Integrated Models of Signal and Background with Application to Speaker Identification in Noise

R. C. Rose, E. M. Hofstetter, and D. A. Reynolds

Abstract—**This paper is concerned with the problem of robust parametric model estimation and classification in noisy acoustic environments. Characterization and modeling of the external noise sources in these environments is in itself an important issue in noise compensation. The techniques described here provide a mechanism for integrating parametric models of acoustic background with the signal model so that noise compensation is tightly coupled with signal model training and classification. Prior information about the acoustic background process is provided using a maximum likelihood parameter estimation procedure that integrates an *a priori* model of acoustic background with the signal model. An experimental study is presented in the paper on the application of this approach to text-independent speaker identification in noisy acoustic environments. Considerable improvement in speaker classification performance was obtained for classifying unlabeled sections of conversational speech utterances from a 16-speaker population under cross-environment training and testing conditions.**

I. Introduction

THE growing need for automation in complex work environments and the increased need for voice-operated services in many commercial areas have motivated recent efforts in reducing laboratory speech-processing algorithms to practice. While many existing systems for speech recognition, word spotting, and speaker identification have demonstrated good performance in relatively constrained environments, performance invariably deteriorates in more difficult noisy environments. The same disparity in performance between laboratory and field applications exists in other areas, including ocean acoustic event classification and optical character recognition. One of the biggest obstacles to operating speech processing equipment in useful applications is the presence of acoustic noise. In many of these applications, acoustic noise can be nonstationary and may depart considerably from traditional broadband or impulsive noise models.

The purpose of this paper is to develop a comprehensive approach to dealing with the effects of acoustic noise in speech processing applications. It is based on the premise that acceptable performance in these difficult acoustic environments will require more robust statistical models of external acoustic sources. However, in addition to developing robust statistical models to characterize external noise sources, it is also necessary to develop a statistical formalism that allows models for external noise sources to be integrated with the model for speech. The goal of this paper is to describe a formalism for integrating models of signal and background that has potential application to a wide variety of signal classification problems. A study is presented describing the application of the techniques to text independent speaker identification from conversational speech utterances that have been corrupted by noise.

A method is described that exploits prior information about background noise to obtain more robust estimates of the parameters of Gaussian mixture classifiers. Prior information about background sources is obtained by providing a mechanism for integrating a broad class of statistical models of background directly with the model for the underlying signal. Gaussian mixtures are investigated in particular because of their ability to provide smooth approximations to arbitrarily shaped underlying densities. When applied to speech processing applications, they have demonstrated good performance as parametric speaker models in speaker identification classifiers [28]–[30], and as observation distributions in hidden Markov model (HMM) speech recognizers [2], [17]. They are also of interest here because of the ease with which the associated theoretical development can be generalized to more structured statistical models like HMM's that are commonly used in speech recognition [26].

Analysis of speech recorded in several different environments has served to illustrate the problems posed by the associated ambient acoustic conditions. Dobroth *et al.* [9] analyzed speech recorded over the telephone from individuals seeking assistance from telephone customer service operators. It was observed that the background environment for 39.3% of conversations contained competing speech, music, or traffic noise. Kishner *et al.* [20] performed a study of voice recordings from a high performance jet aircraft cockpit. The background noise was found to be highly nonstationary with a dynamic range of more than 30 dB, and speech signal-to-noise ratio (SNR) was found to be as low as 5 dB. Signal identification in other domains, such as ocean acoustic event classification, is also complicated by nonstationary background ambient acoustic conditions that are often very similar in character to the desired signal [3].

Current approaches to improving the robustness of speech processing algorithms in noise have been focused in three areas. The first set of approaches seeks to improve the SNR at the input to the processor by improving the characteristics of

Manuscript received June 9, 1992; revised July 2, 1993. The associate editor coordinating the review of this paper and approving it for publication was Dr. Amro El-Jaroudi

R. C. Rose is with the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974-0636.

E. M. Hofstetter and D. A. Reynolds are with the Massachusetts Institute of Technology, Lexington, MA 02173-9108.

IEEE Log Number 9215231.

1063-6676/94$04.00 © 1994 IEEE

the speech transducer and using noise cancellation techniques. Microphone arrays followed by beam forming algorithms have been used to improve speech SNR in large rooms [11], office environments [33], and automobiles [25]. Of course, these techniques are not applicable to those applications such as speech input from a telephone, where the system designer has no control over the choice of microphone.

The goal of the second set of approaches is to improve the SNR using one of a set of noise preprocessing techniques. In noise prefiltering, a spectral representation is obtained for a frame of noisy speech, and each spectral component is modified to remove the effects of the estimated noise power [4], [6]. Projection based distortion measures applied to cepstrum domain observations reduce the sensitivity with respect to noise-induced variability [21]. Signal processing techniques that involve vector space mapping do not attempt to estimate the characteristics of the noise source. Instead, they relate the clean signal to the noisy signal through a vector space mapping trained from utterances that have been observed in both the clean and noise corrupted environments [12], [18].

The last set of approaches for improving the robustness of a speech processing system integrates the process of noise compensation directly into the system. These approaches do not attempt to improve the SNR at the input to the speech processor, but instead incorporate knowledge of the background process when estimating model parameters from noisy observations or performing signal classification in noise. An example of one integrated approach to noise compensation is noise masking, where the goal is to simulate identical noise conditions for observations during both training and recognition by adding an appropriate noise mask level to all observations [15], [19], [36]. A minimax approach to robust speech recognition compensates for mismatch between training and test conditions by modifying the form of the decision rule used in the recognizer. The mismatch is compensated for during recognition by allowing model parameters to occupy a prescribed neighborhood surrounding the parameters estimated during training [22]. A technique for gain adaptation of noisy speech signals in speech recognition attempts to match the energy contour of the signal with the energy contour of the model for that signal [10]. This is done by combining HMM models of speech directly with estimates of the clean speech gain contours during speech recognition. The set of techniques described in this paper is an example of this integrated approach applied in a probabilistic framework.

This work builds on previous efforts in noise robust speech recognition. The problem of estimating the parameters of a Gaussian mixture from speech observations corrupted by broadband Gaussian noise was originally investigated by Nadas *et al.* [23]. The general formulation presented in this paper rederives their approach from first principles. Furthermore, it extends the approach both to allow for a more general definition of the interaction between signal and background and also to allow for more general models of acoustic background. Varga and Moore [35] investigated techniques for decomposition of speech and noise by modeling the noisy signal as the combined output of separate simultaneous HMM's for speech and noise. They achieved excellent digit recognition performance in nonstationary noise environments when HMM word models were trained in a noise-free environment. The development presented here is for estimating signal model parameters from noisy observations and can easily be generalized to HMM's.

The techniques described in this paper have been successfully applied to two separate applications. The first application was text independent speaker identification from unlabeled conversational utterances corrupted by competing speech [28], [29]. Gaussian mixtures, defined over observation vectors formed from the log energies of filterbank channels, were used to model target speaker identity. Improved estimates of speaker model parameters were obtained and significant improvement in speaker identification performance was obtained using integrated models for speech and background. The second application was not related to speech signals, but instead to ocean acoustic event classification [16]. HMM parameters for synthetic wideband acoustic pulses were estimated in the presence of ocean acoustic background. Promising results were obtained for detecting synthetic pulses in the ocean environment.

The outline of the paper is as follows. The next section introduces the integrated model for signal and background and describes a probabilistic formulation for estimating model parameters in the presence of noisy background. This formulation is general in that the expressions obtained for the model parameter estimates are not tied to a particular mode of interaction between signal and background (additive, multiplicative, max operator, etc.), nor are they tied to a particular structure for the model of acoustic background. Section III describes the application of the technique to some specific noise corruption models. Section IV compares the approach presented in this paper with noise masking, another technique where noise compensation is integrated directly into the classifier. An experimental study of the application of this approach to text independent speaker identification is described in Section V. Finally, discussion and summary are provided in Section VI.

II. Mixture Densities with Integrated Noise Model

The purpose of this section is to introduce the integrated model of signal and background and to describe a general procedure for estimating the signal model parameters from noise corrupted observations. The procedure is first motivated as a means for taking advantage of prior knowledge of the background characteristics to obtain better estimates of the signal model parameters in noise. Following that, a general technique for maximum likelihood estimation of the model parameters is described. Both the model and the procedure for estimating the parameters of the model are very general in that the model parameters can be expressed in terms of a general noise corruption process. While the formulation is very general, the procedure is limited in the specific types of noise corruption models that can applied. In Section III, the expressions for the signal model parameters are used to obtain the expressions for some specific examples of noise corruption processes.

