Statistical-Model-Based Speech Enhancement Systems

YARIV EPHRAIM, SENIOR MEMBER, IEEE

Speech enhancement has been a challenge for many researchers for almost three decades. The problem involves improving the performance of speech communication systems in noisy environments. Since the statistics of the speech signal as well as of the noise are not explicitly available, and the most perceptually meaningful distortion measure is not known, model-based approaches have recently been extensively studied and applied to the three basic problems of speech enhancement. These problems comprise 1) signal estimation from a given sample function of noisy speech, 2) signal coding when only noisy speech is available, and 3) recognition of noisy speech signals in man–machine communication. In this paper, the recent research on the model-based approach is integrated and put into perspective with other more traditional approaches for speech enhancement. A unified statistical approach for the three basic problems of speech enhancement is developed using composite source models for the signal and noise and a fairly large set of distortion measures.

I. Introduction

Speech enhancement attempts to improve the performance of voice communication systems when their input or output signal is corrupted by noise. The improvement is in the sense of minimizing the effects of the noise on the performance of these systems. The need for enhancing speech signals arises in many situations in which the speech either originates from some noisy location or is affected by the noise over the channel or at the receiving end. Both digital and analog channels are possible, and communication can be either between people or with a machine. Hence, speech enhancement is the problem of enhancing a given sample function of noisy speech signal, as well as the problem of enhancing the performance of speech coding and recognition systems whose input signal is noisy. These problems have been a challenge for many researchers for almost three decades (see [1]–[5] and the references therein, [6]–[54], [57], [63], [79]–[83], and [85]–[107]).

Examples of important applications of speech enhancement include improving the performance of 1) cellular radio telephone systems, which usually suffer from background noise in the car as well as from channel noise; 2) pay phones located in noisy environments (e.g., airports); 3) air–ground communication systems in which the cockpit noise corrupts the pilot's speech; 4) teleconferencing systems where noise sources in one location may be broadcast to all other locations; 5) long-distance communication over noisy radio channels; 6) paging systems located in noisy environments (e.g., airports, machine rooms); 7) ground–air communication in which the cockpit noise corrupts the received messages; and 8) suboptimal speech quantization systems.

In the cellular radio telephone example, the original speech is corrupted by the noise generated by the engine, fan, traffic, and wind [53], [94], as well as by the channel noise. The signals delivered by cellular systems may therefore be noisy with impaired quality and intelligibility. If the cellular system encodes the signal prior to its transmission, then further degradation in its performance results, since speech coders rely on some model for the clean signal and normally that model is not suitable for the noisy signal. Similarly, if the cellular system is equipped with a speech recognition system which is used for automatic dialing, then the recognition accuracy of such system deteriorates in the presence of noise, since the noisy input signal is unlikely to obey the statistical model for the clean signal used by the recognizer. Similar problems are encountered with pay phone communication, air–ground communication, and teleconferencing systems. In the air–ground communication example, however, the messages of low quality and intelligibility delivered to the air traffic controllers may have disastrous effects. The situation in long-distance communication, paging systems, and ground–air communication is somewhat simpler, since the noise is added to the speech at the channel and at the receiving end, respectively, rather than at the source location. Hence, the clean signal can be "immunized" prior to being affected by the noise [6]–[13]. In suboptimal quantization of speech signals, the quantized signal is considered a noisy version of the clean signal [14], [15]. Hence, enhancement can be applied to reduce the quantization noise, provided that quantization was not optimally performed.

The foregoing discussion demonstrates that speech enhancement has three major goals:

t1) to improve perceptual aspects (e.g., quality, intelligi-

Manuscript received June 19, 1990; revised November 26, 1991.

The author is with the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974, and with the C^3I Research Center, George Mason University, Fairfax, VA 22030.

IEEE Log Number 9205056.

bility) of a given sample function of degraded speech signal;

t2) to increase robustness of speech coders to input noise;

t3) to increase robustness of speech recognition systems to input noise.

The quality of speech signals is a subjective measure which reflects on the way the signal is perceived by listeners. It can be expressed in terms of how pleasant the signal sounds or how much effort is required on behalf of the listeners in order to understand the message. Intelligibility, on the other hand, is an objective measure of the amount of information which can be extracted by listeners from the given signal, whether the signal is clean or noisy. A given signal may be of high quality and low intelligibility, and vice versa. Hence, the two measures are independent of each other. Both the quality and the intelligibility of a set of given signals are evaluated based on tests performed on human listeners. Since no mathematical quantification of these measures, in terms of closed-form perceptually meaningful distortion measures, is known, algorithms for tasks t1 and t2 above are difficult to design and evaluate. Task t3 is significantly simpler since the problem is that of decoding the signal into a finite number of classes, and the ultimate goal can be easily formulated in mathematical terms. Usually the problem is that of designing decoders which minimize the probability of recognition error.

Speech enhancement systems which can operate on the clean signal prior to its degradation by noise achieve significant improvement in the intelligibility of the noisy signal [6]–[13]. This is done by first applying high-pass filtering to the clean signal, then normalizing the amplitude of the filtered signal so that it remains constant within a certain range, and, finally, by restoring the original power of the clean signal [7], [8]. These operations emphasize perceptually important components of the signal. High-pass filtering emphasizes the signals in the frequency band of the second formant (resonant) compared with the much stronger signals in the frequency band of the first formant, since the former signals are significantly more important to the intelligibility of the speech signal [9], [10]. Amplitude normalization and power restoration attempt to increase the relative amplitude of consonants compared with vowels, since consonants are considerably more important for signal intelligibility than vowels [11]–[13].

Speech enhancement systems which can operate on the signal only after it has been contaminated by noise primarily improve the quality of the noisy signal at the expense of some intelligibility loss [3], [17]. Increasing the intelligibility of the noisy signal has been accomplished by only one speech enhancement system [16], but in that case the quality of the enhanced signal was significantly worse than that of the noisy signal. Greater success has been achieved in applying speech enhancement to coding and recognition of noisy signals. For speech coding, algorithms have been demonstrated which improve both the quality and the intelligibility of the encoded signal [3], [17], [54]. Similarly, significant improvement in recognition accuracy has been achieved (at least for wideband additive noise) when enhanced classification was used [80]–[82], [85]–[107].

The speech enhancement problem consists of a family of subproblems characterized by the type of the noise source, the way the noise interacts with the clean signal, the number of voice channels, or microphone outputs, available for enhancement, and the nature of the speech communication system. The noise, or the interfering signals, may, for example, be due to competitive speakers, background sounds (music, fans, machines, door slamming, wind, traffic), room reverberation, or random channel noise. The noise may accompany the original signal at the source location, over communication channels, or at the receiving end. It may affect the original signal in an additive, multiplicative, or convolutional manner. Furthermore, the noise may be statistically dependent or independent of the clean signal. The number of voice channels available for enhancement is an important factor in designing speech enhancement systems. In general, the larger the number of microphones, the easier the speech enhancement task. For example, if an auxiliary microphone whose primary purpose is to pick up the noise is available, then the output of that microphone can be used in adaptive cancellation of the noise from the main microphone [18]. The communication system for which speech enhancement is designed can simply be a recording which has to be played to audience, a man–machine communication system (speech recognizer), a digital communication system, etc.

This paper focuses on enhancement of speech signals which have been degraded by statistically independent additive noise. The discussion is not restricted to any particular type of noise (e.g., white noise) but rather considers a wide class of noise sources, which will be specified below. Furthermore, the output of only one microphone which contains the noisy signal is assumed available for enhancement. This setup constitutes one of the most difficult situations of speech enhancement, since no reference signal to the noise is assumed available, and the clean speech cannot be preprocessed prior to being affected by the noise. Information theoretic approaches to the three major problems, t1–t3, of speech enhancement are considered. Specifically, problem t1 is studied from an estimation point of view in which the clean signal is estimated from the noisy signal. Problems t2 and t3 are treated as source coding and signal classification problems, respectively. Using this approach, speech enhancement becomes a set of particular problems in estimation and information theory, for which solutions could be found if the joint statistics of the signal and noise, and a perceptually meaningful distortion measure, were *explicitly* available. Since in practice neither the statistics of the signal and noise nor the most meaningful distortion measure is explicitly known, suboptimal solutions which capitalize on statistical models for the speech signal and noise and on distortion measures which are either mathematically tractable or capture some properties of the auditory system [156], [129]–[131] have been proposed. Such approaches result in model-based speech enhancement systems.

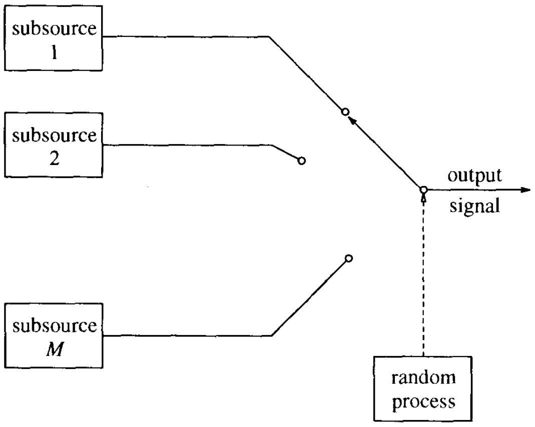


Fig. 1. Composite source model.

The purpose of this paper is to integrate the recent research on the model-based speech enhancement approach, which has dominated the field in recent years, and to provide a unified statistical framework for the three basic problems of speech enhancement. In particular, speech enhancement approaches which use composite source models for the signal and noise, and a wide class of distortion measures, are studied. The composite source model is the most general statistical model known for speech signals, and it has proven extremely useful in speech recognition and enhancement applications. This model can also be useful for a wide class of noise sources encountered in practice, e.g., wideband noise, a mixture of noise sources, and competitive speech.

A block diagram of the composite source model is shown in Fig. 1. A composite source model comprises a finite set of statistically independent subsources which are controlled by a switch [55, p. 177]. Each subsource represents a particular class of statistically similar speech sounds, and the probability distribution (PD) of the subsource is assumed parametric of some given form. The position of the switch at each time instant is randomly selected according to some probability law. The selected position defines the state of the source at the given time instant. Normally, each subsource is assumed a statistically independent identically distributed (iid) Gaussian vector source, and the switch is assumed to be controlled by a first-order Markov chain. Thus, speech signals are assumed to be composed of sounds generated by a finite number of Gaussian vector sources, and transitions from one sound to another are done in a Markovian manner. The model obtained in this way is referred to as a hidden Markov model (HMM) in the speech literature [108]–[110]. This model is also known as a Markov source [111], [112, p. 63], or as a probabilistic function of the Markov chain [113]–[116].

Two distortion measures have been commonly used in enhancing speech signals. These are the mean squared error (MSE) and the uniform distortion measure [156, p. 55] which leads to the maximum *a posteriori* (MAP) estimation approach. These measures are mathematically tractable, and under certain conditions on the PD's of the signal and noise (see Section III), they result in estimators which are optimal for a large set of distortion measures, e.g., all convex difference distortion measures. Hence, these measures are potentially useful for speech signals, since the most meaningful distortion measure is likely to be approximated by some member of that set.

This paper is organized as follows. In Section II some milestones in the development of speech enhancement are reviewed. Section III provides the rationale for the HMM-based speech enhancement approach. In Section IV the three problems of speech enhancement are mathematically formulated, the relationships between them are discussed, and the statistical knowledge needed for their solutions is summarized. Section V deals with the acquisition of such statistical knowledge from training data of clean signals and noise using HMM's. In Section VI the solutions of the signal estimation problem using the MSE and the MAP distortion measures are studied. In Section VII the source coding problem is studied using the MSE and the Itakura–Saito distortion measures. In Section VIII the minimum probability of error signal classification is studied. The performance of these speech enhancement systems in a standard test case, where the clean signal is corrupted by additive Gaussian white noise, is provided in Section IX. A discussion on the model-based approach, which leads to several future research directions, is given in Section X.

