Gain-Adapted Hidden Markov Models for Recognition of Clean and Noisy Speech

Yair Ephraim, Senior Member, IEEE

Abstract—A key issue in applying hidden Markov modeling for recognition of speech signals is the matching of the energy contour of the signal to the energy contour of the model for that signal. Energy matching is normally achieved by appropriate normalization of each vector of the signal prior to both training and recognition. This approach, however, is not applicable when only noisy signals are available for recognition. A unified approach is developed for gain adaptation in recognition of clean and noisy signals. In this approach, hidden Markov models (HMM's) for gain-normalized clean signals are designed using maximum likelihood (ML) estimates of the gain contours of the clean training sequences. The models are combined with ML estimates of the gain contours of the clean test signals, obtained from the given clean or noisy signals, in performing recognition using the maximum a posteriori decision rule. The gain-adapted training and recognition algorithms are developed for HMM's with Gaussian subspaces using the expectation-maximization (EM) approach, and tested in recognition of clean and noisy speech signals.

I. INTRODUCTION

Let \{H_i, i = 1, \ldots, W\} be a sequence of W words in a given vocabulary. Let \(z = \{z_t, t = 0, \ldots, T - 1\}, z \in \mathbb{R}^K\), be a sample function of noisy acoustic speech signals corresponding to some word in the vocabulary, where \(\mathbb{R}^K\) is the K-dimensional Euclidean space. The problem of interest is that of classifying \(z\) as one of the words in the vocabulary such that the resulting probability of classification error is minimum. This is the problem of finding the partition of the sample space of the acoustic noisy signals from all words in the vocabulary which yields minimum probability of classification error.

Let \(\Omega = \{\omega_1, \ldots, \omega_W\}\) be a partition of the sample space of the noisy signals. The probability of classification error associated with this partition is given by

\[
P_e(\Omega) = \sum_{i=1}^{W} P(H_i) \int_{\mathbb{R}^K} p(z|H_i) dz
\]

where \(P(H_i)\) is the a priori probability of occurrence of the \(i\)th word, and \(p(z|H_i)\) is the probability density function (pdf) of the noisy signal \(z\) given that the word \(H_i\) was spoken. The minimization of \(P_e(\Omega)\) is achieved by the well-known maximum a posteriori (MAP) decision rule given by

\[
\max_{1 \leq i \leq W} P(H_i)p(z|H_i).
\]

According to (2), an utterance \(z\) will be classified as coming from the \(i\)th word if \(P(H_i)p(z|H_i)\) is greater than \(P(H_j)p(z|H_j)\) for all \(j \neq i\). Ties in (2) are arbitrarily broken. Note that in the absence of noise, \(z\) will be a sample function of the clean signal, \(p(z|H_i)\) will be the pdf of the clean signal given that the \(i\)th word was spoken, and (2) will be identical to the decision rule used for classification of clean signals (see, e.g., [1]). Thus, in principle, the problems of classification of clean and noisy speech signals are identical.

For continuous time signals with integrable variance, and statistically independent additive white Gaussian noise in the mathematical sense (i.e., its integral is a Wiener, or Brownian motion, process), it was shown in [2] that minimum probability of error classification can be achieved by applying the MAP decision rule to causal minimum mean square error (MMSE) estimates of the clean signals. This theoretical result provides the intuitive basis for the commonly used approach to recognition of noisy speech (see, e.g., [3]-[6]), in which the clean signal or some function of that signal (e.g., sample spectrum, autocorrelation, cepstrum) is first estimated, and then the MAP decision rule is applied. We opt for the direct recognition approach of (2), since for the statistical framework of hidden Markov modeling used here, this approach is simpler than that of applying the MAP decision rule to the MMSE estimator of the clean signal [7].

Let \(y = \{y_t, t = 0, \ldots, T - 1\}, y_t \in \mathbb{R}^K\), denote a sample function of the clean signal from some word in the vocabulary. Let \(v = \{v_t, t = 0, \ldots, T - 1\}, v_t \in \mathbb{R}^K\), denote a sample function of the noise process. Assume that \(z = y + v\), and that the signal and noise are statistically independent. In this case we have that

\[
p(z|H_i) = \int p(y|H_i)p_v(z - y) dy
\]

where \(p(y|H_i)\) denotes the pdf of the clean signal \(y\) given that the \(i\)th word was spoken, \(p_v(\cdot)\) denotes the pdf of the noise process, and the integral in (3) is defined on \(\mathbb{R}^{K_T}\). In practice, the pdf's \(p(y|H_i)\) and \(p_v(\cdot)\) are unknown, and hence they are replaced by models whose parameters are estimated from training data of speech and noise. Let \(p_{\lambda_i}(y|H_i)\) be the model for \(p(y|H_i)\), where \(\lambda_i\) denotes the
parameter set of the signal model. Similarly, let \( p_{\lambda}(r) \) be the model for \( p_{\lambda}(r) \), where \( \lambda \) denotes the parameter set of the noise model. In this case, the pdf (3) of the noisy signal from the \( i \)th vector is approximated by

\[
p(z|H) = \int p_{\lambda}(r|H)p_{\lambda}(z - y) \, dy
\]

\[
= p_{\lambda}(z|H)
\]

(4)

where \( \lambda \) is \((\lambda_1, \lambda_2) \). Hence, for equiprobable words, the MAP decision rule (2) is approximated by

\[
\max_{\lambda \in \Lambda} p_{\lambda}(z|H).
\]

(5)

A hidden Markov model (HMM) with a mixture of Gaussian pdf’s at each state is normally chosen for \( p_{\lambda}(r|H) \) [8]. Furthermore, the covariance matrices of the Gaussian processes are normally parameterized either by autoregressive (AR) modeling or simply by assuming that each matrix is diagonal [8]. Here we shall focus on the AR parameterization which has been proven useful for speech signals (see, e.g., [9]), though the approach used in this paper is applicable to other parameterizations as well. HMM’s with a mixture of Gaussian AR pdf’s at each state will be referred to as Gaussian AR HMM’s, and their mathematical description will be given in Section II. The model for the pdf \( p_{\lambda} \) depends on the nature of the noise. For zero mean stationary wide-band Gaussian noise processes considered here, we use a single state single mixture component Gaussian AR HMM.

A key issue in designing speech recognition systems which are based upon statistical modeling, is the matching of the energy contour of the test clean signal to the energy contour of the model for that signal. Such energy matching is essential for two reasons. First, speech signals are quasi-stationary and therefore have time-varying energy. Hence, the energy contour of the test signal may not be reliably estimated from the training data even if the test and training data were recorded under similar gain conditions. Second, the training and test data are often recorded under different gain conditions which may also be time varying. Such situations occur, for example, when training is done under laboratory conditions and recognition is performed over time varying communication channels. A mismatch between the energy contours of the clean signal and the model for that signal results in misrepresentation of the statistics of the signal by the model, and hence in erroneous calculation of probabilities of speech events.

When clean signals are available for recognition, the energy matching is usually achieved by normalizing each vector of the clean signal prior to both training and recognition, thus neutralizing any possible effect of gain mismatch. In speech recognition systems which are based upon Gaussian AR HMM’s, each vector \( y \) is normalized by \( \sigma_y \), the standard deviation of the residual signal obtained from AR modeling of \( y \) [10]. In speech recognition systems that are based upon cepstral HMM’s (see, e.g., [8]), the zeroth cepstral coefficient of each vector of the signal, which is proportional to the energy of the signal in that vector, is excluded. Higher order cepstral coefficients are obtained from the AR coefficients of the signal in that vector using the recursion [9, p. 230]. Thus, for each \( t, y \) is effectively normalized by \( r^2_{y_0} \), the square root of the zeroth autocorrelation coefficient of the signal \( y \). When only noisy speech is available for recognition, these approaches of gain normalization are not applicable, and the problem becomes significantly harder. For example, normalizing \( z \) by an estimate of \( \sigma_z \) is not a trivial task, since this normalization requires the estimation of the entire AR model for each vector of the clean speech given the noisy speech. This problem was treated in [12] where algorithms for maximum likelihood (ML) estimation of AR models from given noisy signals were developed using the EM (expectation-maximization) algorithm [11].