A. Speech-Background Model

It is important to incorporate prior knowledge of external noise sources when considering the two problems of maximum likelihood decoding and signal model parameter estimation in noisy acoustic environments. In both of these problems, prior knowledge of external noise sources is obtained through an explicit parametric model of these sources. The first problem, maximum likelihood decoding, is to find the signal model that is most likely to have generated a set of noisy observations. In addressing the first problem, we begin with parametric model representations for the signal source λ_s and a parametric model representation for the acoustic background λ_b. It is often assumed that all models are estimated from independent measurements so that the interaction between the signal and background processes are not considered an issue in training. The second problem, signal model parameterl estimation, is to estimate the signal model parameters from noise corrupted observations where the interaction between the signal and background processes is considered an important issue. It is assumed in parameter estimation that the background model parameters are estimated from independent observations so that prior knowledge of the background process is assumed to exist.

Of the two problems of maximum likelihood decoding and maximum likelihood estimation of signal model parameters with an integrated model of the background, we are more concerned with the latter. It is often the case that the only available observations for a signal source are taken from a noisy environment, and it may further be the case that signal classification is to be performed in a different noisy environment with completely different background characteristics. In these applications, it would be desirable to be able to separate the effects of the noisy acoustic background from the signal when estimating signal model parameters. The following discussion will describe a Gaussian mixture model whose output vectors are observed in the presence of a noise process.

The signal source is modeled as an independent sequence of random vectors $\vec{x}$ with a pdf that is a mixture of M state pdf's $b_i(\vec{x})$ combined with state probabilities p_i for $i = 1, \cdots, M$. This situation is depicted in Fig. 1. If viewed as a generative process, the model in Fig. 1 generates a sequence of hidden states, $I = (i_1, i_2, \cdots, i_T)$, where i_t denotes an integer between 1 and M, and this sequence of states is turned into a set of D-dimensional signal vectors, $X = (\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_T)$ through a set of state-dependent continuous probability densities. The density for a particular model λ_s is represented as

$$p(\vec{x}|\lambda_s) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{1}$$

where each $b_i(\vec{x})$ is a continuous probability density function and p_i is the pdf associated with the hidden sequence of states. Both the sequence of states and the sequence of signal vectors are assumed to be independent in time, which means that the joint probability density for the sequence X can be expressed as

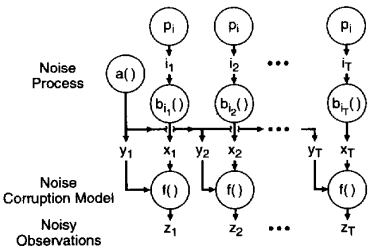


Fig. 1. Probabilistic mixture density with integrated noise process. Hidden signal values x_t generated by signal densities $b_{i_t}(\,)$ are combined with hidden background values y_t generated by background noise density $a(\,)$ according to a general function of signal and noise, $f(x, y)$.

$$p(X|\lambda_s) = \prod_{t=1}^{T} p(\vec{x}_t|\lambda_s). \tag{2}$$

The densities $b_i(\vec{x}_t)$ are assumed here to be jointly Gaussian, although the following discussion could easily be expanded to include a larger class of exponential densities. It is further assumed that the joint Gaussian densities have diagonal covariance matrices, which implies that the $b_i(\vec{x})$ are completely defined by their means $\mu_i[d]$ and variances $\sigma_i^2[d]$, where $i = 1, \cdots, M$ and $d = 1, \cdots, D$. In (1), there is an implicit dependence on the signal model parameters that include the acoustic state probabilities and the means and variances for each of the component Gaussian densities

$$\lambda_s = \{p_i, \mu_i[d], \sigma_i[d]\}, \qquad i = 1, \cdots, M \qquad d = 1, \cdots, D. \tag{3}$$

In the following discussion, an arbitrary vector component of a vector $\vec{x}_t$ will be represented simply as x_t with the vector index dropped for notational simplicity.

Following Nadas *et al.*, the notion of a probabilistic mixture density can extended by including a noise process as shown in Fig. 1 [23]. The noise process in this model is assumed to consist of time-independent identically distributed D-dimensional random vectors with probability-density function $p(\vec{y}|\lambda_b)$. Furthermore, $p(\vec{y}|\lambda_b)$ is assumed to be Gaussian with means $\mu_b[d]$ and diagonal covariance matrix with variances $\sigma_b^2[d], d = 1, \cdots, D$. The parameters of the background noise process are assumed known from previous measurements. Finally, the output sequence $Z = (\vec{z}_1, \vec{z}_2, \cdots, \vec{z}_T)$ is observed through some component-wise function of the signal and noise, $\vec{z}_t = f(\vec{x}_t, \vec{y}_t)$.

The single Gaussian background model shown in Fig. 1 is insufficient for modeling the difficult noisy environments described in Section I. It is necessary to provide a more flexible parametric representation of acoustic background. The integrated signal background model can be made more general by allowing the background process to be represented by a mixture of Gaussians as is illustrated in Fig. 2. In this case, there is also a hidden sequence of background states, $J = (j_1, j_2, \cdots, j_T)$, where j_t denotes an integer between 1 and N, and the sequence of background states is turned into a set of D-dimensional signal vectors, $Y = (\vec{y}_1, \vec{y}_2, \cdots, \vec{y}_T)$ through a set of state-dependent probability densities. The

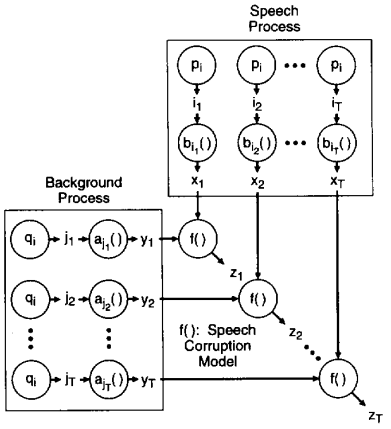


Fig. 2. Integrated speech-background speaker model with background process represented using a mixture of Gaussians.

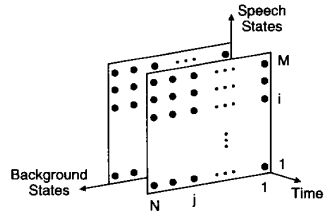


Fig. 3. Signal-background state space lattice formed from M signal states and N background states.

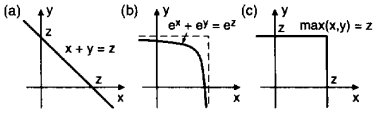


Fig. 4. Contours of integration for (a) additive noise model, (b) additive noise model with log domain observations, and (c) max noise model.

background noise density can be expressed as

$$p(\vec{y}|\lambda_b) = \sum_{j=1}^{N} q_j a_j(\vec{y}) \tag{4}$$

where $a_j(\vec{y})$ is a continuous probability-density function associated with background and q_j is the probability density function associated with the hidden sequence of states. Again, it is assumed that the parameters of the background pdf are obtained from observation vectors known to contain only samples of the background process.

Since each observation $\vec{z}_t$ is formed from a general function of speech and background $f(x_t, y_t)$, the observed signal density for state (i_t, j_t) is given as

$$p(\vec{z}_t|i_t, j_t, \lambda_s, \lambda_b) = \int\int_{C_t} b_{i_t}(\vec{x}_t) a_{j_t}(\vec{y}_t) d\vec{x}_t d\vec{y}_t \tag{5}$$

where C_t denotes the contour defined by $z = f(x, y)$. The functions $f(\,)$ to be considered here are not arbitrary but confined to those for which the equation $z = f(x, y)$ defines a 1-D contour in the x- y plane. For example, if the noisy observation sequence were an additive function of signal and background, the observation density in (5) would be obtained by integrating along the contour in Fig. 4(a).

The total likelihood of a noisy observation vector for the integrated signal-background model is a function of the combined signal and background densities. The observation likelihood for the simple M-state mixture density without considering background is given by (1), since the signal vector is assumed to be directly observable. However, for the model of Fig. 2 the signal observations $\vec{x}_t$ are hidden, and the likelihood of a noise corrupted observation $\vec{z}_t$ is given as

$$p(\vec{z}_t|\lambda_s, \lambda_b) = \sum_{i=1}^{M}\sum_{j=1}^{N} p_i q_j p(\vec{z}_t|i, j, \lambda_s, \lambda_b). \tag{6}$$

Equation (6) is the total observation probability computed over an entire $M \times N$ signal-background state space lattice illustrated by the diagram in Fig. 3. In the following, as a notational shorthand, reference to λ_b will be dropped and the signal model will be referred to simply as λ, where $\lambda = \{p_i, \vec{\mu}_i, \vec{\sigma}_i\}, i = 1, \cdots, M$.