II. Milestones in Speech Enhancement Development

A pioneering contribution to the theory and applications of speech enhancement was made by Lim and Oppenheim [1], [3]. In their landmark paper [3], Lim and Oppenheim formulated the speech enhancement problem, classified the different speech enhancement systems known at that time according to the principles they have been based upon, and evaluated and compared these systems. The major conclusions in [3] were as follows:

c1) The usefulness of speech enhancement systems whose task is t1 is in improving the quality of the degraded speech rather than in raising its intelligibility. Quality improvement is important since it reduces listeners' fatigue and hence indirectly contributes to elevating the intelligibility of the given speech material.

c2) Speech enhancement is useful in improving both the quality and the intelligibility of autoregressive (AR) model based speech coding systems [157]. In these systems, speech is modeled as a time-varying AR process whose parameters remain constant over relatively short time frames of 20–40 ms. These parameters, as well as the pitch period of voiced signals (e.g., vowels), are estimated from the noisy signal and transmitted over the channel. The received parameters are used for synthesizing the signal using the linear predictive vocoder [157].

c3) The accomplishments of speech enhancement discussed in c1 and c2 were achieved using the "spec-

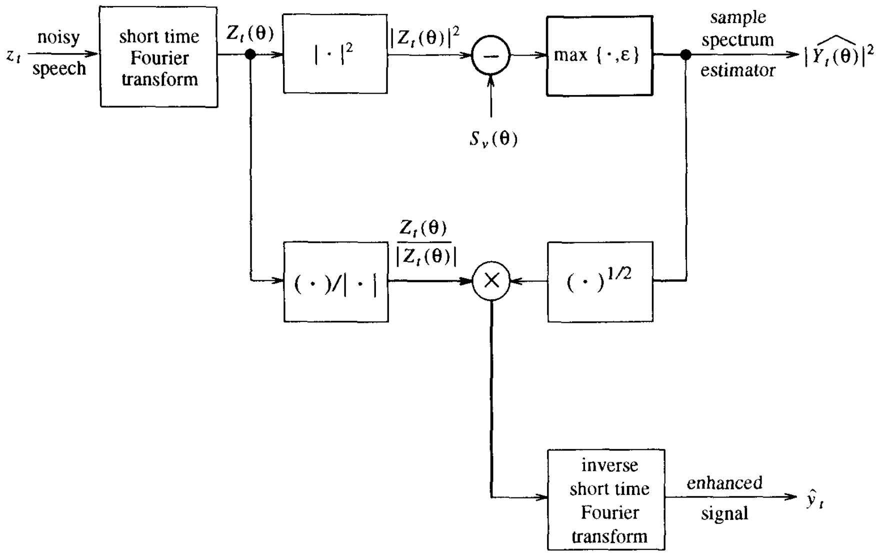


Fig. 2. The spectral subtraction approach.

tral subtraction" estimation approach and its derivatives [17], [20]–[26], [54], [56]. Linear predictive encoding of degraded speech was performed by first estimating the sample autocorrelation of the clean signal and then applying standard AR modeling [157] to the estimated autocorrelation function.

The spectral subtraction estimation approach is suitable for enhancing speech signals degraded by uncorrelated additive noise. It is an approach for estimating the power spectral density of the clean signal by subtracting an estimate of the power spectral density of the noise process from an estimate of the power spectral density of the degraded signal. The estimation is performed on a frame-by-frame basis, where each frame consists of 20–40 ms of speech samples. The sample spectrum (or the periodogram) estimator of the power spectral density of the noisy signal is usually employed in the spectral subtraction approach, thus resulting in an estimator for the sample spectrum of the clean signal. An estimator for the sample autocorrelation function of the clean signal is obtained from the inverse Fourier transform of the sample spectrum estimator. The square root of the sample spectrum estimator is considered an estimator for the spectral magnitude of the speech signal. A signal estimator is obtained by combining the spectral magnitude estimator with the complex exponential of the phase of the noisy signal.

A block diagram of the spectral subtraction approach is shown in Fig. 2. The noisy vector z_t is given by $z_t = y_t + v_t$, where y_t denotes a K-dimensional vector of the clean signal, v_t denotes a K-dimensional vector of the noise process, and y_t and v_t are assumed uncorrelated. The Fourier transform of z_t normalized by $K^{1/2}$, or the short time Fourier transform of the noisy signal z_t [158], is denoted by $Z_t(\theta)$, $0 \leq \theta \leq 2\pi$. The sample spectrum of z_t is given by $|Z_t(\theta)|^2$. The estimate of the power spectral density of the noise process v_t is denoted by $S_v(\theta)$. The spectral subtraction estimate $\hat{Y}_t(\theta)$ of the Fourier transform of y_t normalized by $K^{1/2}$ is given by

$$\hat{Y}_t(\theta) = \left[|Z_t(\theta)|^2 - S_v(\theta)\right]^{1/2} \frac{Z_t(\theta)}{|Z_t(\theta)|}, \tag{1}$$

provided that the difference of spectral estimates of the noisy signal and the noise process is nonnegative. If this difference becomes negative, then it is usually replaced by an arbitrarily small nonnegative number, say ϵ. The power spectral density of the noise is normally estimated from portions of the noisy signal during which speech is absent and only noise is present. The spectral subtraction signal estimator affects the spectral magnitude of the noisy signal in each frame (vector) while it keeps the phase of that signal intact. From a perceptual point of view this is a desirable property, since the short time spectral magnitude of the clean signal is considerably more important than its short time phase [2], [3], [27], and optimal estimation of the short time spectral magnitude and phase of the clean signal cannot be simultaneously performed [28], [29].

Many variations on the basic spectral subtraction approach have been proposed [3], [17], [20]–[26]. The most popular modifications are those that involve averaging or smoothing of the sample spectrum estimator [17], [26], controlling the amount of subtracted noise [3], [22], and using different degrees of nonlinearity in estimating the spectral magnitude of the clean signal [3], [20]. The latter two modifications are accomplished by the estimator [3]

$$\hat{Y}_t(\theta) = [|Z_t(\theta)|^\alpha - \beta E\{|V(\theta)|^\alpha\}]^{1/\alpha} \frac{Z_t(\theta)}{|Z_t(\theta)|}, \tag{2}$$

where $V(\theta)$ is the Fourier transform of some realization of the noise process normalized by $K^{1/2}$, $\alpha > 0$, and $\beta > 0$. Equation (2) degenerates to the standard spectral subtraction estimator (1) when $\alpha = 2$ and K is sufficiently large so that $S_v(\theta) \approx E\{|V(\theta)|^2\}$.

The spectral subtraction estimation approach has an intuitive basis and is relatively easy to implement. When applied for suppressing wideband noise, its major drawback is that it results in noticeable annoying residual noise which consists of narrow-band signals with time-varying frequencies and amplitudes. This residual noise is therefore referred to as musical noise. The speech in the enhanced signal is, however, reasonably clear. The musical residual noise is obtained, since the spectral subtraction approach often results in negative estimates of randomly selected components of the sample spectrum of the clean signal, and these estimates are arbitrarily replaced by some small nonnegative number. Thus, spectral components of the clean signal are randomly being turned on and off, and an effect of musical noise is perceived.

The simplicity of the spectral subtraction approach, on the one hand, and the fact that it "almost" does the job, on the other hand, attracted the attention of many researchers, who tried to provide a better theoretical basis for this approach and to improve its performance by affecting the tonal characteristic of the residual noise [3], [17], [20], [22], [24], [25]. Notable work in this area is that of McAulay and Malpass [24], who formulated the spectral subtraction approach as a maximum likelihood (ML) estimation problem of the variance of each spectral component of the original clean signal. This has also been done in a closely related work by Feit [21]. McAulay and Malpass achieved significant improvement in the performance of the spectral subtraction approach by modifying the estimator to take into account the fact that speech is not always present in the noisy observations. Thus, the notion that the clean signal can be in different "states," i.e., silence and nonsilence, and that the estimator of the clean signal should be composed of individual estimators for the signals in these states was applied. The composite estimator in this approach comprises a weighted sum of the individual estimators for the signals in the two states, where the weights are the posterior probabilities of the two states given the noisy signal. Since the optimal estimator of the clean signal given that this signal is absent from the noisy observations equals zero, the resulting composite estimator is simply the product of the estimator for the clean signal given that this signal is present in the noisy observations and the posterior probability of signal presence given the noisy signal. A block diagram of this estimator is shown in Fig. 3. In [24], the product of a spectral subtraction signal estimator and a parametric posterior probability was used. The two-state signal model was used in [28] and [30] in developing minimum mean squared error (MMSE) signal estimators for speech enhancement applications.

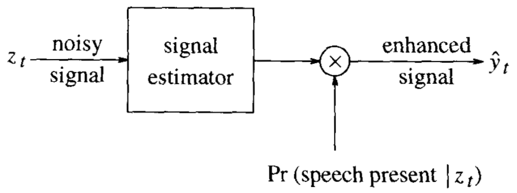


Fig. 3. Two-state spectral subtraction approach (McAulay and Malpass [24]).

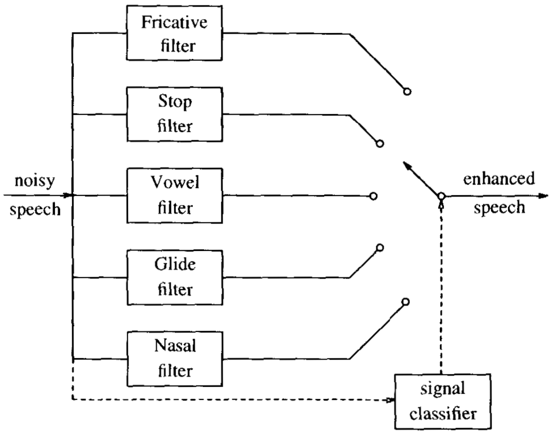


Fig. 4. Detection–estimation approach. (Drucker [31]).

The idea of using different estimators for different classes or states of the speech signal was first introduced by Drucker [31]. In his speech enhancement approach, five categories of speech signals, comprising fricatives, stops, vowels, glides, and nasals, are considered. The noisy speech is first classified into one of the five sound groups and then enhanced using a "matched filter" designed for the chosen class. A block diagram of this system is shown in Fig. 4. The filters used in [31] were simply low-pass, high-pass, and band-pass filters whose bandwidths were determined from known properties of the signals in each class. A major difference between the approaches of McAulay and Malpass [24] and Drucker [31] is that in [24] a "soft decision" is made in combining the outputs of the filters for the two classes of speech, while in [31] a "hard decision" is applied in choosing the appropriate filter for the given signal. As we shall see, the speech enhancement approaches studied in this paper are closely related to these two fundamental approaches, with the advantage that the filters and the decision rules are systematically designed using training sequences of speech and noise data.

A different rather original approach for speech enhancement was proposed by Lim and Oppenheim [19]. In this approach, a time-varying AR model is attributed to the speech signal, and both the model and the signal are estimated from the given noisy signal using the MAP estimation approach. The maximization of the appropriate likelihood function is iteratively performed, once over the AR model assuming that the clean signal is available and then over the clean signal using the estimated model and an assumed known estimate of the power spectral density of the noise. If a vector of the speech signal and its AR model are estimated from the corresponding vector of noisy signal, however, the number of unknowns, i.e., the number of samples of the clean signal plus the number of model parameters, is larger than the number of noisy observations. Hence, the variance of the estimator of either the signal or the model parameters cannot be made arbitrarily small. This difficulty has been observed in practice and has led to algorithms in which the signal and its model for a given vector are estimated from the corresponding vector of noisy signal as well as from adjacent vectors of that signal [32], [91].

As we shall see, a similar time-varying AR model for the speech signal is used in the model-based speech enhancement approach considered here. This model, however, is estimated from training data of clean speech signals rather than from the given noisy process. Thus, the problem of insufficient available data for estimation of the signal and its model does not arise.

Lim and Oppenheim's approach was used in [35] and [57, case 7]. A variation on this approach was developed in [33] and [34], where the time-varying AR model is first estimated from the noisy signal and then used for constructing the estimator for the clean signal.

III. The Model-Based Approach Rationale

As we have seen in Section I, speech enhancement comprises a family of problems in estimation and information theory. It involves estimation of the clean signal from a given sample function of the noisy signal (task t1), encoding of clean speech signals when only noisy versions of these signals are available, and recognition of degraded acoustic signals (task t3). Its solution within this framework requires explicit knowledge of

i. the joint statistics of the clean signal and the noise process, and
ii. a perceptually meaningful distortion measure for speech signals.

If such knowledge were available, task t1 could be optimally accomplished using the estimator which minimizes the expected value of the distortion measure between the clean and the estimated signals. The design of such an estimator can be performed by minimizing the conditional mean of the original distortion measure given the degraded signal [159]. Task t2 could be similarly fulfilled by using the encoder which operates on the noisy signal but approximates the clean signal so that the resulting distortion is minimum. This is not a standard source coding problem, since the encoder does not operate on the same source it represents. It was shown in [58]–[62], however, that the problem can be transformed into a standard source coding problem, in which the encoder operates and approximates the noisy source, if encoding is performed by minimizing the expected value of a modified distortion measure. The modified distortion measure is defined similarly to [159] as the conditional mean of the original distortion measure between the clean and the encoded signal given the degraded signal. Task t3 could be carried out by using the classifier which minimizes the probability of classification error, exactly in the same way this is done when clean signals are available. For, after all, there is no principal difference between the two problems, since the goal in each case is to partition the sample space of the given signal into a finite number of classes which correspond to the number of words in the vocabulary. Minimum probability of classification error is achieved in both cases by applying the MAP decision rule [156] to the given signals, where in each case this decision rule is implemented using the PD's of the given signals from the different words in the vocabulary [126], [80]–[82].