In this paper we generalize the notion of gain normalization of clean signals, and develop algorithms for gain adaptation in recognition of clean and noisy speech signals. In the gain-adapted approach, an ML estimate of the gain contours of the training data from each word is used for designing an HMM for gain-normalized signals from that word. The resulting models are combined with ML estimates of the gain contours of the clean test signals, obtained from the given signals, in applying the MAP decision rule. Thus, recognition is performed using models for gain-normalized clean signals designed from the training data, and estimates of the gain contours of the clean signals obtained from the given clean or noisy test signals. The mathematical formulation of the gain-adapted approach is now given.

Let \( g \) be \([g_0, \ldots, g_{T-1}] \), \( g > 0 \), be a sequence of gain factors, or a gain contour, for the signal \( y = [y_0, \ldots, y_{T-1}] \). The pdf of the sequence \( y \) expressed in terms of the HMM for the gain-normalized signal \( \{y_0/g_0, \ldots, y_{T-1}/g_{T-1}\} \), assuming that the \( i \)th word was spoken, is given by

\[
p_{\lambda}(y|H, g) = p_{\lambda}(y_0/g_0, \ldots, y_{T-1}/g_{T-1}|H).
\]

(6)

Note that \( \lambda_i \) in (6) denotes the parameter set of the HMM for the gain-normalized signal. The pdf of the model for the noisy signal \( z = [z_0, \ldots, z_{T-1}] \) assuming that the \( i \)th word was spoken is obtained from (4) and (6) and is given by

\[
p_{\lambda}(z|H, g) = \int p_{\lambda}(y|H, g)p_{\lambda}(z - y) \, dy.
\]

(7)

Gain-adapted training of the HMM for the \( i \)th word results from ML estimation of the parameter set \( \lambda_i \) from a training sequence \( y \) from that word using an ML estimate of the gain contour \( g \). Specifically, \( \lambda_i \) is estimated from

\[
\max_{\lambda_i} \max_{g} p_{\lambda_i}(y|H, g).
\]

(8)

Gain-adapted recognition of the noisy signal \( z \) is performed by applying the MAP decision rule to the pdf (7)
using ML estimates of the gain contour $g$ obtained from the given noisy signal $z$. Specifically, the noisy signal $z$ is associated with the word $H_i$ which results from

$$\max_{1 \leq i \leq W} \max_{g} p_n(z|H_i, g). \quad (9)$$

In the absence of noise, the recognition test (9) becomes

$$\max_{1 \leq i \leq W} \max_{g} p_n(y|H_i, g). \quad (10)$$

The likelihood ratio test associated with this decision rule is referred to as the generalized likelihood ratio test [21, p. 92].

In Sections II and III of this paper, we develop algorithms for the gain-adapted training (8), and the gain-adapted recognition of noisy signals (9), respectively, using the EM approach, and study their efficient implementation in the frequency domain. In Section IV we study the performance of the proposed algorithms in recognition of clean and noisy speech signals corresponding to the English digits and $E$-set words (b, c, d, e, g, p, t, v, z), at signal-to-noise ratios (SNR's) greater than or equal to 5 dB. For digit recognition at input SNR greater than or equal to 10 dB, it is shown that the gain-adapted recognition approach performed significantly better than other approaches in which either a gain normalization is applied or the gain mismatch is simply ignored. A comparison with the latter approach shows that the gain-adapted approach elevated the recognition accuracy by 2%–13% when the training and test data were recorded under similar gain conditions, and by 5%–21% when a recording gain mismatch existed. In the first case where the training and test data were recorded under similar gain conditions, gain adaptation is needed due to the quasi-stationarity of the speech signals as explained earlier. In the second case, gain adaptation due to both the quasi-stationarity of the signals and recording conditions was studied.

Speech recognition approaches which use decision rules similar to (5) were studied in [13]–[15]. In [13], a Gaussian mixture model was assumed for the clean signal, and a Gaussian model was attributed to the noise process. The model for the noise was independently estimated from a given sample function of the noise process. The models for the noisy signals from the different words in the vocabulary, were estimated from training data of noisy signals using the EM algorithm. In [14], [15], HMM's were assumed for the clean signal and the noise process. These models were independently estimated from training data of clean speech and noise samples, respectively. The models for the noisy signals from the different words in the vocabulary were obtained from the individually estimated models for the clean signals and noise process. The latter approach is conceptually similar to the approach taken in this paper. The advantage of this approach compared to [13], is that the models for the clean signals from the different words in the vocabulary have to be estimated once only, regardless of the particular noisy environment.

In [13], however, a new set of models for the noisy signals must be estimated for any paradigm of noisy signals, since these models are estimated from noisy training data.

The problem of gain adaptation in speech recognition considered here complements similar problems in speech coding [16]–[19]. AR modeling of speech signals [20], and enhancement of noisy speech [7]. In all cases, the gain of the signal enjoyed a special treatment. In [16]–[19], several approaches were developed for designing vector quantizers using shape-gain product code books. In [19], an approach for gain-adapted vector quantization was developed. The principal difference between the two cases is that in the shape-gain vector quantization approach the gain can take a finite number of values, while in the gain-adapted vector quantization approach the gain can theoretically take an infinite number of values. In both cases, the separation of the shape and gain code books was proved useful as more efficient coding schemes were obtained for a given complexity. In [7], MAP gain-adapted signal estimators for speech enhancement applications were developed, using a similar approach to that considered here.

II. GAIN-ADAPTED HIDDEN MARKOV MODELING

A. Gaussian AR HMM's

Let $p_n(y)$ be the pdf of the HMM for the acoustic signal from some word in the vocabulary. For simplicity of notation, we shall suppress the conditioning on $H_i$ as the discussion will be applicable for any word in the vocabulary. We consider first-order Gaussian AR HMM's with $M$ states, $L$ mixture component per state, and AR processes of order $N$. For this class of models, the pdf $p_n(y)$ is given by

$$p_n(y) = \sum_x p_n(x)p_n(y|x) \quad (11)$$

where $x = \{x_t, t = 0, \cdots, T - 1\}, x_t \in \{1, \cdots, M\}$, denotes a sequence of states corresponding to the sequence of clean signal vectors $y = \{y_t, t = 0, \cdots, T - 1\}$, $p_n(x)$ denotes the probability of the sequence of states $x$, and $p_n(y|x)$ is the pdf of the sequence of output vectors $y$ given $x$. For first-order HMM's, $p_n(x)$ is given by

$$p_n(x) = \prod_{t=0}^{T-1} a_{x_{t-1}x_t} \quad (12)$$

where $a_{x_{t-1}x_t}$ denotes the transition probability from state $x_{t-1}$ at time $t - 1$ to state $x_t$ at time $t$, and $a_{x_{t-1}x_0} = \pi_0$ denotes the probability of the initial state $x_0$. For the pdf $p_n(y|x)$ we make the standard assumption that

$$p_n(y|x) = \prod_{t=0}^{T-1} p_n(y_t|x_t) \quad (13)$$

where $p_n(y_t|x_t)$ is the pdf of the vector $y_t$ given that this vector was generated from state $x_t$. For Gaussian AR HMM's, i.e., HMM's with a mixture of Gaussian AR
pdf's at each state, \( p_{\lambda_t}(y_t|x_t) \) is given by

\[
p_{\lambda_t}(y_t|x_t) = \sum_{h} c_{h|x_t} b(y_t|x_t, h_t)
\]