B. Estimating Model Parameters

The goal of this section is to obtain estimates of the parameters of an underlying signal model by maximizing the likelihood of a set of noisy observations, Z, shown in Fig. 2. The maximum likelihood parameter estimation procedure is based on the use of the auxiliary Q function introduced by Baum, Petrie, *et al.* [1] for finding a model λ that improves the likelihood of the data sequence $P(Z|\lambda)$. In Section II-B-1, the form of the auxiliary Q function corresponding to the corrupted signal source shown in Fig. 2 is derived. In Section II-B-2, general expressions for the signal model parameters are obtained by maximizing the auxiliary function with respect to the model parameters. Simply observing the form of the expressions for p_i, μ_i, and σ_i in Section II-B-2 provides considerable insight into how the estimation procedure deals with the interaction between signal and background. Finally, the iterative parameter estimation procedure is described in Section II-B-2 as a special case of the expectation-maximization (EM) algorithm [8].

Estimating the model parameters in the maximum likelihood (ML) sense involves finding the model λ that maximizes $p(Z|\lambda)$ where

$$P(Z|\lambda) = \sum_{I}\sum_{J}\int\int_{C} P(X, Y, I, J|\lambda) dX dY \tag{7}$$

and

$$P(X, Y, I, J|\lambda) = \prod_{t=1}^{T} b_{i_t}(\vec{x}_t) p_{i_t} a_{j_t}(\vec{y}_t) q_{j_t}. \tag{8}$$

In (7), the double summation is over all possible length T state sequences through the signal-background state space lattice shown in Fig. 3. For example, the generative model in Fig. 2 has generated a single state sequence shown as $(i_1, j_1), (i_2, j_2), \cdots, (i_T, j_T)$. The notation $\int\int_C$ stands for the T-fold iterated integral, each component of which is along the contour C_t defined by $f(x_t, y_t) = z_t$.

Equation (7) expresses the probability of the observable data in terms of the probability of the complete data, which includes the noise, the signal, and the underlying state processes. It is important to note that one can define a particular model for the corruption of the signal by noise simply by defining particular contours of integration C_t in the x_t- y_t planes. For instance, in Fig. 4 simple contours are shown for several noise models.

1) Auxiliary Q Function: It is not possible to obtain the ML estimates directly. However, it is possible to iteratively improve on an initial model λ and find a new model $\overline{\lambda}$ such that $P(Z|\overline{\lambda}) \geq P(Z|\lambda)$. Baum *et al.* [1] showed that for a broad class of densities that includes those of the type given in (7), the desired improvement in likelihood can be obtained by finding a new model $\overline{\lambda}$ that maximizes an auxiliary function

$$Q(\lambda,\overline{\lambda}) = E\{\log P(X,Y,I,J|\overline{\lambda})\} \tag{9}$$

$$= \sum_I \sum_J \iint_C P(X,Y,I,J|\lambda) \cdot \log P(X,Y,I,J|\overline{\lambda}) dX dY. \tag{10}$$

A proof of this property of the Q function is contained in [1] and will not be repeated here. A simple expression for $Q(\lambda,\overline{\lambda})$ is obtained here in terms of the relevant densities. Assuming again that the observations are independent in time, and further assuming that the random processes representing X, Y, I, and J are independent, $Q(\lambda,\overline{\lambda})$ can be rewritten as

$$Q(\lambda,\overline{\lambda}) = \sum_I \sum_J \iint_C P(X,Y,I,J|\lambda) \cdot \log \left(\prod_{t=1}^{T} \overline{b}_{i_t}(\vec{x}_t)\overline{p}_{i_t}\overline{a}_{j_t}(\vec{y}_t)\overline{q}_{j_t} \right) dX dY \tag{11}$$

$$= \sum_{t=1}^{T} \sum_I \sum_J \iint_C P(X,Y,I,J|\lambda) \cdot \sum_{k=1}^{M} \sum_{l=1}^{N} n_t(k,l,I,J) \log (\overline{b}_k(\vec{x}_t)\overline{p}_k\overline{a}_l(\vec{y}_t)\overline{q}_l) dX dY \tag{12}$$

where $n_t(k,l,I,J)$ is the counting function defined by

$$n_t(k,l,I,J) = \begin{cases} 1 & \text{if } i_t = k, j_t = l \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

The Q function can then be reduced to

$$Q(\lambda,\overline{\lambda}) = \sum_{t=1}^{T} \sum_{k=1}^{M} \sum_{l=1}^{N} \iint_C \log (\overline{b}_k(\vec{x}_t)\overline{p}_k\overline{a}_l(\vec{y}_t)\overline{q}_1 \times \sum_I \sum_J n_k(k,l,I,J) P(X,Y,I,J|\lambda) dX dY \tag{14}$$

$$= \sum_{t=1}^{T} \sum_{k=1}^{M} \sum_{l=1}^{N} \iint_C \log (\overline{b}_k(\vec{x}_t)\overline{p}_k\overline{a}_l(\vec{y}_t)\overline{q}_l)\gamma_t(k,l) dX dY \tag{15}$$

where

$$\gamma_t(k,l) = \sum_I \sum_J n_t(k,l,I,J) P(X,Y,I,J|\lambda). \tag{16}$$

2) General Expressions for Model Parameters: Individually maximizing the expression for $Q(\lambda,\overline{\lambda})$ in (15) with respect to each of the model parameters in (3) is straightforward. The estimates of the model parameters will be obtained in two steps. First, general expressions for the model parameters will be derived. These are general expressions in that they do not depend on any specific noise corruption model. Second, parameter estimates for specific noise corruption models will be discussed in Section III. Different noise models are considered by substituting different functional forms for the noise corruption operation, $f(x_t, y_t)$, given in Fig. 1. In practice, it turns out that there is a relatively small number of functional forms that are numerically tractable according to the formulation in this section. However, some of these functional forms provide accurate representations of physically meaningful noise corruption mechanisms in speech processing.

To find $\overline{p}_i$, maximize (15) with respect to $\overline{p}_i$ under the constraint that $\Sigma_{k=1}^{M} \overline{p}_k = 1$. This yields the expression

$$\overline{p}_i = \frac{\sum_{t=1}^{T} \sum_{l=1}^{N} \iint_C \gamma_t(i,l) dX dY}{\sum_{t=1}^{T} \sum_{k=1}^{M} \sum_{l=1}^{N} \iint_C \gamma_t(k,l) dX dY}. \tag{17}$$

The dependence of the model parameters on the noise model can be made more clear by writing

$$\iint_C \gamma_t(k,l) dX dY = \prod_{\tau \neq t} p(z_\tau|\lambda) \cdot \iint_{C_t} b_k(\vec{x}_t) p_k a_l(\vec{y}_t) q_l d\vec{x}_t d\vec{y}_t \tag{18}$$

$$= p(Z|\lambda) p(i_t = k, j_t = l|\vec{z}_t, \lambda) \tag{19}$$

where

$$p(i_t = k, j_t = l|\vec{z}_t, \lambda) = \frac{\iint_{C_t} b_k(\vec{x}_t) p_k a_l(\vec{y}_t) q_l d\vec{x}_t d\vec{y}_t}{p(\vec{z}_t|\lambda)} = \frac{p(\vec{z}_t|i_t = k, j_t = l, \lambda) p_k q_l}{\sum_{k=1}^{M} \sum_{l=1}^{N} p(\vec{z}_t|i_t = k, j_t = l, \lambda) p_k q_l}. \tag{20}$$

Substituting (19) into (17), the class probability can be written as

$$\overline{p}_i = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} p(i_t = i, j_t = j|\vec{z}_t, \lambda). \tag{21}$$

Hence, the class probability is implicitly dependent on the noise corruption process through the dependence on the density of the noise corrupted observations.

