp. 4–16, Jan. 1986.

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

[7] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 947–954, July 1987.

[8] K. K. Paliwal and M. M. Sondhi, "Recognition of noisy speech using cumulant-based linear prediction analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (Toronto, Canada), May 1991, pp. 429–432.

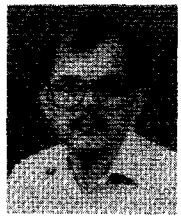


**Biing-Hwang Juang** (S'79–M'80–SM'87–F'92) received the B.Sc. degree in electrical engineering from National Taiwan University, Taipei, in 1973 and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of California, Santa Barbara, in 1979 and 1981, respectively.

In 1978, he did research on vocal tract modeling at Speech Communications Research Laboratory (SCRL). He then joined Signal Technology, Inc., in 1979 as Research Scientist, working on signal and speech related topics. Since 1982, he has been with AT&T Bell Laboratories where he is engaged in a wide range of speech related research activities. He has published extensively in the area of speech communication and holds two sets of patents.

Dr. Juang has served on several IEEE technical committees and chaired IEEE workshops. He was an Associate Editor for the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING from 1986 to 1988. He

currently chairs the Technical Committee on Neural Networks for Signal Processing in the IEEE Signal Processing Society. He also serves on several international advisory boards outside the United States and is Associate Editor of the *Journal of Speech Communication.*



**Kuldip K. Paliwal** (M'89) was born in Aligarh, India, in 1952. He received the B.S. degree from Agra University, India, in 1969, the M.S. degree from Aligarh University, India, in 1971, and the Ph.D. degree from Bombay University, India, in 1978.

Since August 1972, he has been with the Tata Institute of Fundamental Research, Bombay, India, where he has worked on various aspects of speech processing; e.g., speech recognition, speech coding, and speech enhancement. From September 1982 to October 1984, he was an NTNF Fellow at the Department of Electrical and Computer Engineering, Norwegian Institute of Technology, Trondheim, Norway. He was a Visiting Scientist at the Department of Communications and Neuroscience, University of Keele, U.K., during June to September 1982 and January to March 1984, and at the Electronics Research Laboratory (ELAB), Norwegian Institute of Technology, Trondheim, Norway, during April to July 1987, April to July 1988, and March to May 1989. From May 1989 to December 1991, he was at the Acoustics and Speech Research Departments, AT&T Bell Laboratories, Murray Hill, NJ. His work has been concentrated on vector quantization of linear predictive coding parameters, fast search algorithms for vector quantization, feature analysis and distance measures for speech recognition and robust spectral analysis techniques. His current research interests are directed towards automatic speech recognition using hidden Markov models and neural networks.

Dr. Paliwal is a Fellow of the Acoustical Society of India. He is a member of the IEEE Signal Processing Society's Technical Committee on Neural Networks.