where \( h_t \in \{1, \ldots, L\} \) denotes the mixture component chosen at time \( t \), \( c_{h|x_t} \) is the probability of choosing the mixture component \( h_t \) given that the process is in state \( x_t \), and \( b(y_t|x_t, h_t) \) is the pdf of the Gaussian AR output vector \( y_t \), given \( x_t, h_t \). This pdf is given by

\[
b(y_t|x_t, h_t) = \frac{\exp \left\{-\frac{1}{2} y_t^T S_{h|x_t}^{-1} y_t \right\}}{(2\pi)^{K/2} \det^{1/2}(S_{h|x_t})}
\]

where \( \# \) denotes vector Hermitian transpose, \( S_{h|x_t} = \sigma_{h|x_t}^2 (A_{h|x_t}^T A_{h|x_t} A_{h|x_t})^{-1} \), \( \sigma_{h|x_t}^2 \) is the variance of the innovation process of the AR source, \( A_{h|x_t} \) is a \( K \times K \) lower triangular Toeplitz matrix in which the first \( N_t + 1 \) elements of the first column constitute the coefficients of the AR process, which are denoted here by \( \xi_{h|x_t}(0), \xi_{h|x_t}(1), \ldots, \xi_{h|x_t}(N_t) \), and \( \xi_{h|x_t}(0) = 1 \) [22]. This formulation of the AR model for \( y_t \) is identical to the ML formulation of linear prediction using the autocorrelation method as given in [9, pp. 20-22].

The pdf \( p_{\lambda_t}(y) \) can be equivalently written in the following form which will prove useful in our derivation:

\[
p_{\lambda_t}(y) = \sum_{x} \sum_{h} p_{\lambda_t}(x, h, y) = \sum_{x} \sum_{h} p_{\lambda_t}(x|h) p_{\lambda_t}(h|x) p_{\lambda_t}(y|x, h)
\]

where \( h = \{h_t, t = 0, \ldots, T-1\} \) denotes a sequence of mixture components, and

\[
p_{\lambda_t}(h|x) = \prod_{t=0}^{T-1} \prod_{i=0}^{T-1} \sum_{x_t} c_{x_t, h_t} \Delta_{x_t, h_t}
\]

\[
p_{\lambda_t}(y|x, h) = \prod_{t=0}^{T-1} \prod_{i=0}^{T-1} b(y_t|x_t, h_t).
\]

The parameter set of the HMM for the clean speech signal is given by \( \lambda_t = (\pi, a, c, s) \), where \( \pi = [\pi_j] \), \( a = [a_{i,j}] \), \( c = [c_{i,j}] \), and \( s = [S_{i,j}] \), for \( \alpha, \beta = 1, \ldots, M \) and \( \gamma = 1, \ldots, L \). Note that since the individual gain of each vector of the training data is estimated in the proposed gain-adapted approach (see (8)), we could restrict the gain \( \sigma_{\gamma|\lambda} \) of each AR model to be unity. We have chosen not to do so since, in general, an AR model designed from the average of the autocorrelation functions of gain-normalized signals, say \( \{y_i/g_i\} \), has a nonunitary gain \( \sigma_{\gamma} \). Thus, when the AR model for \( S_{\gamma|\lambda} \) is estimated from the average of the autocorrelation functions of the gain-normalized vectors assigned to state \( \beta \) and mixture component \( \gamma \) (see, [23, eqs. (28), (33)]), the resulting \( \sigma_{\gamma|\lambda} \) need not be one. By letting \( \sigma_{\gamma|\lambda} \) to take values different than one, the estimated model \( \lambda_t \) will represent the true model for the gain-normalized signal, and the estimated gain \( \gamma \) will be the actual gain contour of the signal. Thus, the two procedures for estimating \( \lambda_t \) and \( g_t \) are not mixed together. The approach taken here has also proven slightly better than the approach of [10], where \( \sigma_{\gamma|\lambda} \) was restricted to be unity, in applications of recognition of prenormalized clean speech signals [24].

B. Gain-Adapted Training

The gain-adapted hidden Markov modeling problem was defined in (8). The gradient equations of \( p_{\lambda_t}(y|g) \) with respect to \( \{\lambda_t, g\} \) are nonlinear and therefore have no simple solution. Hence, the estimation of \( \{\lambda_t, g\} \) is performed here iteratively using the EM approach. The application of this approach for joint maximization of \( p_{\lambda_t}(y|g) \) over \( \{\lambda_t, g\} \) however, results in an auxiliary function whose gradient equations are also nonlinear. Hence, alternate maximization of \( p_{\lambda_t}(y|g) \), once over \( \lambda_t \) assuming \( g \) is known and then over \( g \) assuming \( \lambda_t \) is available, is performed, using the EM algorithm. Since, by definition, each such iteration can either increase the value of \( p_{\lambda_t}(y|g) \) or keep it constant, the proposed approach results in a sequence of HMM's \( \{\lambda_t(m)\} \) with nondecreasing likelihood values. The iterative procedure is terminated if either a fixed point is reached (\( \lambda_t(m+1) = \lambda_t(m), g(m+1) = g(m) \)), or the values of the likelihood function \( \ln p_{\lambda_t}(y|g) \) in two consecutive iterations are sufficiently close to each other. We now show how the EM approach can be applied for alternate estimation of \( \lambda_t \) and \( g_t \).

We begin by assuming that the gain contour \( g_t \) is known, and apply the Baum algorithm [25] for estimating \( \lambda_t \). Let

\[
\ln p_{\lambda_t}(y|g) = \ln \sum_{x_t,g} p_{\lambda_t}(x_t, y_t|g)
\]

where from (6), (12), and (15)-(17) we have that

\[
p_{\lambda_t}(x_t, y_t|g) = \prod_{t=0}^{T-1} a_{x_{t-1},c_{x_{t-1}}b(y_t|x_t, h_t, g_t)}
\]

and \( b(y_t|x_t, h_t, g_t) = b(y_t|g_t|x_t, h_t)g_t^\gamma \). This pdf of \( y_t \), given \( x_t, h_t, g_t \), is Gaussian with zero mean and covariance matrix \( g_t^\gamma S_{\gamma|\lambda} \). Let \( \lambda_t \) and \( g_t \) be a current and a new estimate of the parameter set of the model, respectively. Using Jensen's inequality we have that

\[
\ln \lambda_t - \ln \lambda_t' = \ln \sum_{x_t} \frac{p_{\lambda_t}(x_t|y_t, g_t) p_{\lambda_t}(y_t|g_t)}{p_{\lambda_t'}(x_t|y_t, g_t) p_{\lambda_t'}(y_t|g_t)}
\]

\[
\leq \sum_{x_t} p_{\lambda_t}(x_t|y_t, g_t) \ln \frac{p_{\lambda_t}(y_t|g_t)}{p_{\lambda_t'}(y_t|g_t)}
\]

\[
\ln (\lambda_t, g) - \ln (\lambda_t', g) \leq \phi(\lambda_t, g) - \phi(\lambda_t', g)
\]

where

\[
\phi(\lambda_t, g) = \sum_{x_t} p_{\lambda_t}(x_t, y_t|g) \ln p_{\lambda_t}(x_t, y_t|g)
\]

\[
= \sum_{i=0}^{T-1} \sum_{x_{i-1},x_{i},h_{i}} p_{\lambda_t}(x_{i-1}, x_{i}, h_{i}) \ln [\ln a_{x_{i-1},c} + \ln b(y_{i}|x_{i}, h_{i}, g_t)]
\]

(20)
and \(p_\lambda(x_i, y_i, h_i | y, g)\) is the posterior probability of the states \((x_i, y_i)\) and the mixture component \(h_i\) given the signal \(y\) and the gain contour \(g\). From (20) we see that maximization of the auxiliary function \(\phi(\lambda, g)\) over \(\lambda\) makes the right-hand side of (20) nonnegative. Hence, the likelihood \(l(\lambda, g)\) associated with the parameter set estimate obtained from the maximization of \(\phi(\lambda, g)\) is greater than or equal to the original likelihood \(l(\lambda', g)\). Thus, maximization of the auxiliary function \(\phi(\lambda, g)\) provides means for reestimation of \(\lambda\) from \(\lambda'\) in a manner which increases the likelihood unless a fixed point \(\hat{\lambda} = \lambda'\) is reached. The maximization of \(\phi(\lambda, g)\) over \(\lambda\) has been considered in detail in [8], [10], [23], [26]. The reestimation formulas for \(\lambda\), given some fixed \(g\), can be obtained from [23, eqs. (18)-(21)] using the posterior probability \(p_\lambda(x_{i-1}, x_i, h_i | y, g)\) [23, eqs. (22)-(27)], where \(S_{h|0}\) in these equations is replaced by \(g; S_{h|0}\). The mapping \((\lambda', g) \rightarrow (\lambda, g)\) for a given \(g\) is a “point to set” mapping [28, p. 183], since if \(p_\lambda(x_{i-1}, x_i, h_i | y, g) = 0\) for some \((x_{i-1}, x_i, h_i)\), then \(\lambda\) cannot be uniquely estimated from \(\lambda'\) [23].