The signal mean and variance are also estimated by maximizing the expression for $Q(\lambda,\overline{\lambda})$ given in (15). Since all densities are assumed to have diagonal covariance matrices, each vector component can be estimated independently, and the shorthand notations $\overline{\mu}_i$ and $\overline{\sigma}_i^2$ are used to refer to arbitrary vector components of the reestimated mean and variance

vectors, respectively. To find $\overline{\mu}_i$, maximize (15) with respect to $\overline{\mu}_i$

$$\frac{\partial Q(\lambda,\overline{\lambda})}{\partial \overline{\mu}_i} = \frac{\partial}{\partial \overline{\mu}_i} \sum_{t=1}^{T}\sum_{k=1}^{M}\sum_{l=1}^{N} \iint_C \gamma_t(k,l) \cdot \log \overline{b}_k(x_t) dX dY \tag{22}$$

$$= \frac{\partial}{\partial \overline{\mu}_i} \sum_{t=1}^{T}\sum_{k=1}^{M}\sum_{l=1}^{N} \iint_C \gamma_t(k,l) \cdot \left[-\frac{1}{2}\frac{(x_t-\overline{\mu}_k)^2}{\overline{\sigma}_k^2}\right] dX dY. \tag{23}$$

Setting the above expression equal to zero yields the following expression for $\overline{\mu}_i$

$$\overline{\mu}_i = \frac{\sum_{t=1}^{T}\sum_{j=1}^{N} \iint_C \gamma_t(i,j) x_t dX dY}{\sum_{t=1}^{T}\sum_{j=1}^{N} \iint_C \gamma_t(i,j) dX dY}. \tag{24}$$

The expression for the mean can be made more clear by first noting that, following the derivation of (19), the integral in the numerator of (24) can be rewritten as (25)–(27), which appear at the bottom of this page, resulting in the following expression for $\overline{\mu}_i$

$$\overline{\mu}_i = \frac{\sum_{t=1}^{T}\sum_{j=1}^{N} p(i_t=i, j_t=j|z_t,\lambda) E\{x_t|z_t, i_t=i, j_t=j, \lambda\}}{\sum_{t=1}^{T}\sum_{j=1}^{N} p(i_t=i, j_t=j|z_t,\lambda)}. \tag{28}$$

Through the same procedure, an expression is obtained for the variance $\overline{\sigma}_i^2$

$$\overline{\sigma}_i^2 = \frac{\sum_{t=1}^{T}\sum_{j=1}^{N} p(i_t=i, j_t=j|z_t,\lambda) E\{x_t^2|z_t, i_t=i, j_t=j, \lambda\}}{\sum_{t=1}^{T}\sum_{j=1}^{N} p(i_t=i, j_t=j|z_t,\lambda)} - \overline{\mu}_i^2. \tag{29}$$

These estimates for the signal model parameters provide a new model $\overline{\lambda}$, which results in an increase in the likelihood of the given noisy observations. The new model parameters can then be used to compute the conditional expectations in (28) and (29) and the *a posteriori* class probabilities in (21). This process of iteratively computing conditional expectations and using them to obtain the maximum likelihood estimates of the model parameters is known as the EM algorithm [8]. As would be expected, a probabilistic mixture of the type used for speaker identification in [30] corresponds to the degenerate clean channel case of the model presented here. There is no noise process in this case, so the signal observations are directly observable. As a result, the expected values in (28) and (29) are replaced by the observations, and the model degenerates to a Gaussian mixture.

III. Specific Noise-Corruption Functions

This section obtains closed form expressions for the signal model parameters by assuming several specific noise corruption functions $f(\,)$. A particular noise corruption function is chosen for a given problem and a given set of features based on the how the signal and background observations interact. General expressions have been obtained in Section II for the signal model parameters in the integrated signal-background model shown in Fig. 2. In order to obtain closed form expressions for the model parameters, a specific noise corruption function must be defined. Both the noisy observation probability, given in (5), and the conditional expectations in (28) and (29), are contour integrals. The contour of integration C_t in these integrals is defined by the general noise corruption function $f(\,)$.

Three separate noise corruption functions and the associated contours in the signal-noise plane are considered. These three cases are summarized in Fig. 4. It is assumed for the sake of simplicity in the examples shown here that the background model is a single Gaussian ($N = 1$ as in Fig. 1) and that the background mean and variance $\lambda_b = \{\mu_b, \sigma_b^2\}$ are known for each channel. As in the previous section, the vector-component index is dropped when referring to an individual vector component, i.e., an arbitrary vector component of the vector $\vec{x}_t$ is referred to as x_t.

A. Additive-Noise Model: $f(x_t, y_t) = x_t + y_t$

A purely additive noise model is appropriate if the components of the observation vector $\vec{z}_t$ represent linear filter-bank energies. Assuming independence of signal and background noise, it is well known that the noise corrupted observation density is given by the convolution of the signal and noise densities

$$p(z_t|i_t=i,\lambda) = \int_{-\infty}^{\infty} b_i(x_t) a(z_t - x_t) dx_t. \tag{30}$$

The conditional expectation required for the estimate of the mean in (28) is obtained by integrating (25) along the contour

$$\iint_C x_t \gamma_t(k,l) dX dY = \prod_{t\neq\tau} p(z_\tau|\lambda) \iint_C x_t p(x_t, y_t, i_t=k, j_t=l|\lambda) dx_t dy_t \tag{25}$$

$$= p(Z|\lambda) \frac{\iint_C x_t p(x_t, y_t|i_t=k j_t=l,\lambda) p(i_t=k, j_t=l|\lambda) dx_t dy_t}{p(z_t|\lambda)} \tag{26}$$

$$= P(Z|\lambda) p(i_t=k, j_t=l|z_t,\lambda) E\{x_t|z_t, i_t=k, j_t=l, \lambda\} \tag{27}$$

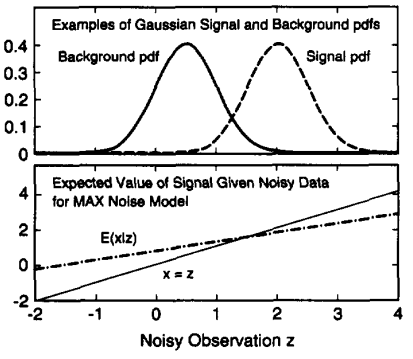


Fig. 5. Example of the expected value of the underlying signal conditioned on noisy observations that arise from the additive noise corruption function $z = x+y$. The signal and background processes are represented by univariate Gaussian densities displayed in the upper plot of the figure. The conditional expectation corresponding to $E(x|z)$ plotted versus z is displayed as the dashed line in the lower plot.

shown in Fig. 4(a)

$$E\{x_t|z_t, i_t = i, \lambda\} = \frac{\iint_{C_t} x_t p(x_t, y_t|i_t = k, \lambda) dx_t dy_t}{p(z_t|i_t = k, \lambda)} \quad (31)$$

$$= \frac{\int_{-\infty}^{\infty} x_t b_i(x_t) a(z_t - x_t) dx_t}{\int_{-\infty}^{\infty} b_i(x_t) a(z_t - x_t) dx_t} \quad (32)$$

$$= \frac{\sigma_i^2}{\sigma_i^2 + \sigma_b^2}\left[z_t + \left(\frac{\sigma_b^2}{\sigma_i^2}\mu_i - \mu_b\right)\right]. \quad (33)$$

The estimate of the signal variance given in (29) can be obtained in a similar way by computing the conditional expectation

$$E\{x_t^2|z_t, i_t = i, \lambda\} = \frac{\sigma_i^2\sigma_b^2}{\sigma_i^2 + \sigma_b^2} E\{x_t|z_t, i_t = i, \lambda\}^2. \quad (34)$$

The expected value of the signal conditioned on the noisy observations generated using the additive noise corruption function can be observed by plotting the values of $E\{x|z, \lambda\}$ with respect to the values of the noisy observations. This is plotted in Fig. 5 for a simple example where both the signal and background densities are represented as single Gaussians. Note that since the density of the sum of Gaussians is also Gaussian, this expected value, shown in the lower plot of Fig. 5, is a linear function of the noisy observations.

B. *Additive Noise with* log *Spectrum Observations:* $e^{f(x_t, y_t)} = e^{x_t} + e^{y_t}$

This model is appropriate under the assumption that noise is additive in the power spectrum or filter-bank domain, but the measurements are actually in the log power-spectrum domain. The problem with this noise model is that the integral in (25) evaluated along the appropriate contour shown in Fig. 4(b) has no closed form solution. Van Compernolle investigated the estimation of the conditional expectation $E\{x_t|z_t, \lambda\}$ in a similar context, assuming a uniform instead of a Gaussian distribution for the underlying signal density [5]. No closed form solution was found, and the integrals were evaluated through a series expansion.