For speech signals which have been degraded by statistically independent noise, the marginal PD's of the clean signal and the noise process must be explicitly known. In practice, however, these PD's are not explicitly available, but they can be estimated from training data of speech and noise. Furthermore, the most perceptually meaningful distortion measure is not explicitly known. Hence, the above theoretical approach can be applied as a two-step procedure in which the statistics of the signal and noise are first estimated and then used together with currently available distortion measures to solve the enhancement problem of interest. The optimality of the two-step enhancement approach depends on the specific estimators used for the unknown PD's. For example, if the sample distributions of the clean signal and the noise process are used to implement the conditional mean estimator, then under appropriate stationary and ergodic assumptions, the sample average estimator can be shown to converge to the true conditional mean with probability 1 when the number of observations tends to infinity [160]. Two versions of the sample average estimator were proposed in [36] and [63] and applied to estimation and encoding of degraded speech signals, respectively.

In spite of its potential asymptotic optimality under certain conditions, the two-step enhancement approach with the sample distribution estimator is impractical, since it usually requires the speech enhancement system to have substantial memory and computational resources. Consider for example the estimation of the conditional mean of a given distortion measure using the sample average estimator. Since the conditional mean is a function of the degraded signal, the sample average must be recalculated for *each* newly observed vector of the degraded signal. This means that the training data of the clean signal and of the noise process must be available at the speech enhancement system, and they must be reapplied for each new vector of the degraded signal. This results in a speech enhancement system whose complexity is proportional to the amount of training data, which normally is very large in order to provide a rich collection of examples from the speech signal and the noise process.

A tractable alternative to the sample distribution estimator is an estimator obtained from parametric modeling of the PD of the source [161, pp. 83–98]. The source here refers to either the clean signal or the noise process. The parametric model is chosen based upon *a priori* theoretical knowledge about the nature of the source as well as empirical observations obtained from that source. Once the model is chosen, its parameters are estimated from training data generated by the source. This approach has three major advantages. First, an appropriately chosen model whose parameters were reliably estimated can provide a better estimate of the PD than the sample distribution estimate, since the chosen model will be capable of characterizing any sample function of the source, not just the sample functions observed in the training set. Second, since the number of parameters of each model is usually small, these parameters can be reliably estimated from a relatively small amount of training data. Third, once the parameters of the models have been estimated, closed-form expressions for the desired estimators (e.g., of the modified distortion measure) and the decision rule can be derived. Thus, only the parameters of the models rather than the entire training data have to be stored in the speech enhancement system.

A parametric model for the speech signal must be capable of providing a reasonable representation of at least the second-order statistics of this signal. These statistics have been shown to be extremely useful in speech processing as they can be used for synthesizing intelligible signals [138], [140], and for recognizing speech signals [108], [110]. By the second-order statistics we mean the different spectra of speech signals as well as the time–frequency correlation of those signals. This correlation can be extremely useful for speech enhancement applications, since it imposes smoothness constraints and thus significantly improves the robustness of the signal estimator.

A useful class of models for speech signals, which can be designed to satisfy the above requirements, is that of HMM's introduced in Section I. The Gaussian subsources represent clusters of spectrally similar acoustic signals. Hence, the power spectral densities of those subsources represent the different spectra of the speech signal or spectral prototypes of that signal. The Markovian states in the HMM provide a Markov model for the evolution of speech spectra or a model for the correlation between those spectra, since states are associated with signal spectra. Correlation between samples of each vector of the signal can be taken into account by the HMM if the covariance matrix of each Gaussian subsource is not diagonal. In the standard HMM described in Section I, vectors generated from a given sequence of states are assumed statistically independent, since each subsource is an iid vector source and the subsources are statistically independent. This model can be extended to take into account correlation between speech vectors, for example, by assuming that a sequence of vectors generated from a given sequence of states is a first-order Markov process [119], [120]. Thus, the HMM can represent both the intervector and the intravector correlation.

HMM's have been widely accepted as reliable statistical models for speech signals. They have been substantiated by speech production theory [108]–[110], [117], and spoken language theory [118]. In the first interpretation, each Gaussian subsource represents acoustic signals generated from a fixed configuration of the vocal tract. The vocal tract is considered an all-pole filter for the spectrally flat (Gaussian white noise for unvoiced signals and an impulse function for one pitch period of voiced signals) excitation signal [157]. The set of all possible vocal tract shapes is represented by a finite number of configurations which equals the number of states in the HMM. This number should approximately be 1024, as has observed in speech coding using vector quantization (VQ) [135]–[140]. In the spoken language interpretation, each subsource represents acoustic signals corresponding to a particular phoneme. Hence, the number of states equals the number of different phonemes in the language (approximately 42 for English [158]). Each interpretation of the HMM results in different constraints which affect the design of the model from training data. The first interpretation is useful for speech enhancement applications, while the second interpretation is commonly used in speech recognition applications.

HMM's can also be useful for many interfering processes [29], [80]. For example, a single-state Gaussian HMM is a suitable model for iid Gaussian vector noise sources, of which Gaussian white noise is a particular case. Furthermore, an HMM similar to that used for the speech signal can be a model for a competitive speech signal. The speech enhancement approach which uses HMM's for both the speech signal and the noise is referred to as the statistical-HMM-based speech enhancement approach.

Given HMM's for the signal and noise, a distortion measure must be chosen in order to solve the speech enhancement problems t1 and t2. A mathematically tractable choice is the MSE distortion measure. This distortion measure is potentially useful for speech enhancement applications for the following reasons:

1) If the HMM's for the signal and noise are such that the posterior probability density function (pdf) of the clean signal given the noisy signal is symmetric about its mean, then the MMSE signal estimator is optimal for a large class of difference distortion measures, not only the MSE measure [156, pp. 60–63]. This class includes all convex difference distortion measures. Since the most perceptually significant distortion measure for speech signals is unknown, it would be desirable to optimize the estimator for the largest possible class of distortion measures, in the hope that the most meaningful measure either belongs to that class or can be well approximated by members of that class.
2) The MMSE estimator of the sample spectrum of the signal is the optimal preprocessor in AR model VQ using the Itakura–Saito distortion measure [62]. In this important application, a finite number of AR models for the clean signal is designed [130], [137], [138], and used for low-bit-rate speech coding. The Itakura–Saito distortion measure (see subsection IV-B) is commonly used, since it is believed to be perceptually significant [157]. This measure is a function of the ratio of the power spectral densities of the source and the model.
3) The MMSE estimator of the signal is the optimal preprocessor in MMSE speech coding given noisy signals [64]. In particular, this estimator is the optimal preprocessor in MMSE waveform VQ using the generalized Llyod algorithm [62].

4) The *causal* MMSE estimator of the signal is the optimal preprocessor in minimum probability of error classification of *any* finite variance *continuous* time signal contaminated by white Gaussian noise [84].

The MMSE estimator using HMM's may, however, be difficult to derive if elaborate HMM's are used [119], [120]. In this case, the MAP estimator of the signal, which can be efficiently calculated using the EM (expectation-maximization) algorithm [147], [148], is often considered a good alternative. MAP estimation is an approximate minimum average distortion estimation approach for the uniform difference distortion measure (see subsection IV-A) [156, p. 55]. This distortion measure assigns zero distortion for estimates in the immediate neighborhood of the clean signal, and uniform distortion for estimates outside this neighborhood. Assuming that the MAP estimator is optimal (not approximately optimal) for this nonconvex distortion measure, then it is also optimal for all symmetric nondecreasing distortion measures, provided that the posterior pdf of the clean signal given the noisy signal is unimodal, that it is symmetric about its mean, and that both the distortion measure and the posterior pdf satisfy

$$\lim_{\epsilon \to \infty} d(\epsilon) p_{Y|Z}(\epsilon|z) = 0, \tag{3}$$

where $d(\epsilon)$ is the difference distortion measure, ϵ is the estimation error, and $p_{Y|Z}(y|z)$ is the posterior pdf of the clean signal Y given the noisy signal Z [156, p. 61].

The HMM-based speech enhancement approach was developed and studied in [29], [37], [40], [41], [79], [80], [82], [83], and [107]. In [29], [37], [40], and [41], MMSE and MAP signal estimators were developed, and their theoretical as well as experimental performance was studied. In [79], signal coding using AR model VQ in the Itakura–Saito sense was studied. In [80], [82]–[83], and [107] minimum probability of error signal classification systems were developed. These approaches resulted in very intuitive speech enhancement systems. For example, the signal estimators resulted in a set of predesigned filters, one for each class or state of acoustic signals, and in a set of posterior probabilities for the occurrences of the states given the noisy signal. In the MMSE estimation approach, all filters are concurrently applied to each vector of the noisy signal, and a weighted sum of the filters' outputs is used as the signal estimator, where the weights are the posterior probabilities of the states. Thus this approach results in a "soft-decision" filter for the noisy signal. In the MAP estimation approach, the most likely filter at each time instant is effectively applied to the noisy signal, thus resulting in a "hard-decision" filter selection rule.

In [38], [39], [57], [81], and [85]–[87], speech enhancement systems were designed using the closely related class of mixture models [162] for the clean signal. The noise is assumed Gaussian with known power spectral density. In this model, vectors of the clean signal are assumed statistically independent, and the PD of each vector of the signal is assumed a finite mixture of Gaussian PD's. Each mixture component represents a cluster of spectrally similar speech signals. The power spectral density of each cluster is referred to as a template of the speech signal. The collection of all templates is commonly estimated from VQ of the training data. The mixture model is equivalent to a single-state HMM whose state-dependent PD constitutes a mixture of Gaussian PD's (see subsection V-A) [121], [37]. The major difference between the mixture model and the standard HMM is that the latter model imposes a Markovian constraint on the evolution of states of the signal, while no such constraint exists in the mixture model. Multiple-state Markovian models are better for speech signals, since the signal may naturally be in different states, and these states must be statistically dependent due to the temporal spectral correlation of the signal.

In [38] and [39], the templates of the mixture model for the clean signal were supplemented by signal parameters estimated from the noisy signal, and the resulting models were used for synthesizing the clean signal. In [38], templates corresponding to a finite set of AR models for the speech signal were considered, while in [39] templates corresponding to a finite set of harmonic models were used. The signal parameters in both cases are the pitch period and the signal voicing. In [38], synthesis was performed using the linear predictive vocoder [157], where each signal vector was synthesized using the nearest neighbor AR template to the noisy vector. In [39], the harmonic zero-phase sine-wave coder was used [163], and each signal vector was synthesized as a weighted sum of harmonic signals, where the weights were the posterior probabilities of the templates given the noisy signal. The approaches in [38] and [39] are therefore based upon "hard" and "soft" decision rules, respectively.

A major difference between model-based estimation and model-based synthesis is that in the former approach the model is used for designing *filters* for the noisy signal, while in the latter approach the model is used for *synthesizing* the desired signal. Hence, under ideal conditions in which the input signal is clean, the model-based estimation approach provides the same clean input signal while the model-based synthesis approach provides signals which differ from the clean signal. The reasons are that for infinite input signal to noise ratio (SNR), the filters are transparent, while synthesis models do not normally achieve perfect reconstruction of the original signal but rather an approximation of that signal in some given sense. The ultimate quality of the synthesized signals in [38] and [39] will therefore be upper bounded by the performance of the linear predictive vocoder and the harmonic zero-phase sine-wave coder, respectively.

Mixture models with Gaussian components were also applied to coding and recognizing noisy speech [57], [81], [85]–[87]. In [57, cases 2, 3, 5], a single mixture component Gaussian AR model was attributed to the speech signal, and MMSE estimators of the signal were implemented using Wiener and Kalman filtering. These estimators were applied as preprocessors for waveform VQ of noisy speech. In [85] and [86], approximate MMSE estimators of the logarithm of the sample spectrum of the clean signal were developed and used as preprocessors for recognition of noisy signals. An exact HMM-based MMSE estimator for the logarithm of the sample spectrum of the clean signal was derived in [29] and [30]. In [87], an approximate MAP signal estimator was developed and applied to word spotting in speech recognition applications. In [81] minimum probability of error recognition of noisy speech was performed using estimates of the PD's of the noisy signal given each word in the vocabulary.

Since mixture models are particular case HMM's, as noted before, we continue our discussion focusing only on HMM's.

IV. Speech Enhancement Problem Formulation

In this section the mathematical formulations of the three basic problems of speech enhancement are provided. The HMM-based solutions are reviewed in subsequent sections.

Let $y \triangleq \{y_t, t = 0, \cdots, T\}$, $y_t \in R^K$, be a sequence of K-dimensional vectors of the clean signal. Let $v \triangleq \{v_t, t = 0, \cdots, T\}$, $v_t \in R^K$, be a sequence of K-dimensional vectors of the noise process. Let $z \triangleq \{z_t, t = 0, \cdots, T\}$, $z_t \in R^K$, be a sequence of K-dimensional vectors of the noisy signal. We assume that $z_t = y_t + v_t$ for $0 \leq t \leq T$, and that y and v are statistically independent. Let $p_{\lambda_s}(y)$ denote the pdf of the HMM for the clean signal, where λ_s denotes the parameter set of that model. Let $p_{\lambda_v}(v)$ denote the pdf of the HMM for the noise process, where λ_v denotes the parameter set of that model. Let $p_\lambda(z)$ denote the pdf of the model for the noisy signal, where $\lambda \triangleq (\lambda_s, \lambda_v)$. The three speech enhancement problems are formulated as follows.