Assume next that \(\lambda\) is known. Let \(g'\) and \(g\) be a current estimate and a new estimate of the gain contour of the signal \(y_i\), respectively. Using an argument similar to (20), and (19), we obtain

\[
l(\lambda, g) - l(\lambda, g') \geq \phi(\lambda, g) - \phi(\lambda, g')
\]

where now

\[
\phi(\lambda, g) = \sum_{i, h} p_\lambda(x_i, h_i | y_i, g') \ln p_\lambda(y_i | x_i, h_i, g).
\]

\[
= \sum_{i=0}^{t-1} \sum_{x, h} p_\lambda(x_i, h_i | y_i, g') \ln b(y_i | x_i, h_i, g).
\]

Similarly to what we have seen before, maximizing the auxiliary function \(\phi(\lambda, g)\) over \(g\) results in an estimate \(g\) of the gain contour whose likelihood \(l(\lambda, g)\) is greater than or equal to the original likelihood \(l(\lambda, g')\). On substituting \(b(y_i | x_i, h_i, g)\) from (19) into (23), and setting the gradient of \(\phi(\lambda, g)\) with respect to \(g\) to zero, we arrive at the following gain reestimation formula

\[
g'(m+1) = \sum_{x, h} p_\lambda(x, h | y_i, g(m)) \frac{1}{K} y_i y_i y_i S_{h|0}^{-1} y_i
\]

where \(g(m)\) is the gain contour estimate obtained at the \(m\)th iteration.

The term \(\frac{1}{K} y_i y_i y_i S_{h|0}^{-1} y_i\) in the reestimation formula (24) is normally calculated using the following approximation

\[
\frac{1}{K} y_i y_i y_i S_{h|0}^{-1} y_i = \sum_{n=-\infty}^{\infty} r_n(n) r_{h|0}(n) / \phi_{h|0}.
\]

where

\[
r_n(n) \triangleq \frac{1}{K} \sum_{l=0}^{K-1} y_l y_l y_l (l + |n|)
\]

is the sample autocorrelation function of the vector \(y_i\) and

\[
r_{h|0}(n) \triangleq \sum_{i=0}^{\infty} \xi_{a|0}(i) \xi_{b|0}(i + |n|)
\]

is the autocorrelation function of the vector \(\xi_{a|0}\) of AR coefficients corresponding to state \(x_i\) and mixture component \(h_i\). This approximation results from the assumption that \(A_{h|0} y_i\) is the convolution of \(\xi_{a|0}\) with \(y_i\), or from neglecting the last \(N_t - 1\) samples in the \(K + N_t - 1\) sample convolution vector. The approximation is known to be satisfactory for \(K >> N_t\), e.g., for \(K \geq 128\) and \(N_t = 10\) [23], [24].

Note from (24), (25) that \(g'(m+1)\) constitutes the average of the variances of the residual signals obtained from inverse filtering of \(y_i\) by \(S_{h|0}\). In contrast with [10], no individual linear predictive analysis of the signal in each vector is performed, and the estimated gain contour is adapted to the given set of AR models. Furthermore, the gain in (24) obtained for Gaussian AR HMM’s is different from the gain used in cepstral hidden Markov modeling, where in that case it equals to \(r_{y|0}^{-1}(0)\) as explained in Section I.

The application of the EM algorithm for maximizing the likelihood (18) can be done in two different ways, which may lead to different model estimates, due to the fact that only local maximization is performed. In the first approach, we start by fixing \(g\) and perform many EM iterations for estimating \(\lambda\) using the auxiliary function (21) until a fixed point is reached. Then, we keep the resulting \(\lambda\) constant and perform many EM iterations for estimating \(g\) using the reestimation formula (24) until a gain fixed point is achieved. This process is repeated until either a fixed point \((\lambda, (m+1) = \lambda, (m), g(m+1) = g(m))\) is reached or the difference in likelihood in two consecutive iterations is sufficiently small. In the second alternative, each iteration constitutes one EM iteration for estimating \(\lambda\) for a given \(g\) and one EM iteration for estimating \(g\) for the resulting \(\lambda\). Since each EM iteration (over \(\lambda\) and over \(g\)) increases the likelihood function, unless a fixed point is reached, this approach is also an ascent approach for maximizing (18). It is difficult to predict which approach will work better when applied to speech signals, in the sense that it will converge faster to a higher local maximum of the likelihood function. We have chosen here to focus on the second alternative maximization approach.

The algorithm for gain-adapted hidden Markov modeling can be summarized as follows:

**Gain-Adapted HMM Training Algorithm:**

0) Initialization: Given \(\lambda_0(0), g(0), \) and \(\epsilon > 0\), calculate \(l(\lambda_0(0), g(0))\) and set \(m = 0\).

1) **Gain estimation:** Estimate \(g'(m+1)\) for the given \(\lambda, (m)\) and \(g(m)\) using the reestimation formula (24).

2) **Parameter estimation:** Estimate \(\lambda, (m+1)\) for the resulting \(g(m+1)\) and the given \(\lambda, (m)\) using the reestimation formulas (18)-(21) in [23].

3) **Convergence test:** If \(l(\lambda, (m+1), g(m+1)) - l(\lambda, (m), g(m)) \leq \epsilon\), stop. Otherwise, set \(m \rightarrow m + 1\) and go to (1).
Convergence of the gain-adapted training algorithm to a stationary point of \( P(x, y | g) \) can be shown using the "global convergence theorem" [27], [28], provided that for all \( m \geq 0 \), the parameters of the model \( \{ \lambda(m), a_{ij}(m), c_{ij}(m), \alpha_{ij}(m) \} \), and the gain \( \{ g(m) \} \), are strictly positive, and \( (\lambda(m), g(m)) \) is contained in a compact (or bounded) subset of the product space of the parameter set and the gain. This is done by defining the gain-adapted training algorithm, say \( A : (\lambda', g') \rightarrow (\lambda, g) \), as a composition of two algorithms [28, p. 186], \( A = A_2(A_1) \), where \( A_1 : (\lambda', g') \rightarrow (\lambda', g) \) is a point-to-point mapping, and \( A_2 : (\lambda', g) \rightarrow (\lambda, g) \) is a point-to-set mapping, and showing that the hypotheses of the global convergence theorem are satisfied. In particular we have to show that i) the algorithm \( A \) is closed [28, p. 185]. ii) If \( (\lambda', g') \) is not a stationary point of \( p_{\lambda}(y | g) \), then \( p_{\lambda}(y | g) > p_{\lambda'}(y | g') \). Since \( p_{\lambda}(y | g) \) is continuous on the product space of the parameter set and gain, and \( p_{\lambda}(y | g) \geq p_{\lambda'}(y | g') \) by construction of \( A \), then proving ii) will show that \( p_{\lambda}(y | g) \) is the ascent function for the algorithm \( A \). Assumption i) results from the facts that the algorithm \( A_1 \) is a continuous mapping, and \( A_2 \) is closed as can be shown by using the continuity of \( \phi(\lambda', g) \) in \( \lambda \) and \( \lambda' \) [28, corollary 2, p. 187]. To prove ii) it suffices to show that either \( p_{\lambda}(y | g') > p_{\lambda}(y | g) \) or \( p_{\lambda'}(y | g') > p_{\lambda'}(y | g') \) and this can be done as in the proof of [27, theorem 2]. Under these conditions the global convergence theorem states that all limit points of any instance \( \{ \lambda(m), g(m) \} \) of the algorithm \( A \) are stationary points of \( p_{\lambda}(y | g) \), and \( p_{\lambda}(y | g(m)) \) converges monotonically to \( p_{\lambda}(y | g_{\ast}) \) for some stationary point \( \lambda_{\ast}, g_{\ast} \). 