C. MAX *Noise Model:* $f(x_t, y_t) = \max(x_t, y_t)$

The "max" noise model used by Nadas *et al.* assumes that a noisy observation z_t is formed as the maximum of the signal and background observations [23]. The contour in Fig. 4(c) shows that the MAX noise model is a reasonable approximation to the case where there is additive noise with log domain observations. If the actual (linear) signal $\tilde{x}_t$ is corrupted by additive background noise $\tilde{y}_t$ and the parameter vectors are formed from log filterbank channel energies, $x_t = \log(\tilde{x}_t)$ and $y_t = \log(\tilde{y}_t)$, then one can approximate the noise process using a max operator

$$\log(\tilde{x}_t + \tilde{y}_t) \approx \max(x_t, y_t).$$

With independent signal and background processes, the noisy observation density is obtained by integrating along the contour in Fig. 4(c)

$$\begin{aligned} p(z_t|i_t = i, \lambda) &= \iint_{C_t} b_i(x_t) a(y_t) dx_t dy_t \\ &= a(z_t) \int_{-\infty}^{z_t} b_i(x_t) dx_t + b_i(z_t) \int_{-\infty}^{z_t} a(y_t) dy_t \\ &= a(z_t) B_i(z_t) + b_i(z_t) A(z_t) \end{aligned} \quad (35)$$

where $B_i(\,)$ is the cumulative distribution for mixture component i of the signal and $A(\,)$ is the cumulative distribution for the background noise. The first term in (35) is the integral along the horizontal branch of C_t in Fig. 4(c), and the second term is the integral along the vertical branch of C_t.

To obtain a closed form expression for the component signal mean $\overline{\mu}_i$ in (28) under the assumption of a *max*( ) noise corruption function, an expression for the conditional expectation of the signal given the noisy observations must be derived. This conditional expectation can be computed by integrating along the contour in Fig. 4(c)

$$E\{x_t|z_t, i_t = i, \lambda\}$$

$$= \iint_{C_t} x_t p(x_t|z_t, i_t = i, \lambda) dx_t dy_t \quad (36)$$

$$= \frac{a(z_t)\int_{-\infty}^{z_t} x_t b_i(x_t) dx_t + z_t b_i(z_t) \int_{-\infty}^{z_t} a(y_t) dy_t}{p(z_t|i_t = i, \lambda)} \quad (37)$$

$$= \frac{a(z_t) B_i(z_t) E\{x_t|x_t < z_t, i_t = i, \lambda\} + z_t b_i(z_t) A(z_t)}{p(z_t|i_t = i, \lambda)}. \quad (38)$$

Here again, the first term in (37) is the integral along the horizontal branch of C_t in Fig. 4(c), resulting in the expected value of a truncated Gaussian. The second term is the integral along the vertical branch of C_t. Equation (38) follows from (37) by the definition of the mean of a truncated Gaussian

$$E\{x_t|x_t < z_t, i_t = i, \lambda\} = \frac{\int_{-\infty}^{z_t} x_t b_i(x_t) dx_t}{B_i(z_t)} \quad (39)$$

$$= \mu_i - \sigma_i \frac{b_i(z_t)}{B_i(z_t)}. \quad (40)$$

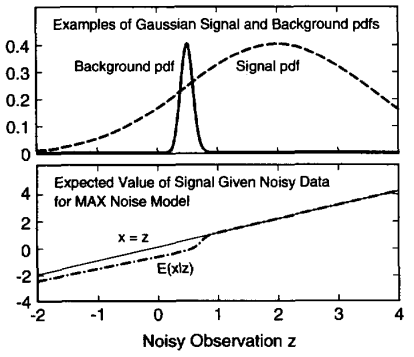


Fig. 6. Example of the expected value of the underlying signal conditioned on noisy observations that arise from the max noise corruption function $z = \max(x, y)$. Example signal and background processes are represented by Gaussian densities shown in the upper plot of the figure. The conditional expectation corresponding to $E(x|z)$ plotted versus z is displayed in the lower plot.

Since $b_i(z_t)A(z_t)$ is the probability mass in Fig. 4(c) where $x_t = z_t$, then

$$p(x_t = z_t|i_t = i, \lambda) = \frac{b_i(z_t)A(z_t)}{a(z_t)B_i(z_t) + b_i(z_t)A(z_t)}. \quad (41)$$

The conditional expectation of the signal given the noisy observations can be obtained directly from (38) and (41) as

$$\begin{aligned} E\{x_t|z_t, i_t = i, \lambda\} = & p(x_t = z_t|i_t = i, \lambda)z_t \\ & + (1 - p(x_t = z_t|i_t = i, \lambda)) \\ & \cdot E\{x_t|x_t < z_t, i_t = i, \lambda\} \end{aligned} \quad (42)$$

Fig. 6 shows the expected value of the signal conditioned on the noisy observations for the max noise corruption function. The upper plot in Fig. 6 displays Gaussian densities representing the background process and an example signal process. The conditional expectation corresponding to $E(x|z, \lambda)$ plotted versus z is displayed in the lower plot of the figure. It is helpful to interpret these curves in the context of (42). For those values of z where $p(x = z|\lambda) \approx 1$ in Fig. 6, $E(x|z, \lambda) \approx z$. For observation values that are likely to be noise and not signal, the expected value of the signal is determined by $E\{x|x < z, \lambda\}$. This expected value defines the input/output relationship when a noisy observation is the input and the estimated signal is the output. It is intuitively satisfying that this relationship is defined directly in terms of the underlying signal and background densities.

IV. Relation to Noise Masking

When signal vectors are observed in the presence of noise, it is inevitable that some useful signal information may be completely lost to the noise corruption process. The integrated signal–background model is a framework that probabilistically selects noisy observations for training that are not likely to have been corrupted by background. Stated another way, the probabilistic framework provides a means for rejecting observations where the signal information has been lost. This section provides more intuition into this process by comparing the probabilistic approach described in Section II with another class of techniques that attempts to integrate background directly into the classifier. It is shown that, when certain simplifying assumptions are applied to the probabilistic model of Section II, the model is equivalent to a straightforward implementation of a technique known as noise masking.

Noise masking techniques have been used for noise compensation in speech recognition [15], [19], [36]. It is always assumed that noise masking is performed on log energies of filter-bank channels for a speech frame. This stems from the approximation that the noise in a particular filter band will only affect the signal energy within that filter band, and that signal energy within the adjacent filter-bank channels is statistically independent. The effect of all of the techniques is to simulate identical noise conditions for observations during both training and recognition by adding an appropriate noise-mask level to all observations.

Noise masking techniques have been implemented in HMM speech recognizers by altering the form of the Gaussian observation probability density functions in the speech recognizer [15], [36]. If a filter band contains noise, it is likely that the measured filter-band energy will contain little information about the underlying speech. Holmes and Sedgwick [15] assumed that if the signal level within a filter band fell below a specified noise-mask level B, then the signal could not be trusted at all and should be replaced by the noise-mask level. If Gaussian observation density parameters are estimated from observations that have been partially masked by noise, the resulting parameters will be biased. Holmes and Sedgwick addressed this problem by estimating the means and variances of continuous Gaussian HMM observation densities using only those vectors from the training data whose log filter-bank energy falls above a threshold B. The threshold was specified from an estimate of the noise level found during the training process. This problem of estimating statistical model parameters from incomplete data, where a known portion of the data is discarded, corresponds to the censored data problem as posed by Dempster *et al.* [8] and Little and Rubin [31].

In the following development, two approaches will be used for estimating the mean μ for the single Gaussian density under the conditions described for noise masking. In the first approach, a ML model will be obtained directly using a special case of an approach suggested in [8]. In the second approach, the noise masking problem is treated as a degenerate case of the integrated signal–background model of Section II. This exercise serves to illustrate the relationship between the integrated model and the more traditional noise masking approaches.

For the first approach, the EM algorithm is applied to estimating the mean μ of a single Gaussian density $p(z|\lambda)$ from training samples where a known portion of the sample values fall below a threshold B. The problem is depicted in Fig. 7, and the reader is referred to the appendix for the derivation. It is assumed that a total of T training vectors are available, but only T_o of these vectors, $Z = (z_1, z_2, \cdots, z_{T_0})$, are actually observable. The remaining $T - T_o$ vectors $X = (x_1, x_2, \cdots, x_{T-T_o})$ are assumed to have been masked by noise at a known level B. The mean of the complete data, which includes both the observed vectors and those masked

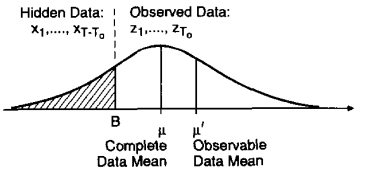


Fig. 7. Estimation of mean vector for Gaussian density from partially censored data.