A. Signal Estimation (Problem t1)

Let $z_r^s \triangleq \{z_r, \cdots, z_s\}$, $0 \leq r \leq s \leq T$, be a sequence of observed vectors from the noisy signal. Let $w_t(z_r^s) : R^{K(s-r+1)} \rightarrow R^K$ be an estimator of the clean signal vector y_t given the observations z_r^s. If $r = s = t$, "point estimation" of y_t from z_t is considered. For $\{r = 0, s = t\}$, causal estimation of y_t from the entire set of vectors of the noisy signal observed up till time t is performed. If $s > t$, then noncausal estimation of y_t is considered. Note that causality here is defined with respect to vectors of the signal rather than with respect to the samples of each vector. In fact, when $\{r = 0, s = t\}$, all samples of the vector y_t but the last one are estimated in a noncausal manner. Noncausal estimation should be better than causal estimation, but it can only be applied when delays in the estimated signal can be tolerated, e.g., in speech recognition applications where classification is naturally performed based on the entire sample function of the acoustic signal. Causal and noncausal estimation are also referred to as filtering and smoothing, respectively.

Let $d(y_t, w_t)$ be a given distortion measure between y_t and $w_t(z_r^s)$ that satisfies $E\{d(y_t, w_t)\} < \infty$, where $E\{\cdot\}$ denotes the mathematical expectation. The signal estimation problem is that of finding w_t which minimizes the average estimation distortion $E\{d(y_t, w_t)\}$. Since [160, eq. 5.9.4, lemma 5.9.2]

$$E\{d(y_t, w_t)\} \geq E\{\min_{w_t} E\{d(y_t, w_t)|z_r^s\}\}, \tag{4}$$

the estimator can be obtained from [159]

$$\min_{w_t} E\{d(y_t, w_t)|z_r^s\} = \min_{w_t} \int d(y_t, w_t) p_\lambda(y_t|z_r^s) dy_t, \tag{5}$$

where $E\{d(y_t, w_t)|z_r^s\}$ is the modified distortion measure, and $p_\lambda(y_t|z_r^s)$ is the conditional pdf of y_t given z_r^s. For $r = 0$ and $s = t$, for example,

$$p_\lambda(y_t|z_0^t) = \frac{\int p_{\lambda_s}(y_0^t) p_{\lambda_v}(z_0^t|y_0^t) dy_0^{t-1}}{\int p_{\lambda_s}(y_0^t) p_{\lambda_v}(z_0^t|y_0^t) dy_0^t}. \tag{6}$$

Two distortion measures are of particular interest here, as discussed in Section III. The first is the squared error distortion measure given by

$$d(y_t, w_t) \triangleq \|y_t - w_t\|^2, \tag{7}$$

where $\|\cdot\|$ denotes the usual norm in the Euclidean space. The estimator which minimizes the expected value of (7) can be easily obtained from (5). This results in the well-known MMSE estimator [159]

$$\begin{aligned} \hat{y}_t &= E\{y_t|z_r^s\} \\ &= \int y_t p_\lambda(y_t|z_r^s) dy_t. \end{aligned} \tag{8}$$

The second distortion measure is the uniform distortion measure given by [156, p. 55]

$$d(y_t, w_t) \triangleq \begin{cases} 0 & \|y_t - w_t\| \leq \epsilon/2 \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

for some positive constant ϵ. On substituting (9) into (5) we obtain, for sufficiently small ϵ,

$$\begin{aligned} &\min_{w_t} E\{d(y_t, w_t)|z_r^s\} \\ &\quad = 1 - \max_{w_t} \int_{\|y_t - w_t\| \leq \epsilon/2} p_\lambda(y_t|z_r^s) dy_t \\ &\quad \approx 1 - V_{\epsilon/2} \max_{w_t} \{p_\lambda(y_t|z_r^s)|_{y_t = w_t}\}, \end{aligned} \tag{10}$$

where [164, eq. 4.632.2]

$$V_{\epsilon/2} \triangleq \int_{\|y_t - w_t\| \leq \epsilon/2} dy_t = \frac{\pi^{K/2}}{\Gamma(K/2 + 1)} (\epsilon/2)^K.$$

This means that minimizing the average of the distortion measure (9) is approximately equivalent to MAP estimation of y_t from z_r^s defined by [156]

$$\hat{y}_t = \arg\max_{w_t} \{p_\lambda(y_t|z_r^s)|_{y_t = w_t}\}. \tag{11}$$

A minimum average distortion interpretation for joint MAP estimation of any subset of the vectors y from any subset of the noisy observation vectors z can be similarly given.

The HMM-based MMSE and MAP estimators (8) and (11), respectively, will be studied in Section VI. For each case, causal and noncausal estimation of y_t from z_0^τ, $\tau \geq t$, is considered.

B. Source Coding (Problem t2)

An encoder is a mapping of signal vectors onto a finite number of predesigned vectors called code words. When the encoder operates on the noisy signal, it maps vectors from that signal onto a set of code words designed for the clean signal. Let $\{u_i \in R^K, i = 1, \cdots, N\}$ denote a set of N code words for the clean signal. Let z_r^s be the sequence of vectors of the noisy signal which is used in choosing the code word for the clean vector y_t. Let $Q_t(z_r^s) : R^{K(s-r+1)} \rightarrow \{u_1, \cdots, u_N\}$ be an encoder for the clean vector y_t given the noisy observations z_r^s. We shall focus on the class of memoryless encoders, since the design of encoders with memory is significantly more complicated even when the clean signal is available (see, e.g., [140]–[143]). Furthermore, we shall consider encoding using VQ, since this approach has been mostly used in encoding speech signals and vectors of AR parameters of those signals (see [140]). In this case, $r = s = t$, and the encoder is simply a memoryless vector quantizer. This vector quantizer, denoted here by $Q(z_t)$, can be described by

$$Q(z_t) = \sum_{i=1}^{N} u_i 1_{S_i}(z_t), \tag{12}$$

where $S_i \subset R^K$ denotes the ith partition cell, or the collection of all noisy vectors which are mapped onto the ith code word, and $1_{S_i}(z_t)$ is the characteristic function of the set S_i:

$$1_{S_i}(z_t) \triangleq \begin{cases} 1 & z_t \in S_i \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Let $d(y_t, Q(z_t))$ be a distortion measure between the clean vector y_t and the encoded vector $Q(z_t)$. The design problem is that of finding the quantizer $Q(z_t)$, i.e., $\{u_i, S_i\}$, which minimizes the average distortion $E\{d(y_t, Q(z_t))\}$ obtained in representing y_t by $Q(z_t)$. This problem is similar to designing point estimators for the clean signal, where the difference is that here we are looking for estimators which constitute simple functions of the noisy signal that can only take N different values. Since [160, eq. 5.9.4]

$$E\{d(y_t, Q(z_t))\} = E\{E\{d(y_t, Q(z_t))|z_t\}\}, \tag{14}$$

the quantizer can be designed by minimizing the expected value of the modified distortion measure defined by [55, p. 79], [58]–[62]

$$\begin{aligned} d'(z_t, Q(z_t)) &\triangleq E\{d(y_t, Q(z_t))|z_t\} \\ &= \int d(y_t, Q(z_t)) p_\lambda(y_t|z_t) dy_t. \end{aligned} \tag{15}$$

This is a standard problem in designing vector quantizers for a given source and distortion measure, provided that the modified distortion measure can be explicitly evaluated. Here, the source to be quantized is the noisy source and the distortion measure is the modified distortion measure.

For waveform VQ the squared error distortion measure (7) is normally used [134]–[136], [139]–[141], [143]–[146]. For AR model VQ [135]–[140], where a finite number (N) of AR models for the clean signal is designed, the Itakura–Saito distortion measure is normally used [138], [129]–[131]. The designed set of AR models are used as N code words for low-bit-rate speech coding in linear predictive vocoding [157], [138]. Let $f_t(\theta)$ be the power spectral density of y_t, and $g_i(\theta)$ be the power spectral density of the ith AR code word. Then the Itakura–Saito distortion measure between y_t and the ith code word is given by

$$d(f_t, g_i) = \int_0^{2\pi} \left[\frac{f_t(\theta)}{g_i(\theta)} - \ln \frac{f_t(\theta)}{g_i(\theta)} - 1 \right] \frac{d\theta}{2\pi}. \tag{16}$$

This measure is closely related to the asymptotic minimum discrimination information[1] (MDI) measure and was found to be perceptually meaningful [138], [132], [157].

In practice, $f_t(\theta)$ is approximated by the sample spectrum of y_t, denoted here by $|Y_t(\theta)|^2$, where $Y_t(\theta)$ is the Fourier transform of y_t normalized by $K^{1/2}$. The power spectral density $g_i(\theta)$ is obtained from the covariance matrix of the AR process corresponding to the ith code word using its asymptotic Toeplitz approximation [165], [166]. The covariance matrix for an N_s-order AR model is given by $\sigma_i^2 (A_i^{\#} A_i)^{-1}$, where σ_i^2 is the variance of the innovation process of the AR source, A_i is a $K \times K$ lower triangular Toeplitz matrix in which the first $N_s + 1$ elements of the first column constitute the coefficients of the AR process, and $\#$ denotes Hermitian transposition. The coefficients of the AR process are denoted here by $e_i \triangleq (e_i(0), e_i(1), \cdots, e_i(N_s))$, where $e_i(0) = 1$. Using this notation we have that $g_i(\theta) = \sigma_i^2 / |A_i(\theta)|^2$, where $A_i(\theta)$ is the Fourier transform of the vector e_i, and it is assumed that $|A_i(\theta)|^2 \geq m > 0$ [165]. For $g_i(\theta) = \sigma_i^2 / |A_i(\theta)|^2$, we have that [130], [138]

$$\begin{aligned} d(f_t, g_i) = \sum_{m=-N_s}^{N_s} r_t(m) \frac{r_i(m)}{\sigma_i^2} + \ln \sigma_i^2 - 1 \\ - \int_0^{2\pi} \ln f_t(\theta) \frac{d\theta}{2\pi}, \end{aligned} \tag{17}$$

where $r_t(m)$ and $r_i(m)$ are the sample autocorrelation functions of the signal y_t and the vector of AR coefficients e_i, respectively. These autocorrelation functions are obtained from the inverse Fourier transform of $f_t(\theta)$ and $|A_i(\theta)|^2$, respectively.

The computationally difficult term in the Itakura–Saito measure, $\int_0^{2\pi} \ln f_t(\theta) d\theta / 2\pi$, which is the logarithm of the one-step prediction error [166], is often replaced by $\ln \sigma_t^2$, where σ_t^2 is the variance of the residual error signal obtained from AR modeling of y_t [130], [138]. This substitution preserves the nonnegativeness of the Itakura–Saito measure and has the advantage that σ_t^2 can be easily computed using the Levinson–Durbin algorithm [157]. In this case (17) becomes

$$d(f_t, g_i) = \sum_{m=-N_s}^{N_s} r_t(m) \frac{r_i(m)}{\sigma_i^2} - \ln \frac{\sigma_t^2}{\sigma_i^2} - 1. \tag{18}$$

The distortion measure (18) will be used in this paper only when the Itakura–Saito distortion measure is numerically

[1] The discrimination information measure is also known as the cross-entropy, relative entropy, directed divergence, I-divergence, and Kullback-Leibler number.

calculated, i.e., in reporting experimental results. Otherwise, we shall refer to (16) as the Itakura–Saito distortion measure.

The design of HMM-based MMSE waveform vector quantizers, and of HMM-based AR model vector quantizers in the Itakura–Saito sense, by minimizing the expected value of the modified distortion measures corresponding to (7) and (16), respectively, will be studied in Section VII.

C. Signal Classification (Problem t3)

Let $\{W_i, i = 1, \cdots, J\}$ be a sequence of words in a vocabulary of J words. Assume that z denotes a noisy version of the acoustic signal from some word in the vocabulary. The classification problem is that of associating z with one of the words so that the resulting probability of classification error is minimum. This is a problem of partitioning the sample space of the noisy acoustic signals from all words in the vocabulary into J partition cells. Let $\Omega \triangleq \{\omega_1, \cdots, \omega_J\}$ be a partition of the sample space of the noisy signals. The probability of error associated with this partition is given by

$$P_e(\Omega) = \sum_{i=1}^{J} P(W_i) \int_{z \notin \omega_i} p_\lambda(z|W_i)dz, \qquad (19)$$

where $P(W_i)$ is the *a priori* probability of occurrence of the ith word, and $p_\lambda(z|W_i)$ is the pdf of the model for the noisy signal from the ith word. The minimization of $P_e(\Omega)$ is achieved by the well-known MAP decision rule given by [156]

$$\max_{1 \le i \le J} p_\lambda(z|W_i)P(W_i). \qquad (20)$$

Hence, z is associated with the word W_i for which $p_\lambda(z|W_i)P(W_i)$ is maximum.