An initial estimate of the gain contour \( g(0) \) can be obtained from the normalization procedure proposed in [10]. A procedure for initial estimation of the parameter set of the HMM for non-normalized signals, using a vector quantization approach, was described in [10], [23]. This procedure can be applied here for estimating \( \lambda(0) \) from training data normalized by \( g(0) \).

A simpler alternative to estimating \( (\lambda, g) \) by maximizing (18), is to consider the joint estimation of \( (x, h, \lambda, g) \) by maximizing

\[
\ln p_{\lambda}(x, h, \lambda, g) \overset{!}{=} \ln p_{\lambda}(x, h, y | g).
\]

(28)

The latter can be accomplished by alternate maximization of \( \ln p_{\lambda}(x, h, \lambda, g) \) over \( (x, h, \lambda, g) \). For given \( \lambda(0), g \), the maximization of (28) over \( (x, h, \lambda, g) \) can be efficiently performed using the Viterbi algorithm [8], [29]. For given \( (x, h, g) \) and a current estimate \( \lambda', g' \), the likelihood (28) can be increased, unless a fixed point is reached, if \( \lambda \) is estimated by the same reestimation formulas as before [23, eqs. (18)-(21)], but now with the posterior probability given by [23, eqs. (32), (33)]

\[
p_{\lambda}(x_{t-1}, x_t, h_t | y, g)
= \begin{cases}
  1 & x'_{t-1} = x_{t-1}, x'_t = x_t, h'_t = h_t \\
  0 & \text{otherwise}.
\end{cases}
\]

(29)

Fig. 1. A block diagram of the algorithm for approximate ML estimation of HMM's for gain-normalized signals.

For given \( x, h, \lambda \) and a current estimate \( \lambda' \), the likelihood (28) can be increased, unless a fixed point is reached, if \( g \) is estimated using the reestimation formula (24) with the posterior probability \( p_{\lambda}(x_{t'}, h_{t'} | y, g') \) defined similarly to (29). Since each phase of the alternate maximization can either increase the likelihood function or keep it constant, the above procedure generates a sequence of models with nondecreasing likelihood values and an ascent optimization procedure results. A block diagram of this algorithm is shown in Fig. 1.

This simplified approach for gain-adapted hidden Markov modeling is justified by the fact that the likelihood functions in (18) and (28), normalized by \( KT \), may differ by at most \( K^{-1} \log (ML) \) [30]. The value of this bound is negligible for typical values of \( K, M, \) and \( L \) used in speech recognition, e.g., \( K = 256, M = 10, L = 1 \). In addition, for given \( \lambda, \lambda \), the most likely sequence of state and mixture components, which is estimated by the Viterbi algorithm, occurs with very high probability provided that \( K \) is sufficiently large [30]. Thus, the posterior probabilities of states and mixture components are well estimated by (29), and similar model and gain estimates are obtained with high probability using the two approaches. The major advantage of the simplified approach is that estimating \( p_{\lambda}(x_{t-1}, x_t, h_t | y, g) \) by the Viterbi algorithm is easier than by the forward-backward algorithm which suffers inherent numerical problems [8]. The amount of computation required by both the Viterbi and the forward-backward algorithms, however, is approximately the same.

### III. GAIN-ADAPTED RECOGNITION

#### A. Gain-Adapted Likelihood Estimation

In this section we apply the EM algorithm for estimating the statistics needed for performing gain-adapted rec-
ognition as defined in (9). In particular, we are interested in

$$\max_g p_s(z|g)$$  \hspace{1cm} (30)

where \( g \) is the gain contour of the clean signal. The noise model assumed here is a single state single mixture component Gaussian AR HMM of order \( N_r \). The pdf of this model is given by

$$p_s(v) = \prod_{r=0}^{T-1} p_{h_r}(v_r)$$

$$p_x(v) = \exp\left\{ -\frac{1}{2} v S^{-1} v \right\} \det^{1/2} (S)\right\}$$  \hspace{1cm} (31)

where \( \tau \cong \sigma^2_f (A_f^r A_r)^{-1} \) with \( \sigma^2_f \) and \( A_r \) defined similarly to \( \sigma^2_{h_{1|0}} \) and \( A_{h_{1|0}} \) in Section II, respectively. The vector of coefficients of the AR filter for the noise is denoted by \( \xi \), and is defined similarly to \( \xi_{h_{1|0}} \) in Section II.

Let

$$l(\lambda, g) \cong \ln p_s(z|g)$$

$$= \ln \sum_{x, h} \int p_x(x, h, y, z|g) dy$$  \hspace{1cm} (32)

where from (19) and (31) it is easy to see that

$$p_x(x, h, y, z|g) = \prod_{t=0}^{T-1} a_{x_{t-1}, x_t} b(y_t|x_t, h_t, g_t) p_h(z_t - y_t).$$  \hspace{1cm} (33)

Furthermore,

$$p_x(x, h, z|g) = \prod_{t=0}^{T-1} a_{x_{t-1}, x_t} b(z_t|x_t, h_t, g_t)$$  \hspace{1cm} (34)

where \( b(z_t|x_t, h_t, g_t) \) is the pdf of the noisy vector \( z \) given that the clean vector \( y \) was generated from state \( x \) and mixture component \( h \), and that the gain of \( y \) is \( g \). This pdf is Gaussian with zero mean and covariance matrix \( g^2 S_{h_{1|0}} + S \). Hence, \( p_s(z|g) \) is the pdf of an \( M \)-state, \( L \)-mixture component per state, Gaussian (not AR) HMM, with parameter set \( \lambda = (\pi, a, c, S, S) \). The statistics associated with \( p_s(z|g) \), in particular, the posterior probability of the state and mixture component at time \( t \) given the noisy signal \( z \) and the gain contour \( g \), \( p_s(x_t, h_t, y_t|z_t, g_t) \), can therefore be efficiently calculated similarly to \( p_h(v_t, h_t) \) using the forward-backward procedure [23, eqs. (22)–(27)].

The gradient equations of \( p_x(z|g) \) with respect to \( g \) are nonlinear. Hence, ML estimation of \( g \) is performed here iteratively using the EM algorithm. Let \( g' \) and \( g \) be a current and a new estimate of the gain contour, respectively.