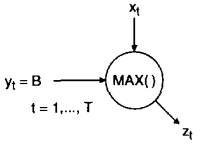


Fig. 8. Degenerate case for the integrated signal-background model equivalent to the censored data model used for noise masking.

by noise, is given as

$$\mu = \frac{1}{T}\sum_{t=1}^{T_o} z_t + \frac{T - T_o}{T} E\{z_t | z_t < B\}. \tag{43}$$

Simply interpreted, the complete data mean given in (43) is a linear combination of the sample mean and the observed data and the sample mean and the unobserved data that has been masked by noise at a fixed level B.

The second approach is based on the fact that censoring data samples that fall below a threshold is equivalent to using the *max*( ) noise corruption process with a background signal equal to the threshold value. This is illustrated in Fig. 8. It is interesting to compare the expression for the mean in the censored data case given in (43) with the expression for the mean obtained for the integrated signal–background model in Section III when similar assumptions are made about the noise corruption process.

An expression similar to (43) can be obtained for this integrated noise model case. If the signal model order is set to $M = 1$, then it is easy to show from (29) and (42) that the estimate for the mean of the *max*( ) noise corruption process is given by

$$\mu = \frac{1}{T}\sum_{t=1}^{T} p(x_t = z_t) z_t + \frac{1}{T}\sum_{t=1}^{T}(1 - p(x_t = z_t)) E\{x_t | x_t < z_t\}. \tag{44}$$

If the background process is replaced by a constant level as shown in Fig. 8, then $p(x_t = z_t)$ in (44) can be replaced by

$$p(x_t = z_t) = \begin{cases} 1 & z_t > B \\ 0 & z_t < B \end{cases} \tag{45}$$

and the expression for μ in (44) is exactly the same as that obtained for the censored data case in (43).

The relationship between (43) and (44) highlight the consequences of differing assumptions between the two methods. In the censored data (noise masking) case, it is assumed that the background level is known with complete certainty and that no information can be derived from observations that fall below this level. However, in the integrated noise case, a measure of uncertainty about the background is included in the model, allowing some level of information to be derived from all noise corrupted observations. In general, one would expect the greatest rewards from the probabilistic approach to integration of speech and background when instantaneous background levels cannot be precisely estimated and only averaged estimates of the background process are available.

V. Application to Speaker Identification

In this section, the integrated model of signal and background is applied to the problem of text independent speaker identification. The particular applications of interest are those involving unconstrained conversational speech utterances observed in noisy acoustic environments. The techniques are evaluated as part of an automated system for classifying unlabeled segments of a continuous utterance according to speaker identity.

There are a number of techniques that have demonstrated good text-independent speaker identification performance in relatively low-noise environments [13], [14], [30], [32], [34]. A system based on a modified maximum likelihood trained Gaussian speaker classifier obtained over 90% correct speaker classification performance with a 16-speaker population for utterances recorded over the public switched telephone network [13]. Maximum likelihood trained Gaussian mixtures were used to represent speakers in [30]. It was found that Gaussian mixtures with diagonal covariance matrices provided a 30% improvement in speaker classification accuracy over unimodal full covariance Gaussians when classifying 5 s segments of utterances from a wideband conversational speech database.

In the following experimental study, the Gaussian mixture is extended for speaker identification to include the integrated background model shown in Fig. 2. The goal of these experiments is to determine how speaker identification performance in noisy acoustic environments is affected when prior knowledge of the acoustic background process is incorporated through the use of integrated models of signal and background.

A. Incorporating Prior Knowledge of Acoustic Background

The experiments in this section address the two problems illustrated in Fig. 9. The first problem, illustrated by Fig. 9(a), is the estimation of speaker model λ_s from noisy speech-observation vectors $\vec{z}_1, \vec{z}_2, \cdots, \vec{z}_T$ given a prior estimate of the background model λ_b. Speaker model parameters are estimated using the maximum likelihood parameter estimation technique described in Section II. The second problem, illustrated by Fig. 9(b), is to compute the likelihood of the noisy observations with respect to a speaker model λ_s and background model λ_b. Assuming independent observations, this likelihood is computed over the signal–background state space lattice as $\Pi_{t=1}^{T} p(\vec{z}_t | \lambda_s, \lambda_b)$ in (6). Speaker identification is performed in this context simply by choosing the speaker model λ_s that maximizes the log of this likelihood for the T length segment of noisy observations.

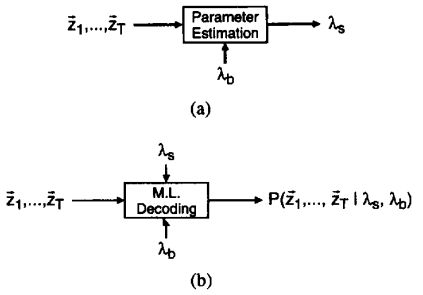


Fig. 9. (a) Speaker model estimation from noisy observation vectors, given prior model of acoustic background; (b) maximum likelihood decoding of noisy observations with respect to speaker and background models.

TABLE I
SPEAKER IDENTIFICATION RESULTS FOR GAUSSIAN MIXTURE CLASSIFER WITH AND WITHOUT THE INTEGRATED BACKGROUND MODEL. SPEECH BABBLE AND BROAD-BAND GAUSSIAN NOISE ENVIRONMENTS AT 10 dB SNR WERE SIMULATED. THE TABULATED RESULTS REPRESENT THE PERCENT CORRECT IN CLOSED SET SPEAKER CLASSIFICATION OF 10 s SEGMENTS FOR A POPULATION OF 16 MALE SPEAKERS

Effect of Integrated Background Model in Speaker Identification				
	Models Trained in Clean, High SNR Environment			Models trained in Speech Babble
Classifier	Test Environment:			Test Environment:
	Clean	Speech Babble	White Noise	White Noise
GMM	95.8	58.0	14.1	12.6
GMM-IB	95.8	79.9	68.8	60.8

B. Experiments

Different acoustic background environments were simulated by adding sampled waveforms representing the noise sources to digitized speech. Recorded examples of two noise conditions, speech babble and broad-band Gaussian noise, were obtained from the NATO RSG10 noise corpus [24]. The discrete time samples of the noise signals were scaled and added to the speech to provide average SNR's over an entire conversation of 0 dB, 5 dB, and 10 dB.

The speaker classification experiments were performed on the KING conversational speech corpus. The corpus contains natural speech elicited from male subjects by having them interact with an interlocutor on one of six simple tasks. The conversations were edited to isolate the subjects' speech from the interlocutor's speech and to remove long silences. Each conversation was approximately 5 min in length before editing, and reduced to approximately 40 s after editing. The speech that was used in these experiments was recorded using an electret microphone and then digitized as 12-bit linear samples at a sampling rate of 10 kHz. Sixteen male speakers were used to define the speaker population. Five conversational sessions per speaker were used including two sessions for training speaker models, and the remaining three sessions were for evaluating classification performance.

The noisy speech observation vectors were obtained directly from the magnitude DFT spectrum of the noisy sampled speech waveform. Twenty-six log filterbank energies were computed over a mel-warped frequency scale in accordance with [7]. Observation vectors were estimated every 10 ms over a 20 ms windowed segment of speech. Speaker models were trained from 6000 observation vectors. All of the experiments compare speaker classification performance of Gaussian mixture speaker models (GMM) with the performance of Gaussian mixtures with the integrated background (GMM-IB). In using the simple GMM for speaker identification in noise, it is implicitly assumed that there is no background model and that the signal observations are directly observable, $\vec{z}_t = \vec{x}_t$.

There is no theoretical means for selecting either speaker model or background model order (the number of mixture densities) that correlates well with classification performance. The empirically optimum model order was found to depend on a number of issues, including the amount of training data per speaker and the length of the test utterance over which the speaker model likelihood is computed. An experimental study was conducted to investigate these issues [27]. It was found that with greater than 60 s of training data per speaker, classification performance for the GMM speaker classifier did not continue to improve when the model order was increased beyond 16 to 32 component Gaussians. Referring to the GMM-IB model in Fig. 9, the speaker model order was set at $M = 25$ for all speakers. There was no formal experimental study to determine an optimum background model order. The background model order was set at $N = 5$ for the speech-babble noise corruption process and $N = 1$ for broad-band Gaussian noise. Informal experiments found no observable improvement in performance when the background model order for the respective background processes was increased.