Note that the problem of classifying noisy signals is essentially the same as that of classifying clean signals. In both cases, the problem is that of partitioning the sample space of the observed signal into J partition cells. Minimum probability of classification error is achieved by applying the MAP decision rule to the given signal, using $p_\lambda(y|W_i)$ when the observed signal is clean and $p_{\lambda_s}(z|W_i)$ when the observed signal is noisy. No estimation of the clean signal or of any parameter of that signal is required in optimal classification of noisy signals. Obviously, the problem of signal classification is significantly simpler than that of signal estimation, as only the association of the signal with one of the partition cells rather than the recovery of the clean signal from the noisy signal is needed.

The application of the HMM-based MAP decision rule (20) to the recognition of clean and noisy signals will be studied in Section VIII.

D. Comments

The solutions of the three speech enhancement problems presented here are contingent on explicit knowledge of the PD's of the speech signals and the noise process. Since these PD's are not explicitly available, as we argued in Section III, they have been replaced in the above formulations by HMM's whose parameters are estimated from training sequences of speech signals and noise samples. The HMM for the clean speech signal applied to problems t1 and t2 is *universal* in the sense that it constitutes an estimate of the PD of the speech signal for all speech sounds and speakers of the same language. This is in contrast to problem t3, where an individual HMM is designed for the acoustic signals corresponding to each word in the vocabulary. Furthermore, models designed for this application are often suitable for a particular group of speakers who participated in the training procedure. Universal models for the PD of the speech signal have been successfully applied, for example, in designing vector quantizers using the sample distribution of the given training data [134]–[146]. For the universal model to be reliable, the training data must be rich enough so that they represent the largest possible class of speech signals from different speakers.

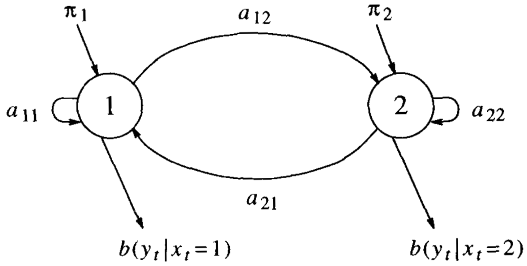


Fig. 5. A two-state first order HMM. For $\alpha, \beta = 1, 2$, π_α denotes the initial state probability, $a_{\alpha\beta}$ denotes the state transition probability, and $b(y_t|x_t = \beta)$ denotes the parametric pdf of the K-dimensional Gaussian output process from state β.

V. Hidden Markov Models

We begin this section by a simple example of a first-order, two-state HMM with stationary Gaussian state-dependent PD's as shown in Fig. 5. The process can begin in the first state with probability π_1 or in the second state with probability $\pi_2 = 1 - \pi_1$. Once in the first state, the process can either stay in this state with probability a_{11} or switch to the second state with probability $a_{12} = 1 - a_{11}$. Similarly, once the second state is visited, the process may stay there with probability a_{22} or return to the first state with probability $a_{21} = 1 - a_{22}$. Each time a state is visited, a K-dimensional zero mean Gaussian random vector y_t whose covariance matrix is Toeplitz is generated. The state dependent pdf of the vector y_t generated from the state x_t at time t is denoted here by $b(y_t|x_t)$. Vectors generated from different states have different covariance matrices, and hence different power spectral densities [165]. Thus, the HMM in this example assumes that the speech signal is composed of a sequence of Gaussian random vectors with two possible power spectral densities, and transitions between vectors are Markovian of first order. Intuitively, since the parameters of the HMM are estimated from training data generated by the signal, the two power spectral densities are prototype spectra of the data, or spectra of the centroids of the data which result from binary clustering using VQ. This intuition was made precise in [122] and [123].

The two-state HMM can be extended to suit very complicated signals such as speech signals. First, a sufficiently large number of states can be used to account for the different spectral prototypes of the speech signal. Second, non-Gaussian state-dependent PD's can be used if one believes that such PD's are more appropriate for the signals generated from each state. Third, different state-dependent PD's for different states can be used. Fourth, each state-dependent PD can be chosen to be a mixture of Gaussian (or any other) PD's. In this case, the signals in each cluster or state will be represented by multiple spectral prototypes. This results in a refined spectral representation of the given training data, while maintaining the convenience of using Gaussian pdf's. Fifth, output vectors from different states can be made statistically dependent. For example, we can assume that the sequence of output vectors from a given sequence of states forms a first-order vector Markov process [119], [120]. This will enhance the capabilities of the HMM to model correlated signals such as speech signals; however, it results in significantly more complicated estimation schemes [120].

In this paper the class of first-order HMM's with Gaussian state-dependent PD's is considered. The subsources of the model are assumed statistically independent. Furthermore, each subsource is assumed an iid vector source. Thus, output vectors from any given sequence of states are statistically independent. Each Gaussian subsource is further assumed to be an AR process of a given order. This allows parameterization (or modeling) of the $K \times K$ covariance matrices of the Gaussian subsources in a way that has proved useful for speech processing applications (see, e.g., [157]). The HMM in this class, and the speech enhancement systems which are based upon such a model, can be easily extended to the case where each state-dependent PD is a mixture of Gaussian AR PD's. Since such model is not conceptually different from a model with a single mixture component per state, only the latter class of models is considered here since it results in significantly simpler notation. Speech enhancement systems which use HMM's with multiple mixture components per state were discussed in [37], [40], and [82]. It should be noted that most of the results discussed here for the AR parameterization of the covariance matrices of the subsources are applicable to other parameterizations of these covariance matrices, for example, the popular parameterization which assumes that these matrices are diagonal [110]. The class of HMM's considered here is mathematically tractable, and has been proven useful in speech recognition [83], [108]–[110], [121], [124] and speech enhancement applications [29], [37], [79], [82].

A more formal definition of HMM's is given next, and the intuition brought up in the beginning of this section is made precise. The HMM's for the speech signal and the noise process are first defined, and then the models for the pdf of the noisy signal, and the pdf of the clean signal given the noisy signal, are provided. As was demonstrated in Section IV, these pdf's play central roles in signal classification, and in signal estimation and coding, respectively. The estimation of the parameter sets of the HMM's from given training data is also discussed in this section.

A. Speech and Noise HMM's

Let M be the number of states of the HMM for the clean signal y. The pdf of this model is given by

$$p_{\lambda_s}(y) = \sum_{x} p_{\lambda_s}(x) p_{\lambda_s}(y|x), \tag{21}$$

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

$$p_{\lambda_s}(x) = \prod_{t=0}^{T} a_{x_{t-1}x_t}, \tag{22}$$

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 $a_{x_{-1}x_0} \overset{\Delta}{=} \pi_{x_0}$ denotes the probability of the initial state x_0. By assumption, the pdf $p_{\lambda_s}(y|x)$ is given by

$$p_{\lambda_s}(y|x) = \prod_{t=0}^{T} p_{\lambda_s}(y_t|x_t) \overset{\Delta}{=} \prod_{t=0}^{T} b(y_t|x_t), \tag{23}$$

where $b(y_t|x_t)$ denotes the pdf of the vector y_t given that this vector was generated from state x_t. For HMM's with Gaussian AR subsources, the pdf $b(y_t|x_t)$ is given by

$$b(y_t|x_t) = \frac{\exp\{-\frac{1}{2} y_t^{\#} S_{x_t}^{-1} y_t\}}{(2\pi)^{K/2} \det^{1/2}(S_{x_t})}, \tag{24}$$

where S_{x_t} is the covariance matrix of the AR process whose order is assumed N_s. This matrix is given by $S_{x_t} = \sigma_{x_t}^2 (A_{x_t}^{\#} A_{x_t})^{-1}$, where $\sigma_{x_t}^2$ and A_{x_t} are defined similarly to σ_i^2 and A_i in subsection IV-B, respectively. The power spectral density of the subsource associated with state x_t is given by $\boldsymbol{S}_{x_t} = \sigma_{x_t}^2 / |A_{x_t}(\theta)|^2$, where $A_{x_t}(\theta)$ denotes the Fourier transform of the first column of A_{x_t}, and it is assumed that $|A_{x_t}(\theta)|^2 \geq m > 0$ (see subsection IV-B).

The parameter set of the HMM for the clean speech is given by $\lambda_s = (\pi, a, S)$, where $\pi \overset{\Delta}{=} \{\pi_\beta\}$, $a \overset{\Delta}{=} \{a_{\alpha\beta}\}$, and $S \overset{\Delta}{=} \{S_\beta\}$, for $\alpha, \beta = 1, \cdots, M$.

When the state-dependent PD constitutes a mixture of L Gaussian PD's, then

$$p_{\lambda_s}(y_t|x_t) = \sum_{h_t} c_{h_t|x_t} b(y_t|x_t, h_t), \tag{25}$$

where $h_t \in \{1, \cdots, L\}$ denotes the mixture component chosen at time t, $c_{h_t|x_t}$ denotes 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). The covariance matrix of this pdf is given by $S_{h_t|x_t}$, and it is defined similarly to S_{x_t}. From (21)–(25), the pdf $p_{\lambda_s}(y)$ can be written as

$$\begin{aligned} p_{\lambda_s}(y) &= \sum_x \sum_h p_{\lambda_s}(x, h, y) \\ &= \sum_x \sum_h p_{\lambda_s}(x) p_{\lambda_s}(h|x) p_{\lambda_s}(y|x, h), \qquad (26) \end{aligned}$$

where $h \overset{\Delta}{=} \{h_t, t = 0, \cdots, T\}$ denotes a sequence of mixture components, and

$$\begin{aligned} p_{\lambda_s}(h|x) &= \prod_{t=0}^{T} p_{\lambda_s}(h_t|x_t) = \prod_{t=0}^{T} c_{h_t|x_t} \\ p_{\lambda_s}(y|x, h) &= \prod_{t=0}^{T} p_{\lambda_s}(y_t|x_t, h_t) = \prod_{t=0}^{T} b(y_t|x_t, h_t). (27) \end{aligned}$$

Comparing (26) and (27) with (21)–(23) shows that there is no principal difference between the two models. The major difference is that the state-dependent pdf's in (26) are double indexed (x_t, h_t), while in (21) they are single indexed (x_t). An HMM with mixture state-dependent PD's can achieve a linear increase in the number of subsources $(M \times L)$ while maintaining the dimension of its state transition probability matrix $M \times M$. In the HMM with a single Gaussian AR PD per state, an increase in the number of subsources (M) implies quadratic increase in the dimension of the state transition probabilities. As mentioned before, only HMM's with a single mixture component per state are considered here.

The model for the PD of the noise process, which is also assumed an HMM with Gaussian AR state-dependent PD's, can be similarly described. Let $\tilde{M}$ be the number of states of this model, and N_v the order of the AR processes. A state sequence of the model corresponding to the noise vectors v will be denoted here by $\tilde{x}$, where $\tilde{x}$ is defined similarly to x. For simplicity of notation, the parameters of the HMM for the clean signal as well as for the noise process will be generically denoted by (π, a, S), where the distinction between the two models will be clear from the indices of these quantities. For example, $a_{x_{t-1}x_t}$ will denote the transition probability for states of the clean signal while $a_{\tilde{x}_{t-1}\tilde{x}_t}$ will denote the transition probability for states of the noise process. Notice that a single-state HMM can be used to model noise sources with statistically iid vectors, e.g., white noise.

B. HMM for Noisy Signal

Given the HMM's for the clean signal and the noise process, p_{λ_s} and p_{λ_v}, respectively, a model for the noisy signal z can be obtained since the signal and noise are assumed statistically independent. Such a model was derived in [29] and is given by

$$p_\lambda(z) = \sum_{\bar{x}} p_\lambda(\bar{x}) p_\lambda(z|\bar{x}), \qquad (28)$$

where $\bar{x} \overset{\Delta}{=} (x, \tilde{x}) = \{(x_t, \tilde{x}_t), t = 0, \cdots, T\}$ denotes the sequence of composite states of the noisy signal. The probability of those states is given by

$$p_\lambda(\bar{x}) = \prod_{t=0}^{T} a_{\bar{x}_{t-1}\bar{x}_t}, \qquad a_{\bar{x}_{t-1}\bar{x}_t} \overset{\Delta}{=} a_{x_{t-1}x_t} a_{\tilde{x}_{t-1}\tilde{x}_t}, \qquad (29)$$

where $a_{\bar{x}_{t-1}\bar{x}_t}$ denotes the state transition probability, and $a_{\bar{x}_{-1}\bar{x}_0} \overset{\Delta}{=} \pi_{\bar{x}_0}$ denotes the initial state probability. The pdf of the noisy vectors given the composite states is given by

$$p_\lambda(z|\bar{x}) = \prod_{t=0}^{T} b(z_t|\bar{x}_t), \qquad (30)$$

where $b(z_t|\bar{x}_t)$ is a Gaussian pdf with zero mean and covariance $S_{\bar{x}_t} \overset{\Delta}{=} S_{x_t} + S_{\tilde{x}_t}$. Hence, the model for the noisy signal is a first-order $M \times \tilde{M}$-state HMM with Gaussian (not AR) state-dependent PD's.