Using Jensen’s inequality, and (33), we have that

$$l(\lambda, g) - l(\lambda, g') = \ln \sum_{x, h} \int \frac{p_x(x, h, y, z|g')}{p_x(x, h, y, z|g)} p_s(x, h, y, z|g) dy$$

$$\geq \sum_{x, h} \int \frac{p_x(x, h, y|z|g')}{p_x(x, h, y|z|g)} \ln \frac{p_x(x, h, y|z|g)}{p_x(x, h, y|z|g')} dy$$

$$\cong \phi(\lambda, g) - \phi(\lambda, g')$$  \hspace{1cm} (35)

where

$$\phi(\lambda, g) = \sum_{x, h} \int p_x(x, h, y|z, g') \ln p_x(y|x, h, g) dy$$

$$= \sum_{x, h} \sum_{y, h} \int p_x(x, h, y|z, g') \ln b(y|x, h, g) dy,$$  \hspace{1cm} (36)

Hence, the likelihood \( l(\lambda, g) \) can be iteratively increased by maximizing \( \phi(\lambda, g) \) over \( g \) for a given \( g' \) as explained in Section II. Using (33), we get from the gradient equations of \( \phi(\lambda, g) \) with respect to \( g \) the following gain reestimation formula:

$$g_t^2 (m + 1) = \sum_{x, h} p_x(x_t, h_t, y_t|g(m)) \frac{1}{K}$$

$$\cdot E[y^T S_{h_{1|0}}^{-1} y_t] x_t, h_t, g_t(m), y_t]$$  \hspace{1cm} (37)

where \( g(m) \) is the gain contour estimate obtained at the \( m \)th iteration. The conditional mean can be calculated as \( tr \{ R_{h_{1|0}, h_{1|0}, g(m)} S_{h_{1|0}}^{-1} \} \), where

$$R_{h, h, g, z} \cong E[y^T y|x, h, g, z]$$  \hspace{1cm} (38)

is the MMSE estimator of the sample covariance of \( y_t \) given \( x_t, h_t, g_t, z_t \). For Gaussian HMM’s for the signal and noise we have that

$$R_{h, h, g, z} = H_{h, h, g} S_{h_{1|0}} + [H_{h, h, g} z] [H_{h, h, g} z]^T$$  \hspace{1cm} (39)

where

$$H_{h, h, g} \cong g^2 S_{h_{1|0}} S_{h_{1|0}}^{-1} + S_{h_{1|0}}^{-1} g$$  \hspace{1cm} (40)

is the Wiener filter for the output process from state \( x \) and mixture component \( h \), assuming gain \( g \) for this process. Since \( S_{h_{1|0}} \) and \( S_{h} \) are positive definite, \( R_{h, h, g, z} \) is positive definite, and \( g^2 (m + 1) > 0 \), provided that \( g^2 (m) > 0 \). The latter result can be shown by specializing [31, theorem 2, p. 185] to the case of two Hermitian positive definite matrices.

The algorithm for local maximization of \( l(\lambda, g) \) over \( g \) can be summarized as follows.

**Gain-Adapted Likelihood Estimation Algorithm:**

0) Initialization: For given \( \lambda, \lambda_g, g(0) \), and \( \epsilon > 0 \), evaluate \( l(\lambda, g(0)) \) and set \( m = 0 \).

1) **Gain Estimation:** Calculate the posterior probabilities \( \{ p_x(x_t, h_t|z_t, g(m)) \} \), and the MMSE sample covariance estimates \( \{ R_{h_{1|0}, h_{1|0}, g(m)|z} \} \), for \( x_i = 1, \ldots, M, h_t = \ldots, \ldots, \ldots, M \).
1, \cdots, L, and \( t = 0, \cdots, T - 1 \), and estimate \( g(m + 1) \) using (37).

2) Convergence test: If \( l(\lambda, g(m + 1)) - l(\lambda, g(m)) \leq \varepsilon \), assign \( \max_{\lambda} l(\lambda, g) = l(\lambda, g(m + 1)) \) and stop. Otherwise, set \( m \to m + 1 \) and go to 1.

Note that if \( S = 0 \), then \( z = y \) and gain-adapted recognition of clean speech is performed. In this case,

\[
p_\mu(x, h_0 | z, g(m)) = p_\mu(x, h_0 | y, g(m)) \quad R_{\mu,\nu,\gamma}(n, z_0) = y_0 y_0^*.
\]

Hence, (37) becomes

\[
g^2(m + 1) = \sum_{x, h_0} p_\mu(x, h_0 | y, g(m)) \frac{1}{K} y_0 y_0^* S^H_{h_0 h_0} y_0
\]

which coincides with the gain reestimation formula (42) for clean signals.

Convergence of the iterative algorithm for gain adaptation (37) can be shown using [27, theorem 2]. Specifically, if for all \( t \) and \( m \), \( g(m) \) is strictly positive and is contained in a compact subset of the real line, then the limit points of any instance \( \{ g(m) \} \) are stationary points of \( p_\mu(z | g) \) and \( p_\mu(z | g(m)) \) converges monotonically to \( p_\mu(z | g^*) \) for some stationary point \( g^* \). This results from continuity and differentiability of \( p_\mu(z | g) \) for any \( \{ g_i > 0 \} \) and continuity of \( \phi(\lambda, g) \) in (36) in both \( g \) and \( g^* \) for any strictly positive \( \{ g_i \} \) and \( \{ g_j \} \).

A useful alternative to maximizing \( \ln p_\mu(z | g) \) over \( g \) in performing recognition, is to consider the joint maximization of

\[
\ln p_\mu(x, h, z | g) = \int p_\mu(x, h, y, z | g) dy
\]

over \( (x, h, g) \). This approach is analogous to the simplified training procedure which results from the maximization of (28), and it can be similarly justified. The joint maximization of (43) can be accomplished by alternate maximization of \( \ln p_\mu(x, h, z | g) \) once over \( (x, h) \) assuming \( g \) is given and then over \( g \) assuming \( (x, h) \) is available. The iterative procedure is terminated when the likelihood values in two consecutive iterations are sufficiently close. The maximization of \( \ln p_\mu(x, h, z | g) \) over \( (x, h) \) for a given \( g \) is performed by the Viterbi algorithm. Given \( (x, h) \) and a current gain contour estimate \( g^* \), a new estimate which increases the likelihood function, unless a fixed point is reached, can be derived similarly to (37) using the EM algorithm. The auxiliary function for this case is given by

\[
\phi(\lambda, g) = \int p_\mu(y | x, h, z, g') \ln p_\mu(x, h, y, z | g) dy
\]

and the gain reestimation formula is given by

\[
g^2(m + 1) = \frac{1}{K} \text{tr} \left\{ R_{\mu,\nu,\gamma}(n, z_0) S^H_{h_0 h_0} \right\}.
\]
(37) can be approximated as

$$g_i^2(m + 1) = \sum_{x_n, h_i} p_i(x_n, h_i | z, g(m))$$

$$\times \frac{1}{K} \sum_{k = 0}^{K-1} \left| Y_{i,k}(x_n, h_i, g(m), z_i) \right|^2 f_{h_i/k}(2\pi k/K)$$

(52)

where

$$\left| Y_{i,k}(x_n, h_i, g(m), z_i) \right|^2 \triangleq w_{x_n, h_i, g}(2\pi k/K) f_{h_i/k}(2\pi k/K)$$

$$+ w_{x_n, h_i, g}(2\pi k/K) Z_{i,k}$$

(53)

and $Z_{i,k}$ is the $k$th component of $Z_i$. Note that

$$\left| Y_{i,k}(x_n, h_i, g(m), z_i) \right|^2 = E\{ \left| Y_{i,k} \right|^2 | x_n, h_i, g(m), z_i \}$$

(54)

where $Y_{i,k}$ is the $k$th component of $K^{-1/2} U y_i$, the complex conjugate of the normalized DFT of $y_i$. Hence, $\left| Y_{i,k} \right|^2$ is the MMSE estimator of the sample spectrum of the clean vector $y_i$ given the noisy vector $z_i$ and the current estimate of the gain, assuming that $y_i$ was generated from state $x_n$ and mixture component $h_i$. Applying Parseval's theorem to (52) gives

$$g_i^2(m + 1) = \sum_{x_n, h_i} p_i(x_n, h_i | z, g(m))$$

$$\times \sum_{n = -N_i}^{N_i} r_{g_m,k}(n) = \frac{\sigma_{h_i|z}}{\sigma_{h_i|z}}$$

(55)

where $r_{g_m,k}(n)$ obtained from the inverse DFT of (53) is the MMSE estimator of the sample autocorrelation of $y_i$ given $(x_n, h_i, g(m), z_i)$.