The GMM-IB introduces prior knowledge of the background process through prior estimates of the parameters of the background model λ_b. In [29] the parameters of a unimodal Gaussian background model ($N = 1$) were estimated from noisy speech by identifying "non-speech" observation frames within the utterance. Frames identified as nonspeech were used to estimate a single background mean and variance. The study summarized in Table I obtained prior estimates of the background model parameters for the GMM-IB from independent samples of the two noise conditions. The problem of continually updating estimates of the background model parameters during GMM-IB classification and training is a topic of future research, but is not directly addressed here. It was found that the classification performance of the GMM, where no knowledge of the background process can be incorporated into the model, was significantly affected by the mismatch in the energy levels between training and testing utterances. To remove this effect, all observation vectors were normalized to unit energy before being presented for GMM training and classification.

Several sets of experiments were performed to evaluate the GMM-IB with respect to three different types of environment mismatch. The first set of experiments investigated the effect of the integrated background model on classification performance alone, as shown in Fig. 9(b). For both the GMM and the GMM-IB, speaker models were trained in clean channel environments so that background noise was not an issue in training.

The last two sets of experiments determined whether the GMM-IB could provide better model parameterl estimates when speaker models were trained from noisy observations, as in Fig. 9(a). The scenario in Fig. 9(b) suggests that, given existing speaker models, we can use these models for classification in a different environment without retraining. All that is required is a representative sample of the new background signal to estimate new background model parameters. The first environmental mismatch scenario considered was a scenario where speaker models were trained using speech corrupted by one *type* of noise (speech babble) and the models were used for identification in speech corrupted by a different *type* of noise (white noise). The second scenario was concerned with SNR mismatch between training and testing environments. Speaker models were trained in one type of noise at a given SNR and then used for identification on speech corrupted by the same noise type at a different SNR. The experiments for both types of mismatch were designed to examine the effectiveness of the GMM-IB training procedure in decoupling the speech and background processes so that a speaker model trained in one environment could be used in a different environment by simply using a new background model.

C. Results

Speaker classification results for the speech babble and broad-band Gaussian noise conditions are shown in Table I. Results are given using GMM-IB and GMM speaker classifiers, labeled in the first column of Table I as Gaussian mixtures with and without background model, respectively, for classifying 10-s length segments of conversational utterances from each speaker. The performance is given as the percent of segments correctly classified in test utterances from all 16 speakers. The second major column in Table I describes the first set of experiments where speaker models were trained in a noise-free (clean) environment and classification was performed under several separate conditions. The third column of the table describes a paradigm where training and classification were performed in the presence of two different *types* of background noise. All simulated noisy-speech environments corresponded to a level of 10 dB SNR.

There are several conclusions that can be derived from the results given in Table I. The first conclusion is that the GMM-IB resulted in a significant improvement in performance relative to the GMM under mismatched conditions. It is clear from the first row of Table I that there was a significant degradation in speaker classification performance for the GMM when the noise environment in testing did not match the training noise environment. Note that Table I shows that for clean channel test environment with very high SNR, the GMM-IB had no effect on performance. The performance of the GMM-IB under mismatched conditions, shown in the second row of Table I, represented a significant improvement over the GMM where no prior knowledge of the background environment was available. This effect can be explained by the implied decoupling of the speech and background processes inherent in the GMM-IB. In speaker classification under a particular noise condition, the same background model is used for all speakers. Hence, any noisy observations that represent the background noise process, and not the speaker, will receive identical likelihood scores across all speakers.

TABLE II
SPEAKER IDENTIFICATION RESULTS USING SPEAKER MODELS TRAINED IN A 10 dB SNR ENVIRONMENT TO CLASSIFY SPEECH FROM NOISE CORRUPTED ENVIRONMENTS AT DIFFERENT SNR'S. RESULTS ARE FOR 10 s TEST UTTERANCES

Speech Babble			
Training Environment	Testing Environment		
10 dB	10 dB	5 dB	0 dB
GMM	89.6	82.8	57.0
GMM-IB	84.3	79.5	72.4
White Noise			
Training Environment	Testing Environment		
10 dB	10 dB	5 dB	0 dB
GMM	77.0	44.3	22.6
GMM-IB	79.8	70.0	47.9

The second conclusion that can be derived from the results shown in Table I is that the GMM-IB can improve performance when a noise *type* mismatch exists across training and testing conditions. The GMM-IB results shown in the third column of Table I represent performance when speaker models were trained in the presence of speech babble and classification was performed in the white noise background environment. Classification was performed in the different environment simply by using a new background model that was representative of the new background process. Speaker classification performance for this paradigm represented a significant improvement over the GMM where no background model was used.

The results of the last set of experiments dealing with SNR environmental mismatch are given in Table II. For both noise types, 10 dB SNR conditions were simulated for the training environment, and 10 dB, 5 dB, and 0 dB SNR conditions were simulated for the testing environments. The background model trained to represent noise conditions at 10 dB SNR was simply scaled to match the testing SNR and used for classification. The results in Table II show how the performance of the GMM begins to degrade as the testing environment SNR departs from that of the training environment. The degradation is much more rapid for the white noise interference than the speech babble interference, which sharply decreases in performance for the 0 dB SNR test environment. For both types of background, the GMM-IB maintains a robustness to the decreasing testing SNR. The GMM-IB has slightly lower performance than the GMM for speech babble at 10 dB and 5 dB SNR tests, but maintains identification performance as the SNR drops to 0 dB. For white noise, the GMM-IB outperforms the GMM for all SNR levels in the test environment, most significantly for the 5 dB and 0 dB SNR levels.

Finally, it can be concluded that the integrated model is equally effective both for "traditional" Gaussian noise and for speech babble that is not well modeled by more well known noise compensation techniques. Indeed, this is one of the major advantages of the integrated model and has served to motivate further investigation of these techniques.

VI. Summary

A maximum likelihood procedure for signal model parameter estimation in noise based on the Expectation–Maximization algorithm has been presented. Expressions were obtained for the ML estimates of Gaussian-mixture signal model parameters when the signal is observed in the presence of a known Gaussian mixture background noise process. The procedure is very general in that it can be applied to a broad class of background noise processes. While the parameter estimation procedure could theoretically be applied to any form of interaction between signal and background, closed form expressions for signal model parameters were obtained for two physically meaningful cases.

The integrated signal–background model was applied as a noise compensation procedure for text-independent speaker identification in noise. The procedure was compared to a Gaussian mixture speaker classifier where the acoustic background was not considered. Two simulated noise environments were investigated. These included both broadband Gaussian noise and speech babble, at SNR's ranging from 0 dB to 10 dB relative to the target speech. Significant performance improvement was obtained for cross-environment training and testing when speakers were represented using integrated models of signal and background.

Appendix

In this Appendix, the EM algorithm is used to obtain a maximum likelihood estimate of the mean of a unimodal Gaussian density from partially censored data. The complete data in this case correspond to the observed data $Z = \{z_1, z_2, \cdots, z_{T_o}\}$ and the noise masked data $X = \{x_1, x_2, \cdots, x_{T-T_o}\}$. The expression for the complete data likelihood $f(X, Z|\lambda)$ is given by

$$f(X, Z|\lambda) = \frac{T!}{T_o!(T-T_o)!} \prod_{t=1}^{T_o} p(z_t|\lambda) \prod_{t=1}^{T-T_o} p(x_t|\lambda) \quad (46)$$

where $p(z_t|\lambda)$ is assumed Gaussian and is defined by its mean μ and variance σ^2.

Proceeding as in Section II, a new model $\overline{\lambda}$ is obtained that will increase the probability of the observed data by maximizing an auxiliary function with respect to the mean μ. The auxiliary function is given by

$$Q(\lambda, \overline{\lambda}) = E\{\log f(X, Z|\overline{\lambda})\}. \quad (47)$$

Expanding those terms of (47) that depend on λ we get

$$Q(\lambda, \overline{\lambda}) = E\left\{\sum_{t=1}^{T_o} \log p(z_t|\overline{\lambda})\right\} + E\left\{\sum_{t=1}^{T-T_o} \log p(z_t|\overline{\lambda}) | z_t < B\right\} \quad (48)$$

$$= \sum_{t=1}^{T_o} \log p(z_t|\overline{\lambda}) + \sum_{t=1}^{T-T_o} \int_{-\infty}^{\infty} \log p(z_t|\overline{\lambda}) p(z_t|z_t < B) dz_t \quad (49)$$

$$= \sum_{t=1}^{T_o} \log p(z_t|\overline{\lambda}) + (T-T_o) \frac{\int_{-\infty}^{B} \log p(z_t|\overline{\lambda}) p(z_t|\lambda) dz_t}{\int_{-\infty}^{B} p(z_t|\lambda) dz_t}. \quad (50)$$

The expression for the re-estimated $\overline{\mu}$ is obtained by maximizing the auxiliary function with respect to $\overline{\mu}$

$$\frac{\partial}{\overline{\mu}} Q(\lambda, \overline{\lambda}) = \sum_{t=1}^{T_o} \frac{z_t - \overline{\mu}}{\overline{\sigma}} \quad (51)$$

$$+ (T-T_o) \frac{\int_{-\infty}^{B} \frac{z_t - \overline{\mu}}{\overline{\sigma}} p(z_t|\lambda) dz_t}{\int_{-\infty}^{B} p(z_t|\lambda) dz_t}. \quad (52)$$

Solving (51) for $\overline{\mu}$, we get

$$\overline{\mu} = \frac{1}{T} \sum_{t=1}^{T_o} z_t + \frac{T-T_o}{T} E\{z_t | z_t < B\}. \quad (53)$$

References

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

[2] J. R. Bellegarda and D. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 12, pp. 2033–2045, 1990.

