VECTOR EQUALIZATION IN HIDDEN MARKOV MODELS FOR NOISY SPEECH RECOGNITION

B.H. Juang and K.K. Paliwal

AT&T Bell Laboratories, Murray Hill, NJ 07974, USA

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 distortion measure 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 norm 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.

1. 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.

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 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.

2. NORM SHRINKAGE AND EQUALIZATION MODEL

Assume that the speech signal $\{x_n\}$ is distorted by an additive white noise $\{v_n\}$, and we observe the distorted signal $\{y_n\}$, where

$$y_n = x_n + v_n . \tag{1}$$

Let $\mathbf{c}' = [c_1\ c_2\ \cdots\ c_k]$ and $(\mathbf{c}')' = [c_1'\ c_2'\ \cdots\ c_k']$ be the k-dimensional cepstral vectors obtained through p-th order LPC analysis of x_n and y_n, respectively. (Usually k is greater than p.) It has been shown [1], both empirically and theoretically, that the presence of white noise v_n causes reduction in the cepstral norm (or energy); i.e.,

$$|\mathbf{c}'| \le |\mathbf{c}| \tag{2}$$

This property of cepstral norm shrinkage was used in [1] for improving speech recognition performance. It was shown [1] that recognition performance can be improved if the calculation of the Euclidean distance involves an equalizing scalar θ. In particular, if $\mathbf{c}$ and η denote the testing cepstral vector (with an unknown amount of noise) and the *clean* reference cepstral vector, respectively, the revised distance is defined by

$$d(\mathbf{c}, \eta) = (\mathbf{c} - \theta\eta)'(\mathbf{c} - \theta\eta) \tag{3}$$

where $\theta = (\mathbf{c}'\eta)/|\eta|^2$.

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 $\{\mathbf{c}_i\}_{i=1}^{M}$, the centroid problem requires calculation of a vector η that minimizes the accumulated distance defined, with the equalizing factor in the current case, as

$$D = \sum_{i=1}^{M} (\mathbf{c}_i - \theta_i\eta)'(\mathbf{c}_i - \theta_i\eta) \tag{4}$$

where θ_i is defined as above for each $\mathbf{c}_i$. Eq. (4) can be rewritten as

0-7803-0532-9/92 $3.00

$$D = \sum_{i=1}^{M} |\mathbf{c}_i|^2 - \sum_{i=1}^{M} (\mathbf{c}_i^t \eta_1)^2 \qquad (5)$$

where $\eta_1 = \eta / |\eta|$. With the constraint that $|\eta_1| = 1$, it is straightforward to show that the solution η_1 satisfies

$$\xi \eta_1 = 2 \Xi \eta_1 \qquad (6)$$

where

$$\Xi = \sum_{i=1}^{M} \mathbf{c}_i \mathbf{c}_i^t, \qquad (7)$$

is the sample covariance. We choose η_1 as the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix to minimize the accumulated distance of Eq. (5). This centroid calculation procedure is a very important one in the following statistical formulation of a first order equalization model of the noisy LPC cepstrum.

The distance measure of Eq. (3) can be generalized to a multivariate probabilistic formulation with the following probability density function (pdf)

$$f(\mathbf{c}) = K \cdot \exp\left[-\frac{1}{2}(\mathbf{c} - \theta\eta)^t W (\mathbf{c} - \theta\eta)\right] \qquad (8)$$

where K is the normalization factor, and $\theta = (\eta^t W \mathbf{c})/(\eta^t W \eta)$. The presence of the equalizing factor θ makes $f(\mathbf{c})$ in Eq. (8) an unusual density since $\left[I - \frac{\eta\eta^t W}{\eta^t W \eta}\right]$ 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 $W\eta = \mathbf{0}$, Eq. (8) becomes

$$f(\mathbf{c}) = K \cdot \exp\left[-\frac{1}{2}\mathbf{c}^t W \mathbf{c}\right]. \qquad (9)$$

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

3. HMM WITH NORM EQUALIZATION

A hidden Markov model (HMM) λ is a triple $\lambda = (\pi, A, F)$, where π 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 π 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 λ for a sequence of observations, $\{\mathbf{c}_t\}_{t=1}^{T} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_T)$, is

$$P_\lambda(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_T) = \sum_s \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1} s_t} f_{s_t}(\mathbf{c}_t) \qquad (10)$$

where a_{ij} are the elements of A, $A = [a_{ij}]_{i,j=1}^{N}$, and π_i are the elements of π, $\pi^t = [\pi, \pi_2, \ldots, \pi_N]$. Quantity a_{ij} is thus the probability of making a transition to state j given that the current state is i and π_i is the probability of staying at state i at the beginning. The equalization mechanism is implemented in some specific forms in the observation density f_i.