The circulant approximation (49) is also useful in evaluating the posterior probability $p_i(x_n, h_i | z, g)$. In particular, it is useful in approximating the inverse and determinant of $g_i^2 S_{h_i|z} + S$, which are needed in evaluating $b(z_i | x_n, h_i, g)$ [22].

The frequency domain implementation of the gain reestimation formula (45) is obtained similarly to (55). Specifically, we have that

$$g_i^2(m + 1) = \sum_{n = -N_i}^{N_i} r_{g_m,k}(n) = \frac{\sigma_{h_i|z}}{\sigma_{h_i|z}}$$

(56)

A block diagram of the algorithm for maximizing the gain-adapted likelihood function (43) using the gain reestimation formula (56) is shown in Fig. 2.

C. Initialization

An initial estimate of $g$ can be obtained as follows. Assume first that the clean signal is available for recognition. In this case, we could first find for each vector $y_i$ the nearest neighbor AR model form the parameter set $X_i$ using the Itakura–Saito distortion measure [33], [34], and then calculate the gain with respect to that model. From (24), (25), the contribution of the AR model $S_{h_i|z}$ to the gain is given by the variance of the residual signal obtained from inverse filtering of $y_i$ by that model. This variance is given by

$$\rho_{h_i|z}(h_i | x_i) = \frac{1}{\sum_{n = -N_i}^{N_i} r(n) r(n) / \sigma_{h_i|z}}$$

(57)

Since the AR models in $h_i$ correspond to gain-normalized signals, $\sigma_{h_i|z} = 1$, and minimizing the Itakura–Saito distortion measure is equivalent to minimizing the variance in (57) [33], [34]. Hence, $g_i^2(0)$ is obtained from

$$g_i^2(0) = \min_{h_i|z} \rho_{h_i|z}(h_i | x_i).$$

(58)

When only noisy signals are available, we estimate $r(n)$ by the correlation subtraction approach and plug the resulting estimator $\rho_{h_i|z}(h_i | x_i)$ into (58). Specifically, let $r(n)$ be the autocorrelation function of the noise process obtained from the model for this process. Using $r(n) = r_x(n) - r(n)$, and $1 - u = 1/(1 + u)$ for sufficiently small $u$, we obtain

$$\rho_{h_i|z}(h_i | x_i) = \rho_{h_i|z}(h_i | x_i) - \rho_{h_i|z}(h_i | x_i)$$

$$= \rho_{h_i|z}(h_i | x_i) \left( 1 - \rho_{h_i|z}(h_i | x_i) \right) \rho_{h_i|z}(h_i | x_i)$$

$$= \rho_{h_i|z}(h_i | x_i) \rho_{h_i|z}(h_i | x_i) + \rho_{h_i|z}(h_i | x_i).$$

(59)

IV. EXPERIMENTAL RESULTS

The gain-adapted hidden Markov modeling approach was tested in recognition of clean and noisy speech signals using computer generated Gaussian white noise at input SNR greater than or equal to 5 dB. Two sets of experiments were performed. In the first set, the gain-adapted approach was applied to training and test signals
which were recorded under similar gain conditions. Thus, gain adaptation due to quasi-stationarity of speech signals was studied (see Section I). In the second set of experiments, the gain contour of the test clean data was modified to simulate recording gain mismatch, and gain adaptation due to both quasi-stationarity of speech signals and recording conditions was studied. The experiments in both sets were performed using the approximate gain-adapted recognition approach described in Figs. 1 and 2. The reasons for using this approach rather than the exact gain-adapted approach, which consists of the gain-adapted HMM training algorithm and the gain-adapted likelihood estimation algorithm (see Sections II, III), are that the two approaches perform similarly (as we shall demonstrate here) but the approximate approach is easier to implement.

The gain-adapted recognition approach was compared with a "gain-normalized" approach and with a "gain-trained" approach. In principle, in the gain-normalized approach, training and recognition are performed using fixed estimates of the gain contours of the training and test data. Training is performed using the gain contour estimate proposed in [10]. Recognition is performed by maximizing (43) for the gain contour obtained from (58), (59). Since the gain-normalized approach performed similarly to the gain-adapted approach in recognition of clean signals, however, we have chosen to compare the two approaches using the same set of HMM’s designed by the gain-adapted training approach. Thus, the difference in performance of the two approaches due to the different gain contour estimates used during recognition is emphasized. Furthermore, since (58), (59) constitute the initial gain contour estimate for the gain-adapted recognition approach, this comparison of the two approaches will demonstrate the effectiveness of the EM algorithm in performing the gain adaptation during recognition. In the gain-trained approach, training and recognition are performed using HMM’s and gain-contours exclusively estimated from the training data. Specifically, training and recognition are performed by maximizing (28) and (43), respectively, using gₜ = 1 for all t. In this approach any possible gain mismatch between the training and the test data is simply ignored.

The three recognition tests obtained using the gain-adapted, gain-normalized, and gain-trained estimation approaches, were applied to acoustic signals corresponding to the English digits and E-set words at input SNR greater than or equal to 5 dB. The data base we used contains speech signals from 4 speakers, two males and two females, where for each word and speaker, 5 utterances are available for training and 10 utterances for recognition. Training of the HMM for each word was performed using all 20 utterances from the 4 speakers simultaneously, and recognition of each word was applied to all 40 utterances from the 4 speakers. Thus, multispeaker recognition tests were performed. Training and recognition were performed on nonoverlapping vectors of the speech utterances whose dimension was K = 256. The order of the output AR processes of the HMM for each word was Nₜ = 10, and the order of the AR model for the noise process was Nₙ = 4. The number of states and mixture components used for each recognition task was experimentally determined. We obtained best results using M = 10 and L = 1 for the digits vocabulary, and M = 8 and L = 5 for the E-set words. The AR Gaussian model for the noise was estimated from an initial interval of each utterance of speech (20 frames) which was known to contain only noise samples.

Tables I-IV show the results obtained in the first set of experiments in which the training and test sequences did not suffer from recording gain mismatch. Table I shows the average accuracy and the standard deviation in digit recognition obtained using the three gain estimation approaches.

The results shown in Table I indicate that the gain-adapted recognition approach performs similarly to the gain-normalized recognition approach when clean signals are available for recognition. However, significantly better results were obtained by the gain-adapted approach when recognition is performed on noisy signals. The gain-adapted approach also outperformed the gain-trained approach at input SNR greater than or equal to 10 dB. In the latter case the gain-adapted approach provided 2%-13% higher recognition accuracy at 10 dB –∞ input SNR, respectively. At the low SNR of 5 dB, the gain-trained approach outperformed the gain-adapted approach by 3%. Note that the 13% improvement in recognition accuracy corresponds to over 96% reduction in the error rate as obtained from (0.135–0.005)/0.135. In the other extreme, the loss of 3% in recognition accuracy at 5-dB SNR corresponds to an increase of 23% in the error rate.

The higher performance obtained using the gain-adapted approach compare to the gain-normalized results from the fact that the gain contour of the clean test signal used in the latter approach is the initial estimate of the gain contour used by the former approach, and this estimate is iteratively improved in the gain-adapted recognition approach. The comparison of the gain-adapted approach with the gain-trained approach is more interesting. Here, the results show that the gain-adapted approach provided significantly higher recognition accuracy when applied to clean signals. Furthermore, a threshold in SNR exists, below which better recognition results were ob-
for gain-adapted recognition of clean speech and 10–20 for gain-adapted recognition of noisy speech at 10-dB SNR.

A surprising result from Table I is that the gain-trained approach performs better at 30-dB input SNR than when the input signal is clean. To further illustrate this phenomenon, Table II shows additional recognition results obtained by applying the gain-trained approach at input SNR greater than or equal to 30 dB.