C. Conditional pdf of Clean Speech Given Noisy Signal

The model $p_\lambda(y_t|z_0^\tau)$, $\tau \geq t$, for the conditional pdf of the clean signal y_t given the noisy signal z_0^τ can be derived similarly to (28). From [29], we have that

$$p_\lambda(y_t|z_0^\tau) = \sum_{\bar{x}_t} p_\lambda(\bar{x}_t|z_0^\tau) b(y_t|z_t, \bar{x}_t), \qquad (31)$$

where $p_\lambda(\bar{x}_t|z_0^\tau)$ is the conditional probability of the composite state of the noisy signal at time t given the noisy observations, and $b(y_t|z_t, \bar{x}_t)$ is the conditional pdf of the clean signal at time t given the noisy signal and its composite state at time t. The conditional pdf $b(y_t|z_t, \bar{x}_t)$ is Gaussian with mean and covariance given by

$$\begin{aligned} E(y_t|z_t, \bar{x}_t) &= S_{x_t}(S_{x_t} + S_{\tilde{x}_t})^{-1} z_t \overset{\Delta}{=} H_{\bar{x}_t} z_t \\ Cov(y_t|z_t, \bar{x}_t) &= H_{\bar{x}_t} S_{\tilde{x}_t} \overset{\Delta}{=} \Sigma_{\bar{x}_t}. \qquad (32) \end{aligned}$$

Notice that $H_{\bar{x}_t}$ is the MMSE estimator (Wiener filter) of the signal y_t given the noisy signal z_t, assuming that the clean signal is in state x_t and the noise is in state $\tilde{x}_t$. Furthermore, $\Sigma_{\bar{x}_t}$ is the MMSE associated with this estimator.

The conditional probability $p_\lambda(\bar{x}_t|z_0^\tau)$ in (31), and the pdf $p_\lambda(z_0^\tau)$ in (28), can be efficiently calculated using the "forward–backward" recursive formulas for HMM's (see, e.g., [37, eqs. (25)–(27)]). Specifically, for $0 \leq t \leq \tau$ we have that

$$p_\lambda(\bar{x}_t|z_0^\tau) = \frac{F(\bar{x}_t, z_0^t) B(z_{t+1}^\tau|\bar{x}_t)}{\sum_{\bar{x}_t} F(\bar{x}_t, z_0^t) B(z_{t+1}^\tau|\bar{x}_t)}, \qquad (33)$$

where the "forward" pdf for the noisy signal is given by

$$\begin{aligned} F(\bar{x}_0, z_0) &\overset{\Delta}{=} p_\lambda(\bar{x}_0, z_0) \\ &= \pi_{\bar{x}_0} b(z_0|\bar{x}_0) \\ F(\bar{x}_t, z_0^t) &\overset{\Delta}{=} p_\lambda(\bar{x}_t, z_0^t) \\ &= \sum_{\bar{x}_{t-1}} F(\bar{x}_{t-1}, z_0^{t-1}) a_{\bar{x}_{t-1}\bar{x}_t} b(z_t|\bar{x}_t), \\ &\qquad 0 < t \leq \tau, \qquad (34) \end{aligned}$$

and the "backward" pdf for the noisy signal is given by

$$\begin{aligned} B(z_{\tau+1}^{\tau}|\bar{x}_{\tau}) &\overset{\Delta}{=} 1 \\ B(z_{t+1}^{\tau}|\bar{x}_t) &\overset{\Delta}{=} p_\lambda(z_{t+1}^{\tau}|\bar{x}_t) \\ &= \sum_{\bar{x}_{t+1}} B(z_{t+2}^{\tau}|\bar{x}_{t+1}) a_{\bar{x}_t\bar{x}_{t+1}} b(z_{t+1}|\bar{x}_{t+1}), \\ &\qquad 0 \le t < \tau. \end{aligned} \tag{35}$$

From (34) we have that

$$p_\lambda(z_0^{\tau}) = \sum_{\bar{x}_{\tau}} F(\bar{x}_{\tau}, z_0^{\tau}). \tag{36}$$

D. HMM Training

The parameter set of the HMM for the clean signal (and similarly for the noise process) is normally estimated from training sequences using the ML estimation approach [108], [113]–[116]. This approach is used for two major reasons. First, if speech were strictly a hidden Markov process of the given class, global maximization of the likelihood function could be performed, all transition probabilities are strictly positive, and the output processes have nonzero discrete PD's, then the ML estimator is consistent [113, theorem 3.4] and asymptotically efficient [113, section 5]. Second, there exists a computationally efficient EM algorithm, the Baum algorithm [113]–[116], for local maximization of the likelihood function. In practice, these sufficient conditions for consistency and asymptotic efficiency of the ML estimator are not always satisfied. For example, speech signals are not strictly hidden Markov processes, the output pdf's of the models are often chosen to be continuous rather than discrete, as is the case here, and the maximization of the likelihood function is local rather than global. Nevertheless, reasonably good modeling of speech signals has been obtained using the ML estimation approach. Other approaches for estimating the parameter set of an HMM from given training data, in which some of the above conditions are relaxed, have been proposed, but normally these approaches are significantly more complicated than the ML approach. For example, the MDI approach for hidden Markov modeling proposed in [125] and [126] does not require the speech to be a hidden Markov process. In the MDI approach, the model whose pdf is closest (in the divergence sense) to the pdf of the signal, as characterized by a given set of moments, is chosen.

An ML estimate of the parameter set λ_s of an HMM, given training data $y \overset{\Delta}{=} y_0^T$, is obtained from local maximization of the likelihood function

$$\ln p_{\lambda_s}(y_0^T) \tag{37}$$

subject to

$$\begin{aligned} &\pi_\beta \ge 0, \qquad \sum_{\beta=1}^{M} \pi_\beta = 1 \\ &a_{\alpha\beta} \ge 0, \qquad \sum_{\beta=1}^{M} a_{\alpha\beta} = 1 \\ &S_\beta = \sigma_\beta^2 (A_\beta^{\#} A_\beta)^{-1} \text{ is positive definite} \end{aligned}$$

for $\alpha, \beta = 1, \cdots, M$. The maximization is performed using the Baum algorithm [113]–[116]. Each iteration of the Baum algorithm starts with an old set of parameters, say λ_s, and generates a new set of parameters, say λ_s', using the following reestimation formulas (see, e.g., [37, eqs. (18)–(21), (28)]):

$$\begin{aligned} \pi_\beta' &= p_{\lambda_s}(x_0 = \beta|y_0^T) \\ a_{\alpha\beta}' &= \frac{\sum_{t=1}^{T} p_{\lambda_s}(x_{t-1} = \alpha, x_t = \beta|y_0^T)}{\sum_{\beta=1}^{M}\sum_{t=1}^{T} p_{\lambda_s}(x_{t-1} = \alpha, x_t = \beta|y_0^T)}, \end{aligned} \tag{39}$$

and the parameters of the AR process at state β can be obtained from AR modeling of the autocorrelation function

$$r_\beta'(m) \overset{\Delta}{=} \frac{\sum_{t=0}^{T} p_{\lambda_s}(x_t = \beta|y_0^T) r_t(m)}{\sum_{t=0}^{T} p_{\lambda_s}(x_t = \beta|y_0^T)}, \tag{40}$$

provided that $K \gg N_s$, where

$$r_t(m) = \frac{1}{K} \sum_{k=0}^{K-|m|-1} y_t(k) y_t(k+|m|), \qquad m = -N_s, \cdots, N_s \tag{41}$$

is the sample autocorrelation of y_t, $p_{\lambda_s}(x_{t-1}, x_t|y_0^T)$ is the conditional probability, induced by p_{λ_s}, of the states (x_{t-1}, x_t) given the clean signal y_0^T, and $p_{\lambda_s}(x_t|y_0^T)$ is the marginal probability of $p_{\lambda_s}(x_{t-1}, x_t|y_0^T)$. The reestimation formulas (39) and (40) are valid provided that the terms in the denominators of these expressions are greater than zero. If any of these conditions is not satisfied, then the affected reestimated parameter can be arbitrarily chosen up to the constraints associated with the problem (37), without affecting the likelihood value. For example, if the denominator of (39) equals zero for a particular α, then any $\{a_{\alpha\beta}, \beta = 1, \cdots, M\}$, which satisfies $\sum_{\beta=1}^{M} a_{\alpha\beta} = 1$ can be chosen. The Baum algorithm is stopped when a convergence criterion is satisfied, e.g., when the difference of the values of the likelihood function (37) in two consecutive iterations is smaller than or equal to a given threshold. Convergence of the model sequence generated by the Baum algorithm can be shown using the results from [148].

The probability measures $p_{\lambda_s}(x_{t-1}, x_t|y_0^T)$ and $p_{\lambda_s}(x_t|y_0^T)$ can be efficiently calculated using the "forward–backward" recursive formulas for HMM's in a way similar to the calculation of $p_\lambda(x_t|z_0^{\tau})$ in (33)–(35). Specifically, let y_r^s, $0 \le r \le s \le T$, be the sequence of $s-r+1$ K-dimensional vectors from the training data of clean signals, and let x_r^s be the corresponding sequence of states. Then,

$$\begin{aligned} p_{\lambda_s}(x_{t-1}, x_t|y_0^T) &\overset{\Delta}{=} p_{\lambda_s}(x_{t-1}, x_t|y_0^T) \\ &= \frac{F(x_{t-1}, y_0^{t-1}) B(y_{t+1}^T|x_t) a_{x_{t-1}x_t} b(y_t|x_t)}{\sum_{x_{t-1},x_t} F(x_{t-1}, y_0^{t-1}) B(y_{t+1}^T|x_t) a_{x_{t-1}x_t} b(y_t|x_t)} \\ &\qquad 0 < t \le T, \end{aligned} \tag{42}$$

$$p_{\lambda_s}(x_t|y_0^T) = \frac{F(x_t, y_0^t) B(y_{t+1}^T|x_t)}{\sum_{x_t} F(x_t, y_0^t) B(y_{t+1}^T|x_t)}, \qquad 0 \le t \le T, \tag{43}$$

where the "forward" pdf for the clean signal is given by

$$\begin{aligned} F(x_0, y_0) &\overset{\Delta}{=} p_{\lambda_s}(x_0, y_0) \\ &= \pi_{x_0} b(y_0|x_0) \\ F(x_t, y_0^t) &\overset{\Delta}{=} p_{\lambda_s}(x_t, y_0^t) \\ &= \sum_{x_{t-1}} F(x_{t-1}, y_0^{t-1}) a_{x_{t-1}x_t} b(y_t|x_t), \\ &\qquad 0 < t \leq T, \end{aligned} \tag{44}$$

and the "backward" pdf for the clean signal is given by

$$\begin{aligned} B(y_{T+1}^T|x_T) &\overset{\Delta}{=} 1 \\ B(y_{t+1}^T|x_t) &\overset{\Delta}{=} p_{\lambda_s}(y_{t+1}^T|x_t) \\ &= \sum_{x_{t+1}} B(y_{t+2}^T|x_{t+1}) a_{x_t x_{t+1}} b(y_{t+1}|x_{t+1}), \\ &\qquad 0 \leq t < T. \end{aligned} \tag{45}$$

The likelihood function (37), which has to be evaluated in checking convergence of the Baum algorithm, can be efficiently calculated using

$$\ln p_{\lambda_s}(y_0^T) = \ln \sum_{x_T} F(x_T, y_0^T). \tag{46}$$

The Baum algorithm presented above can be generalized for parameter estimation from N statistically independent training sequences [37], [121]. Such estimation is often required in practice, when either a set of N training data is naturally available, or a long training sequence is artificially broken into N subsequences in order to reduce the complexity (storage and computation) of the algorithm. The reestimation formulas for this case can be found in [37].

The reestimation formulas have an intuitive basis which can be best understood if the likelihood function $p_{\lambda_s}(y) = \sum_x p_{\lambda_s}(x, y)$ (see (21)), is assumed to be dominated by a unique sequence of states [37], [122], [123], [127], [128]. In this case, $p_{\lambda_s}(y) \approx \max_x p_{\lambda_s}(x, y)$, and $p_{\lambda_s}(x_{t-1}, x_t|y_0^T) \approx 1$ if (x_{t-1}, x_t) belongs to the dominating sequence, and $p_{\lambda_s}(x_{t-1}, x_t|y_0^T) \approx 0$ otherwise [37]. Alternatively, $p_{\lambda_s}(x_{t-1}, x_t|y_0^T)$ is a probability measure which is concentrated on the dominating states. Using this assumption, $a'_{\alpha\beta}$ can be interpreted as the number of transitions from state α to state β, normalized by the total number of transitions from state α to any other state (including state β), as occurred on the dominating state sequence of the training data. Similarly, r'_β can be interpreted as the average, or the centroid, of the sample autocorrelation functions which correspond to signal vectors with dominating state β. Thus, $\{r'_\beta\}$ constitute centroids of an M partition of the sample space of the signals in the training data. Applying AR modeling [130] to $\{r'_\beta\}$ results in estimates $\{S_\beta\}$ of the covariance matrices of the signals in the partition cells (clusters) or in estimates of the covariance matrices of the subsources of the HMM. The power spectral densities $\{\boldsymbol{S}_\beta(\theta) = \sigma_\beta^2/|A_\beta(\theta)|^2\}$ associated with these covariance matrices [165] are estimates of the power spectral densities of signals in the different clusters, estimates of the power spectral densities of the output processes from the HMM, or spectral prototypes of the training data.