3.1 Reestimation Algorithm

Baum's reestimation algorithm [6] is an iterative maximization algorithm in which the model parameters λ, 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 λ':

$$Q(\lambda, \lambda') = \sum_s P_\lambda\left[\{\mathbf{c}_t\}_{t=1}^{T}, s\right] \log P_{\lambda'}\left[\{\mathbf{c}_t\}_{t=1}^{T}, s\right] \qquad (11)$$

   where s is a state sequence, $s = (s_0, s_1, \ldots, s_T)$, 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 λ'.

During the next iteration, the new model $\bar{\lambda}$ is used in place of the old model λ and the two steps repeat again. It can be shown [6] that each iteration guarantees an increase in likelihood. The algorithm stops when it reaches a fixed point solution or when the increase in likelihood falls below a prescribed level.

3.2 Maximization of $Q(\lambda, \lambda')$

The auxiliary function $Q(\lambda, \lambda')$ is defined by Eq. (11), where the logarithmic term can be broken down into individual groups of parameters as follows:

$$\log P_{\lambda'}\left[\{\mathbf{c}_t\}_{t=1}^{T}, s\right] = \log \pi'_{s_0} + \sum_{t=1}^{T} \log a'_{s_{t-1} s_t} + \sum_{t=1}^{T} \log f'_{s_t}(\mathbf{c}_t).$$

Maximization of $Q(\lambda, \lambda')$ over parameters π and A thus remains 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:

$$\begin{aligned} &\sum_s P_\lambda\left[\{\mathbf{c}_t\}_{t=1}^{T}, s\right] \sum_{t=1}^{T} \log f'_{s_t}(\mathbf{c}_t) \\ &= \sum_{i=1}^{N} \sum_{t=1}^{T} P_\lambda\left[\{\mathbf{c}_t\}_{t=1}^{T}, s_t = i\right] \log f'_i(\mathbf{c}_t) \\ &= \sum_{i=1}^{N} Q_f(\lambda, \mathbf{f}'_i). \end{aligned} \qquad (12)$$

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

$$\nabla_{\mathbf{f}'_i} Q_f(\lambda, \mathbf{f}'_i)\Big|_{\mathbf{f}'_i = \bar{\mathbf{f}}_i} = 0, \qquad (13)$$

where

$$Q_f(\lambda, \mathbf{f}'_i) = \sum_{t=1}^{T} P_\lambda\left[\{\mathbf{c}_t\}_{t=1}^{T}, s_t = i\right] \log f'_i(\mathbf{c}_t). \qquad (14)$$

3.3 Observation Density with First Order Equalization

The proposed equalization as suggested in Eq. (3) 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 Eq. (3), 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.

3.3.1 Fixed Dispersion

The particular form of observation probability densities we consider in this class is

$$f(\mathbf{c}) = \frac{1}{(2\pi)^{k/2} \mid \Sigma \mid^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{c}-\theta\eta)^t \Sigma^{-1}(\mathbf{c}-\theta\eta)\right\} \quad (15)$$

where Σ is positive-definite and fixed. When $\Sigma = I$, the identity matrix, this form of density function becomes identical to what Eq. (3) has implied. The fixed dispersion matrix can thus be considered a weighting matrix in the vector space. With

$$f_i'(\mathbf{c}_t) = \frac{1}{(2\pi)^{k/2} \mid \Sigma_i \mid^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{c}_t-\theta_{it}\eta_i')^t \Sigma_i^{-1}(\mathbf{c}_t-\theta_{it}\eta_i')\right\}$$

maximization of the auxiliary function defined in Eq. (12) with respect to η_i' becomes the typical problem outlined in Eqs. (5-7). The factorization $\Sigma_i^{-1} = U_i U_i^t$ facilitates the transformations $\mathbf{y}_t = U_i^t \mathbf{c}_t$ and $\zeta_i' = U_i^t \eta_i' / \mid U_i^t \eta_i' \mid$. The optimization objective becomes minimization of

$$\Omega = \sum_{t=1}^{T} P_\lambda\left[\{\mathbf{c}_t\}_{t=1}^T, s_t = i\right]\left\{(\mathbf{c}_t-\theta_{it}\eta_i')^t \Sigma_i^{-1}(\mathbf{c}_t-\theta_{it}\eta_i')\right\}$$

$$= \sum_{t=1}^{T} P_\lambda\left[\{\mathbf{c}_t\}_{t=1}^T, s_t = i\right]\left\{\mathbf{c}_t^t \Sigma_i^{-1}\mathbf{c}_t - \frac{[(\eta_i')^t \Sigma_i^{-1}\mathbf{c}_t]^2}{(\eta_i')^t \Sigma_i^{-1}\eta_i'}\right\}$$