The results in Table II show a systematic improvement in recognition performance as the input SNR decreases from ∞ to 30 dB. Since for all SNR values examined in Table II the signal is essentially noise-free, this behavior must be attributed to the different pdf's used in performing recognition of clean and noisy signals, i.e., $p_n(x, h, y | g)$ and $p_n(x, h, z | g)$, respectively. From (19) and (34), these two pdf's are the pdf's of HMM's which differ only in their covariance matrices. The covariance matrices of $p_n(x, h, y | g)$ are given by $\{S_{11}\}$, and the covariance matrices of $p_n(x, h, z | g)$ are given by $\{S_{11}, S_{12}\}$. The fact that better results were obtained at 30-dB SNR than at ∞ implies that the model for the clean signal can be improved by slightly contaminating it with white noise. This has a similar effect as modifying the gain contour of the model. To support this conjecture, we performed recognition of the clean speech signals by maximizing (43), pretending that the signal is noisy with 30-dB SNR. For this experiment we obtained a recognition accuracy of 92.75% with standard deviation of 5.17, a significantly better performance than that obtained using the original model for the clean signal.

Table III shows results similar to those given in Table I for the English E-set words. Here the estimation of the gain contour of the clean test signal from the noisy signal is more difficult than in the case of digit recognition, since the acoustic signals from the different E-set words are much more confused. Hence, the gain-adapted approach performed better than the gain-trained approach only at high SNR. Specifically, an elevation of recognition accuracy of 4%–16% was obtained by the gain-adapted approach at SNR greater than or equal to 30 dB. For the lower SNR range of 10–20 dB, the gain-trained approach performed slightly better than the gain-adapted approach. At the low SNR of 5 dB, the gain-trained approach provided recognition accuracy which is higher than that obtained by the gain-adapted approach by 9%.

In Table IV we compare the performance of the approximate and the exact gain-adapted approaches in recognition of the digit vocabulary. The results show that the two approaches provide practically identical results as predicted by the theory discussed in Section II.

We now describe the results obtained in the second set of experiments in which the same data base was used, but the gain contour of each utterance of the test data has been modified by the multiplicative gain function given by

$$1 + \mu \sin \frac{2\pi}{T} t, \quad t = 0, \cdots, T - 1$$  (60)
where \( \mu \) is a modulation index which was chosen to be equal to \( \mu = 0.2 \). This slowly varying gain function constitutes one model for gain mismatch, and other models are of course possible.

Fig. 3 shows a typical example of gain contour estimation performed by the three gain estimation approaches considered here. The utterance used in this example corresponds to the word "zero" in the digit vocabulary, and estimation was performed from the noisy utterance at 10-dB input SNR. In Fig. 3(a), the logarithm of the energy contour \( \ln r_e(0) \) of the utterance is shown. The dash lines in Fig. 3(b)–(d) represent the logarithm of the gain contour of the utterance \( \ln \sigma^2 \). The solid lines in these figures represent the logarithm of the estimates of the gain contour \( \ln \sigma^2 \) obtained from using the gain-normalized, the gain-trained, and the gain-adapted estimation approaches, respectively. These figures show that the gain-adapted approach provides the best estimate of the gain contour of the clean signal in the region where most of the energy of the utterance is concentrated. Since high energy regions of the signal dominate its likelihood function, accurate estimation of the gain in these regions is crucial for MAP recognition which is based upon the likelihood ratio test. Hence, best performance is expected from the approach which provides the most accurate gain contour estimation in those high energy regions. This indeed turns out to be the case here, as is demonstrated in the next set of experiments.

Table V shows recognition results similar to those in Table I for test signals whose gain contours have been perturbed by the modulation function (60). The results indicate that the gain-adapted approach performs significantly better than the gain-trained approach at SNR greater than or equal to 10 dB. Furthermore, the gain-adapted is more robust than the gain-trained approach, as it provided recognition accuracy similar to that obtained in Table I where no recording gain mismatch existed, while the performance of the gain-trained approach dropped at all SNR’s to the level obtained in Table I for input SNR of 5 dB. Similarly to what we have seen in Table I, the gain-
TABLE V
AVERAGE (AVE.) RECOGNITION ACCURACY AND STANDARD DEVIATION (SD) IN DIGIT RECOGNITION WITH THE TIME-VARYING RECORDING GAIN MISMATCH (60) AT SEVERAL INPUT SNR VALUES

<table>
<thead>
<tr>
<th>SNR</th>
<th>Gain Trained Ave.</th>
<th>Gain Normalized Ave.</th>
<th>Gain Adapted Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ave.</td>
<td>sd</td>
<td>Ave.</td>
</tr>
<tr>
<td>5</td>
<td>87.00</td>
<td>7.81</td>
<td>39.25</td>
</tr>
<tr>
<td>10</td>
<td>87.50</td>
<td>4.60</td>
<td>54.75</td>
</tr>
<tr>
<td>15</td>
<td>87.75</td>
<td>7.45</td>
<td>68.50</td>
</tr>
<tr>
<td>20</td>
<td>86.75</td>
<td>9.75</td>
<td>76.25</td>
</tr>
<tr>
<td>30</td>
<td>86.75</td>
<td>11.56</td>
<td>92.00</td>
</tr>
<tr>
<td>∞</td>
<td>78.50</td>
<td>19.11</td>
<td>98.50</td>
</tr>
</tbody>
</table>

adapted approach provided here significantly better results than those obtained by the gain-normalized approach at all input SNR’s.

V. COMMENTS

We developed a gain-adapted approach for recognition of clean and noisy speech signals which is based upon Gaussian hidden Markov modeling of the clean signals, and a single state single mixture component Gaussian hidden Markov modeling of the noise process. Training in this approach comprises ML estimation of HMM's for gain-normalized clean signals. Recognition is performed by applying the MAP decision rule to the noisy signals, using the HMM’s for the gain-normalized clean signals, and ML estimates of the gain contours of the clean signals as obtained from the given test utterances. The proposed approach does not require estimation of the clean signal from the noisy signal. Only the gain contour of the clean signal is estimated from the noisy signal. The gain-adapted approach was efficiently implemented in the frequency domain using the EM algorithm. This recognition approach can be easily extended to noise models with multiple states and mixture components per state [7].

The gain-adapted speech recognition approach was proven useful in recognition of clean and noisy speech signals, even when the training and test data were recorded under similar gain conditions, as in this case gain adaptation is necessary due to the quasi-stationarity of speech signals. It has been demonstrated that the gain-adapted approach elevated the recognition accuracy of the English digits at input SNR greater than or equal to 10 dB, compared to the recognition accuracy obtained by the gain-trained approach, by 2%–13% when the training and test data were recorded under similar gain conditions, and by 5%–21% when a recording gain mismatch existed. A SNR threshold exists, however, below which estimation of the gain contour of the signal from the training data rather than from the noisy signal is preferable. The threshold depends on the statistical variability between the training and test data and on the noise level.

The gain adaptation approach was extended in [35] for robust recognition of speech signals, by treating the models estimated from the training data as “nominal” models, and performing recognition using new models which are constrained to some neighborhood of the old models and their parameters are estimated from the test data. This approach provided 2%–14% higher recognition accuracy compared to the standard non-adaptive recognition approach when applied to digit recognition under different conditions of mismatch between the training and test data.

ACKNOWLEDGMENT

The author thanks L. R. Rabiner, D. Malah, and N. Merhav for many discussions during this work. He also thanks the anonymous referees for many helpful comments which improved the presentation.

REFERENCES


Yariv Ephraim (S'82-M'84-SM'90) received the B.Sc., M.Sc., and D.Sc. degrees from the Technion, Israel Institute of Technology, in 1977, 1979, and 1984, respectively, all in electrical engineering.

During 1984-1985 he was a Rothschild Post-doctoral Fellow at the Information Systems Laboratory of Stanford University, Stanford, CA. Since 1985 he has been a Member of Technical Staff at the Speech Research Department of AT&T Bell Laboratories, Murray Hill, NJ.