The case of a dominating sequence of states is of particular interest, since it has often been encountered in practice, especially when the vector dimension K is sufficiently large [122], [123]. The reason is that when K is sufficiently large, speech signals have fairly distinct clusters. Furthermore, the cluster or state of each signal vector can be reliably estimated when a sufficiently large number of observations generated from that state are available. This situation was observed in practice for K in the order of 256 samples at an 8 kHz sampling rate.

Similar interpretations for the reestimation formulas can be given if the likelihood function is not assumed to be dominated by a unique sequence of states. In that case, the relative number of transitions and the average of the autocorrelation functions are the expected number of transitions and the expected value of the sample autocorrelation, respectively (see, e.g., [117]).

A useful initial model for the Baum algorithm is obtained from AR model VQ of the training data using the generalized Lloyd algorithm with the Itakura–Saito distortion measure (see, e.g., [138]). Specifically, if the likelihood function (21) of the training sequence is approximated using the dominant sequence of states, then we find from (22) and (23) that

$$-\frac{1}{K} \ln p_{\lambda_s}(y) \approx \min_x \sum_{t=0}^{T} \frac{1}{K}[-\ln a_{x_{t-1}x_t} - \ln b(y_t|x_t)]. \tag{47}$$

Since $\ln a_{x_{t-1}x_t}/K$ vanishes as $K \to \infty$ (assuming $a_{\alpha\beta} \geq m > 0$ for all α and β), and $-\ln b(y_t|x_t)/K$ approaches the Itakura–Saito distortion measure (up to some additive term which is independent of x_t) when $K \to \infty$ [129], [122], the likelihood function in (47) is closely related (for large K) to the average Itakura–Saito distortion used in designing AR model vector quantizers. The generalized Lloyd algorithm is first applied to the training data for estimating a set of M AR models which represents the initial estimate of $\{S_\beta\}$. Then, the training data are clustered using the estimated M AR models, and an initial estimate of (π, a) is obtained from the appropriate relative frequencies at which each initial state and state transition are chosen. The relations between model estimates obtained using the Baum algorithm and the generalized Lloyd algorithm were rigorously studied in [122] and [123]. It was shown that with high probability the estimate of the parameter set of the HMM obtained from VQ of the training data is a fixed point of the Baum algorithm provided that K is sufficiently large. Since this condition is satisfied in most speech applications, HMM's can be estimated using the generalized Lloyd algorithm which is significantly simpler than the Baum algorithm.

VI. Signal Estimation

In this section MMSE and MAP estimation of the clean signal using HMM's for the signal and noise is considered.

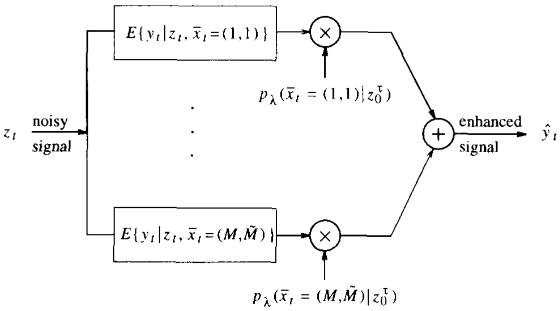

$E\{y_t|z_t, \bar{x}_t=(1,1)\}$
$p_\lambda(\bar{x}_t = (1,1)|z_0^\tau)$
z_t
noisy
signal
enhanced
signal
$\hat{y}_t$
$E\{y_t|z_t, \bar{x}_t=(M,\tilde{M})\}$
$p_\lambda(\bar{x}_t = (M,\tilde{M})|z_0^\tau)$

Fig. 6. HMM-based MMSE estimation. $E\{y_t|z_t, \bar{x}_t\}$ denotes the MMSE estimator of the clean signal y_t given the noisy signal z_t and its composite state $\bar{x}_t$ at time t, and $p_\lambda(\bar{x}_t|z_0^\tau)$ denotes the posterior probability of the composite state of the noisy signal at time t given the noisy observations z_0^τ.

A. MMSE Estimation

The MMSE estimator of y_t given z_0^τ, $\tau \geq t$, was derived in [29]. This estimator is obtained from (8), (31), and (32), and it is given by

$$\begin{aligned}\hat{y}_t &\triangleq E\{y_t|z_0^\tau\}\\ &= \int y_t p_\lambda(y_t|z_0^\tau)dy_t\\ &= \sum_{\bar{x}_t} p_\lambda(\bar{x}_t|z_0^\tau)E\{y_t|z_t, \bar{x}_t\}. \qquad (48)\end{aligned}$$

A block diagram of the MMSE estimator is shown in Fig. 6. This estimator comprises a weighted sum of conditional mean estimators for the composite states of the signal and noise, where the weights are the probabilities of these states given the noisy signal. The probability $p_\lambda(\bar{x}_t|z_0^\tau)$ can be efficiently calculated using (33). The conditional mean $E\{y_t|z_t, \bar{x}_t\}$ is given by (32) for HMM's with Gaussian subsources. In this case, the conditional mean estimator is a linear function of z_t, and the estimate of y_t given $\bar{x}_t$ is obtained from z_t using a Wiener estimator (or filter). Note that since $p_\lambda(\bar{x}_t|z_0^\tau)$ depends on the noisy data z_0^τ in a nonlinear manner (see (33)–(35)), the MMSE signal estimator $\hat{y}_t$ is a nonlinear function of z_0^τ. Note also that the causal MMSE estimator ($\tau = t$) differs from the noncausal MMSE estimator ($\tau > t$) only in the noisy observations from which the composite state probabilities are calculated.

The MMSE signal estimator (48) is unbiased in the sense that the expected value of the estimated signal equals the expected value of the original signal; i.e.,

$$E\{\hat{y}_t\} = E\{E\{y_t|z_0^\tau\}\} = E\{y_t\}. \qquad (49)$$

The MSE associated with this estimator was evaluated in [41] for HMM's with strictly positive state transition probabilities, and fairly general state-dependent PD's which only satisfy very mild regularity conditions. When HMM's with Gaussian AR subsources were considered, it was assumed that the power spectral densities of the subsources, i.e., $\{\boldsymbol{S}_{x_t}(\theta) = \sigma_{x_t}^2/|A_{x_t}(\theta)|^2\}$ and $\{\boldsymbol{S}_{\tilde{x}_t}(\theta) = \sigma_{\tilde{x}_t}^2/|A_{\tilde{x}_t}(\theta)|^2\}$, are bounded from below and above. For this case, it was shown that as $K \to \infty$, the MMSE of $\hat{y}_t$ exponentially approaches a weighted sum of the asymptotic MMSE of the individual Wiener estimators, where the weights are the *a priori* probabilities of the composite states. The asymptotic MMSE is given by

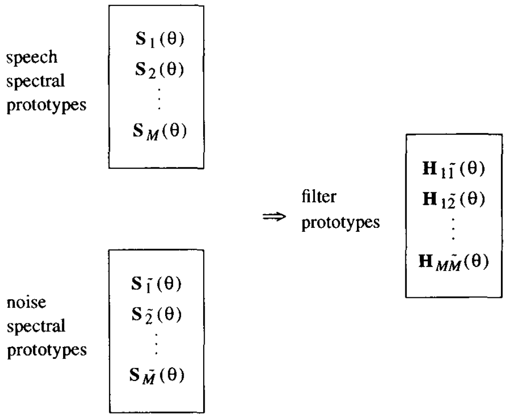


Fig. 7. Model-based estimation. $\{S_\beta(\theta)\}_{\beta=1}^M$ and $\{S_{\tilde{\beta}}(\theta)\}_{\tilde{\beta}=1}^{\tilde{M}}$ are prototype power spectral densities of speech and noise, respectively. $\boldsymbol{H}_{m\tilde{m}}$ is the Wiener filter for the prototype pair $(m, \tilde{m})$.

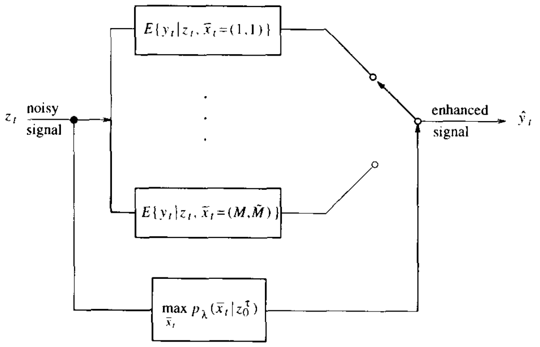
$$\begin{aligned}\bar{\epsilon}_t^2 &\triangleq \lim_{K\to\infty} \frac{1}{K}\mathrm{tr}E\{(y_t - \hat{y}_t)(y_t - \hat{y}_t)^{\#}\}\\ &= \sum_{\bar{x}_t} p_\lambda(\bar{x}_t)\int_{-\pi}^{\pi} \boldsymbol{H}_{\bar{x}_t}(\theta)\boldsymbol{S}_{\tilde{x}_t}(\theta)\frac{d\theta}{2\pi}, \qquad (50)\end{aligned}$$

where $p_\lambda(\bar{x}_t)$ is the *a priori* probability of the composite state $\bar{x}_t$ of the noisy signal, and

$$\boldsymbol{H}_{\bar{x}_t}(\theta) \triangleq \frac{\boldsymbol{S}_{x_t}(\theta)}{\boldsymbol{S}_{x_t}(\theta) + \boldsymbol{S}_{\tilde{x}_t}(\theta)} \qquad (51)$$

is the frequency response of the Wiener filter associated with the composite state $\bar{x}_t$. The *a priori* probability $p_\lambda(\bar{x}_t)$ can be efficiently calculated using (33)–(35) with $b(z_t|\bar{x}_t) \equiv 1$ for all t.

The HMM-based MMSE signal estimator has a very intuitive basis, and it can be interpreted as follows. Suppose that the spectral prototypes of the speech signal and of the noise process were estimated by applying the VQ approach to the training data from the two sources. Using the usual notation, let the resulting spectral prototypes of the signal and noise be $\{\boldsymbol{S}_m(\theta)\}_{m=1}^M$ and $\{\boldsymbol{S}_{\tilde{m}}(\theta)\}_{\tilde{m}=1}^{\tilde{M}}$, respectively. For each pair of spectral prototypes, say $(\boldsymbol{S}_m(\theta), \boldsymbol{S}_{\tilde{m}}(\theta))$, a Wiener filter $\boldsymbol{H}_{m\tilde{m}}$ (similar to (51)) can be designed. This results in a set of $M \times \tilde{M}$ Wiener filters, as shown in Fig. 7, which are optimal for all possible combinations of speech and noise vectors. Now, if the states of the speech and noise for a given vector of noisy signal z_t were known, then the most appropriate filter from the predesigned set of filters could be applied to z_t, and optimal (in the MMSE sense) estimation of y_t could be performed. Since the states of the signal and noise are hidden, however, the most appropriate filter for the given noisy vector is not known. Hence, all filters are tried, and each estimate is assigned a probability of being the best signal estimate.



$E\{y_t|z_t,\bar{x}_t=(1,1)\}$
z_t
noisy
signal
$E\{y_t|z_t,\bar{x}_t=(M,\tilde{M})\}$
$\max_{\bar{x}_t} p_\lambda(\bar{x}_t|z_0^\tau)$
enhanced
signal
$\hat{y}_t$

Fig. 8. Asymptotic MMSE signal estimator. $E\{y_t|z_t,\bar{x}_t\}$ is the MMSE estimator of the clean signal y_t given the noisy signal z_t and its composite state $\bar{x}_t$ at time t, and $p_\lambda(\bar{x}_t|z_0^\tau)$ is the posterior probability of the composite state of the noisy signal at time t given the noisy observations z_0^τ.

The MMSE signal estimate is constructed as the average of the individual estimates weighted by their probabilities.

Other strategies for utilizing the predesigned set of Wiener filters are possible. For example, the filter which is most likely to be the correct filter can be chosen and applied to the noisy signal. The estimation scheme which results from this approach is described in Fig. 8. This estimation approach was shown in [41] to be asymptotically optimal (as $K \to \infty$) in the MMSE sense; i.e., it achieves the asymptotic MMSE (50) at the same exponential rate as the MMSE estimator. The MMSE estimator and the asymptotically MMSE estimator are soft and hard decision estimation approaches, respectively. The MMSE estimator was proven more useful than the asymptotic MMSE estimator in speech enhancement applications, probably because the frequency response of the filter in the soft decision approach is less likely to change drastically from one vector to another compared with the frequency response of the filter in the hard decision approach.

A third strategy for using the set of predesigned filters results from the MAP signal estimation approach. In this case, both the noisy signal and an estimate of the clean signal are used to estimate the posterior probabilities of the states, and harmonic mean of the Wiener filters is applied to the noisy signal. As is shown in part B of this section, this approach can also be considered a hard decision estimation approach, since only one filter is effectively applied to the noisy signal at each given time instant.