[3] A. E. Bisson, "Pattern recognition in ocean acoustics," in *Auditory and Visual Pattern Recognition*, D. A. Waterman and F. Hayes-Roth, Ed. New York: Academic, 1981.

[4] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[5] D. Van Compernolle, "Spectral estimation using a log distance error criterion applied to speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1987, pp. 258–261.

[6] D. Van Compernolle, "Noise adaptation in a hidden Markov model speech recognition system," *Comput., Speech, Language*, vol. 3, pp. 151–167, 1989.

[7] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, 1980.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc.*, vol. 39, pp. 1–38, 1977.

[9] K. M. Dobroth, B. L. Zeigler, and D. Karis, "Future directions for audio interface research: characteristics of human-human order-entry conversations," in *Proc. Am. Voice Input/Output Soc.*, Sept. 1989.

[10] Y. Ephraim, "Gain-adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Trans. Signal Processing*, vol. 40, no. 6, pp. 1303–1316, 1992.

[11] J. L. Flanagan, J. D. Johnson, R. Zahn, and G. W. Elko, "Computer steered microphone arrays for sound transduction in large rooms," *J. Acoust., Soc. Am.*, vol. 78, no. 5, pp. 1508–1518, 1985.

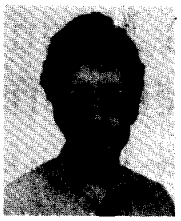
[12] H. Gish, Y. Chow, and J. R. Rohlicek, "Probabilistic vector mapping of noisy speech parameters for HMM word spotting," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1990, pp. 117–120.

[13] H. Gish *et al.*, "Investigation of text-independent speaker identification over telephone channels," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1985, pp. 379–382.

[14] A. L. Higgins and L. G. Bahler, "Text-independent speaker verification by discriminator counting," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, May 1991.

[15] J. N. Holmes and Nigel C. Sedgwick, "Noise compensation for speech recognition using probabilistic models," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1986.

[16] W. M. Huang and R. C. Rose, "Integrated models of signals and background for an HMM/neural net ocean acoustic events classifier," *Asilomar Conf. Signals, Syst., Comput.*, Nov. 1991.
[17] X. D. Huang and M. A. Jack, "Semi-continuous hidden Markov models for speech signals," *Comput., Speech, Language*, vol. 3, pp. 239–257, 1989.
[18] B. H. Juang and L. R. Rabiner, "Signal restoration by spectral mapping," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1987, pp. 2368–2371.
[19] D. H. Klatt, "A digital filter bank for spectral matching," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1976.
[20] W. Kushner, W. Gancharoff, C. Wu, and V. Nguyen, "Cockpit speech enhancement," Rome Air Development Center, Tech. Rep. RADC-TR-90-306, Nov. 1990.
[21] D. Mansour and B. H. Juang, "A family of distortion operators based on projection operation for robust speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 1659–1671, Nov. 1989.
[22] N. Merhav and C. H. Lee, "A minimax classification approach with application to robust speech recognition," To appear in *IEEE Trans. Acoust., Speech, Signal Processing*, 1992.
[23] A. Nadas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. Acoust.,Speech, Signal Processing*, vol. 37, no. 10, pp. 1495–1502, 1989.
[24] H. J. M. Steeneken and F. W. M. Geurtsen, "Description of the RSG-10 noise database," TNO Institute for Perception, Soesterberg, The Netherlands, Tech. Rep. IZF 1988-3, 1988.
[25] S. Oh, V. Viswanathan, and P. Papamichalis, "Hands free voice communication in an automobile with a microphone array," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1992.
[26] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.
[27] D. A. Reynolds, "A Gaussian mixture modeling approach to text-independent speaker identification," Ph.D. dissertation, Georgia Institute of Technology, Atlanta, Aug. 1992.
[28] D. A. Reynolds and R. C. Rose, "An integrated speech–background model for robust speaker identification," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1992.
[29] R. C. Rose, J. Fitzmaurice, E. M. Hofstetter, and D. A. Reynolds, "Robust speaker identification in noisy environments using noise adaptive speaker models," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, May 1991.
[30] R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1990.
[31] Roderick, J. A. Little, and D. B. Rubin, *Statistical analysis with missing data*. New York: Wiley, 1987.
[32] L. Rudasi and S. A. Zahorian, "Text-independent talker identification with neural networks," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1988.
[33] H. F. Silverman, "An algorithm for determining talker location using a linear microphone array and optimal hyperbolic fit," in *Proc. DARPA Speech, Natural Language Workshop*, June 1990.
[34] F. K. Soong, A. E. Rosenburg, B. H. Huang, and L. R. Rabiner, "A vector quantization approach to speaker recognition," *AT& T Tech. J.*, vol. 66, no. 2, pp. 14–26, 1987.
[35] A. P. Varga and R. E. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1990.
[36] A. P. Varga, R. E. Moore, J. S. Bridle, K. M. Ponting, and M. Russell, "Noise compensation algorithms for use with hidden Markov model based speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1988.



R. C. Rose received the B.S. and M.S. degrees in electrical engineering from the University of Illinois in 1979 and 1981, respectively. He received the Ph.D. E.E. degree from the Georgia Institute of Technology in 1988, completing his dissertation work in speech coding and analysis.

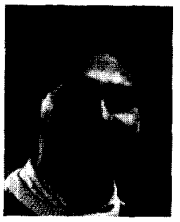
From 1980 to 1984, he was with Bell laboratories, Holmdel, NJ, where he worked on speech processing problems in digital switching environments. From 1988 to 1992, he was a member of the Speech Systems and Technology group at MIT Lincoln Laboratory. While there, he was involved in developing techniques for keyword recognition, improved noise robustness in speech processing, and speaker identification. He is presently a member of technical staff at Bell Laboratories, Murray Hill, NJ, where his work has focused on problems relating to speech recognition and speaker verification.

He is a member of the IEEE SP Technical Committee on Digital Signal Processing, the Acoustical Society of America Technical Committee on Speech, and is an adjunct faculty member with the Georgia Institute of Technology. He is also a member of Tau Beta Pi, Eta Kappa Nu, and Phi Kappa Phi.

E. M. Hofstetter was born in New York, New York on November 17, 1932. He received the B.S. and M.S. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1955. From 1955 to 1956 he was a Teaching Assistant at M.I.T. , and began working toward the doctorate. This work was interrupted during 1956–1957 when he received a Fulbright scholarship to study mathematics at the University of Gottingen, Germany. He received the Sc.D. degree in electrical engineering from M.I.T. in 1959.

He became an Assistant Professor of Electrical Engineering at M.I.T in 1959. In this capacity he was primarily concerned with teaching and research in the areas of signal analysis and the theory of stochastic processes. In 1963, he joined the staff of the M.I.T. Lincoln Laboratory, Lexington, Massachusetts, where he worked on problems associated with the radar examination of reentry vehicles. He is presently working in the area of digital signal processing with emphasis on the hardware and software aspects of speech signal processing.

D. Hofstetter is a member of Sigma Xi.



D. A. Reynolds received the B.E.E. degree (with highest honors) in 1986 and the Ph.D. degree in electrical engineering in 1992, both from the Georgia Institute of Technology.

Currently he is a staff member in the Speech Systems Technology Group at the Lincoln Laboratory, Massachusetts Institute of Technology where his research interests include robust speaker and speech recognition and transient signal classification.

Douglass is a member of IEEE Signal Processing Society, Eta Kappa Nu, and Tau Beta Pi.