$$= \sum_{t=1}^{T} P_\lambda\left[\{\mathbf{c}_t\}_{t=1}^T, s_t = i\right]\left\{\mathbf{c}_t^t \Sigma_i^{-1}\mathbf{c}_t - [(\zeta_i')^t \mathbf{y}_t]^2\right\} \quad (16)$$

subject to $\mid \zeta_i' \mid = 1$. The objective now is identical to Eqs. (5-7) and the solution $\bar{\zeta}_i$ is thus the eigenvector satisfying

$$\xi_{\max}\bar{\zeta}_i = 2\Xi\bar{\zeta}_i \quad (17)$$

where

$$\Xi = \sum_{t=1}^{T} P_\lambda\left[\{\mathbf{c}_t\}_{t=1}^T, s_t = i\right]\mathbf{y}_t\mathbf{y}_t^t \quad (18)$$

and $\xi_{\max}$ is the maximal eigenvalue of Ξ. The difference between Eq. (18) and Eq. (7) is the weighting factor due to $P_\lambda\left[\{\mathbf{c}_t\}_{t=1}^T, s_t = i\right]$.

3.3.2 Singularized Dispersion

There are two parameter categories involved in the estimation of distributions with the form of Eq. (15), η_i' and Σ_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 θ_{it} is chosen to maximize the exponent in Eq. (15). Estimation of these two parameter categories thus has to take the equalizing factor into account.

The equalizing factor results in the following

$$\mathbf{c}_t - \theta_{it}\eta_i = \left(I - \frac{\eta_i\eta_i^t\Sigma_i^{-1}}{\eta_i^t\Sigma_i^{-1}\eta_i}\right)\mathbf{c}_t \quad (19)$$

where the bracketed term is a projector which projects a vector onto a hyperplane perpendicular to η_i. The projection operator results in singularized dispersion and thus necessitates a particular treatment based on the theory of singular distributions. A theory of singular Gaussian distributions has been well-developed by Khatri [4].

If Σ^{-1} in Eq. (19) is singularized along a particular direction and η_i is chosen to be in that direction, the density function collapses to Eq. (9). Therefore, norm equalization can be embedded in the singularization process. This naturally leads to the use of the following singular multivariate density

$$f_i(\mathbf{c}_t) = (2\pi)^{-k/2} \mid \Sigma_i^- \mid^{1/2} \exp\left\{-\frac{1}{2}\mathbf{c}_t^t \Sigma_i^- \mathbf{c}_t\right\} \quad (20)$$

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

$$\Sigma_i^- = V\Lambda V^t \quad (21)$$

where

$$V = [\rho_1, \rho_2, \ldots, \rho_{k-1}] \quad (22)$$

and

$$\Lambda = \begin{bmatrix} \xi_1 & & 0 \\ & \ddots & \\ 0 & & \xi_{k-1} \end{bmatrix} \quad (23)$$

Note that Khatri's results [4] are directly applicable in this case. The reestimation transformation for the covariance matrix, therefore, involves two steps:

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

$$\bar{\Sigma}_i = \frac{\sum_{t=1}^{T} P_\lambda\left[\{\mathbf{c}_t\}_{t=1}^T, s_t = i\right]\mathbf{c}_t\mathbf{c}_t^t}{\sum_{t=1}^{T} P_\lambda\left[\{\mathbf{c}_t\}_{t=1}^T, s_t = i\right]} \quad (24)$$

2. Singularize $\bar{\Sigma}_i^{-1}$ to $\bar{\Sigma}_i^-$ according to Eqs. (21-23).

4. 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 multi-speaker mode and study it for the 10-word English digit vocabulary. The data base consists of speech from 4 talkers (2 males and 2 females). 24 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 8-th 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 studied the following four configurations of the HMM-based speech recognizer:

1. Configuration 1: The recognizer does not incorporate the cepstral norm shrinkage model; i.e., it is a conventional HMM-based speech recognizer.
2. 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., $\Sigma = I$.
3. 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.
4. Configuration 4: The HMM-based speech recognizer uses the first order norm equalization model with singularized dispersion as described in Section 3.3.2.

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 desired SNR. Recognition results for noisy speech at eight different SNR's (∞ dB, 35 dB, 30 dB, 25 dB, 20 dB, 15 dB, 10 dB and 5 dB) are shown in Table 1. 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.

Table 1. Recognition performance as a function of SNR

SNR (dB)	Recognition accuracy (%) with			
	Conf. 1	Conf. 2	Conf. 3	Conf. 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

5. 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 equalization 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.

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. ASSP-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. R. Rabiner, "A linear predictive front-end processor for speech recognition in noisy environments," *ICASSP-87*, 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 Magazine*, vol. 3, no. 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.