It is interesting to note that similar MMSE estimation schemes have been commonly used in control applications where the unknown parameter of the plant is quantized into M values which constitute the states of the systems (see, e.g., [149]–[155]). A filter (Wiener or Kalman) is designed for each state based upon the specific value of the parameter represented by that state, and the probabilities of the states are estimated from the given noisy observations. Since only one state corresponds to the true value of the parameter of the plant, the posterior probability of that state approaches 1 as the number of observations increases, and the filter associated with that state becomes the only active filter of the system. For this reason the MMSE estimator is referred to as an adaptive estimator in control theory. The situation in speech enhancement is different since each vector of the signal may be generated from a different state, and none of the states should be constantly active.

B. MAP Estimation

The MAP estimator of y_t given z_0^τ, $\tau \geq t$, was developed in [29]. The estimator is obtained from local maximization of $p_\lambda(y_t|z_0^\tau)$ in (31) over y_t using the EM algorithm.

The EM algorithm is an iterative procedure for MAP estimation from incomplete data. The "complete data" in this case comprise $\{\bar{x}_t, z_t\}_{t=0}^\tau$ while the "incomplete data" constitute the given noisy signal z_0^τ. If the complete data were available, say the composite state of the noisy signal at time t was known to be $(\bar{x}_t)^*$, then $p_\lambda(y_t|z_0^\tau) = b(y_t|z_t,(\bar{x}_t)^*)$ and the MAP estimate of y_t would simply be the conditional mean given in (32). Since the complete data is not available, direct maximization of $p_\lambda(y_t|z_0^\tau)$ is not trivial. The EM approach attempts to perform this maximization iteratively, by maximizing in each iteration the expected value of $\ln b(y_t|z_t,\bar{x}_t)$ given a current estimate of y_t and the noisy observations. It can be shown that the sequence of signal estimates obtained in this way has nondecreasing likelihood values [147], [148]. Furthermore, conditions for convergence of this sequence were given in [148]. The conditional expected value of $\ln b(y_t|z_t,\bar{x}_t)$ is referred to as an auxiliary function in the EM terminology, and normally it is significantly simpler to maximize than the original likelihood $p_\lambda(y_t|z_0^\tau)$. The evaluation and maximization of the auxiliary function correspond to the expectation and maximization steps of the EM algorithm, respectively.

The auxiliary function for MAP estimation of y_t from z_0^τ, assuming a current estimate y_t' of y_t, is given by

$$\begin{aligned} Q(y_t,y_t') &\triangleq E\{\ln b(y_t|z_t,\bar{x}_t)|y_t',z_0^\tau\} \\ &= \sum_{\bar{x}_t} p_\lambda(\bar{x}_t|y_t',z_0^\tau)\ln b(y_t|z_t,\bar{x}_t), \end{aligned} \tag{52}$$

where $p_\lambda(\bar{x}_t|y_t',z_0^\tau)$ denotes the posterior probability of the composite state of the noisy signal at time t given the current estimate of y_t and the noisy observations. The probability $p_\lambda(\bar{x}_t|y_t',z_0^\tau)$ is calculated using the relation

$$p_\lambda(\bar{x}_t,y_t|z_0^\tau) = p_\lambda(\bar{x}_t|z_0^\tau)b(y_t|z_t,\bar{x}_t), \tag{53}$$

where this expression results from (31), and $p_\lambda(\bar{x}_t|z_0^\tau)$ can be efficiently calculated using the "forward–backward" formulas (33)–(35). Maximization of (52) results in the following signal reestimation formula:

$$\begin{aligned} y_t(n+1) &\triangleq \arg\max_{y_t} Q(y_t,y_t')\Big|_{y_t'=y_t(n)} \\ &= \left[\sum_{\bar{x}_t} p_\lambda(\bar{x}_t|y_t(n),z_0^\tau)H_{\bar{x}_t}^{-1}\right]^{-1} z_t, \end{aligned} \tag{54}$$

where $y_t(n)$ denotes the estimate of y_t as obtained in the nth iteration. At each iteration, the MAP estimate of y_t is obtained by applying a filter which comprises the harmonic mean of Wiener filters to z_t. If the sum in (54) is dominated by a unique composite state, i.e., $p_\lambda(\bar{x}_t|y_t(n), z_0^\tau) \approx 1$ for some $\bar{x}_t$, then y_t is obtained from z_t using the "most likely" Wiener filter among all $M \times \tilde{M}$ possible filters. This situation has often been encountered in practice when enhancing speech signals degraded by white noise. The reason is that the speech signal has distinct clusters and hence the state corresponding to the cluster to which the current estimate of y_t belongs dominates the sum in (54). This situation does not happen so often in MMSE signal estimation, since there the weights are the probabilities of the composite states conditioned only on the noisy signal, i.e., $p_\lambda(\bar{x}_t|z_0^\tau)$, and the clusters of this signal are not distinct due to noise masking. Thus, in practice, MAP estimation is based upon a "hard decision" in choosing the filter for each vector of noisy speech, while MMSE estimation makes a "soft-decision" by taking the weighted sum of all possible filters.

The convergence of the MAP signal estimator was proven in [29]. Specifically, it was shown that every limit point of $y_t(n)$ is a stationary point of $p_\lambda(y_t|z_0^\tau)$, and $p_\lambda(y_t(n)|z_0^\tau)$ converges monotonically to $p_\lambda(y_t^*|z_0^\tau)$, where y_t^* is some stationary point of $p_\lambda(y_t|z_0^\tau)$. The recursion in (54) can be started using any estimate of the clean signal, for example, the noisy signal, i.e., $y_t(0) = z_t$. Since only local convergence is guaranteed, however, the performance of the algorithm will be dependent on the initial estimate of the signal.

A noncausal MAP estimator of the entire signal $y = y_0^t$ from the entire noisy signal $z = z_0^t$ was developed in [37] using the EM approach. An interesting estimating scheme results if the likelihood function $p_\lambda(y|z)$ is assumed to be dominated by a unique sequence of states. In this case, y is estimated by maximizing $p_\lambda(x, y|z)$ over $\{x, y\}$, where $x = x_0^t$. Alternate maximization of $p_\lambda(x, y|z)$ is performed, once over x assuming y is given and then over y assuming x is available. This results in a sequence of signal estimates with nondecreasing likelihood values. The estimation of the most likely sequence of states is performed by applying the Viterbi algorithm [175] to the current estimate of the clean signal. The estimation of the signal is done by applying the sequence of Wiener filters associated with the most likely sequence of states to the noisy signal. A block diagram of this approach is shown in Fig. 9.

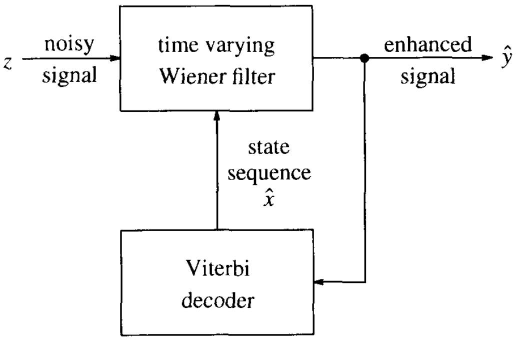


Fig. 9. Wiener filtering-Viterbi decoding enhancement.

C. Gain Adaptation

The HMM-based speech enhancement approach relies solely on statistical properties of the speech signal as inferred from the training data. However, certain properties of the test speech signal, such as the energy contour, cannot be reliably predicted from the training data, and hence must be estimated from the given sample function of the noisy signal. There are two reasons for this difficulty. First, recording conditions during training and testing may be different. This situation occurs, for example, in enhancing pilot's speech recorded through fading channels using a speech enhancement system designed from training data recorded under laboratory conditions. In this example, the recording condition of the test data are not only different from those of the training data but also are time varying. Second, speech signals are not strictly stationary and hence have time-varying energy. A mismatch between the energy contours of the clean signal and the HMM for that signal results in poor enhancement of the noisy signal, since probabilities of speech events cannot be reliably calculated by the model. This is a consequence of the fact that in this case the covariance matrices of the model misrepresent the second-order statistics of the clean test signal.

Two strategies for matching the energy of the model for the clean signal to the energy of the clean signal observed through the noisy signal were studied in [29]. In the first strategy, which is geared to MMSE signal estimation, a global gain factor is estimated by matching the energy of the given noisy signal to the energy of that signal as predicted from its model. Specifically, the global gain factor G_T is determined from

$$\sum_{t=0}^{T} z_t^{\#} z_t = G_T^2 \sum_{t=0}^{T} E\{y_t^{\#} y_t\} + \sum_{t=0}^{T} E\{v_t^{\#} v_t\}, \tag{55}$$

where from (21)–(24),

$$\begin{aligned} E\{y_t^{\#} y_t\} &= \sum_{x_t} p_{\lambda_s}(x_t) \mathrm{tr}\{S_{x_t}\} \\ E\{v_t^{\#} v_t\} &= \sum_{\tilde{x}_t} p_{\lambda_v}(\tilde{x}_t) \mathrm{tr}\{S_{\tilde{x}_t}\}. \end{aligned} \tag{56}$$

A sequential version of this gain estimation approach would be to use a gain factor G_τ estimated from z_0^τ using (55), where τ takes the same value as in (48).

The second strategy is tailored to the MAP signal estimation approach. Here, an individual gain factor is attributed to each vector of the clean signal from the training and test data, and the gain factors, or the gain contours of the training and test signals, are estimated from the given signals during both training and enhancement. During training, an ML estimate of the gain contour of the training data is used for estimating an HMM for gain-normalized clean signals. During enhancement, the HMM for gain-normalized signals is supplemented with an ML estimate of the gain contour of the clean signal obtained from the given noisy signal. The resulting procedures are referred to as gain-adapted training and enhancement, respectively. Gain-adapted training is performed by

$$\max_{\lambda_s} \max_{g_0^T} \frac{p_{\lambda_s}(y_0/g_0, \cdots, y_T/g_T)}{g_0^K \cdots g_T^K}, \tag{57}$$

where g_t denotes the gain factor for the vector y_t, $g_0^T \triangleq \{g_0, \cdots, g_T\}$ is the gain contour of y_0^T, and λ_s is now the parameter set of the HMM for the *gain-normalized* clean signal. Given λ_s, gain-adapted MAP signal estimation is performed by

$$\max_{y_t} \max_{g_t} p_\lambda(y_t, z_0^t | g_0^t), \tag{58}$$

where $p_\lambda(y_t, z_0^t|g_0^t)$ is the joint pdf of y_t and z_0^t given the HMM λ_s for gain-normalized clean signals and the gain contour g_0^t. An expression for this pdf can be obtained by multiplying (31) by $p_\lambda(z_0^\tau)$, using $p_\lambda(\bar{x}_t, z_0^t) = F(\bar{x}_t, z_0^t)$ as given by (34), and by replacing S_{x_t} by $g_t^2 S_{x_t}$ for all $\{x_t\}$. The optimization in (58) is performed only on g_t, since g_0^{t-1} is assumed available from estimation performed at time instants $t' < t$. Furthermore, maximization of the joint pdf $p_\lambda(y_t, z_0^t|g_0^t)$ is considered, since maximization of the conditional pdf $p_\lambda(y_t|z_0^t, g_0^t)$ results in a highly nonlinear gain estimation problem [29].

Gain-adapted training is performed by alternate maximization of the likelihood function in (57) over λ_s and g_0^T. This results in an alternative model-gain estimation procedure as shown in Fig. 10. Similarly, gain-adapted enhancement can be performed by alternate estimation of the signal and its gain at time t as shown in Fig. 11. Convergence in each case is determined by examining the difference of the values of the likelihood function in two consecutive iterations. The algorithms are stopped if this difference becomes smaller than a given threshold. The appropriate reestimation formulas for both algorithms are summarized below.

For a given model λ_s and gain contour g_0^T at the nth iteration of gain-adapted training, a new estimate of the gain at the $(n+1)$th iteration is obtained from

$$g_t^2(n+1) = \sum_{x_t} p_{\lambda_s}(x_t|y_0^T, g_0^T)\sigma_{x_t}^2, \tag{59}$$

where

$$\sigma_{x_t}^2 \triangleq \frac{1}{K} y_t^\# S_{x_t}^{-1} y_t \tag{60}$$

is the variance of the residual signal obtained from AR modeling of y_t using S_{x_t}, and $p_{\lambda_s}(x_t|y_0^T, g_0^T)$ is defined similarly to (43) with S_{x_t} being replaced by $g_t^2 S_{x_t}$. Thus for a given model and gain contour estimate, the new gain estimate for y_t constitutes the square root of the average power of the residual signals obtained from AR modeling of y_t using the models corresponding to the different subsources. Using the resulting gain estimate and the given model, a new model can be estimated by using reestimation formulas similar to (38)-(40) with posterior state probabilities given by $p_{\lambda_s}(x_{t-1}, x_t|y_0^T, g_0^T(n+1))$.

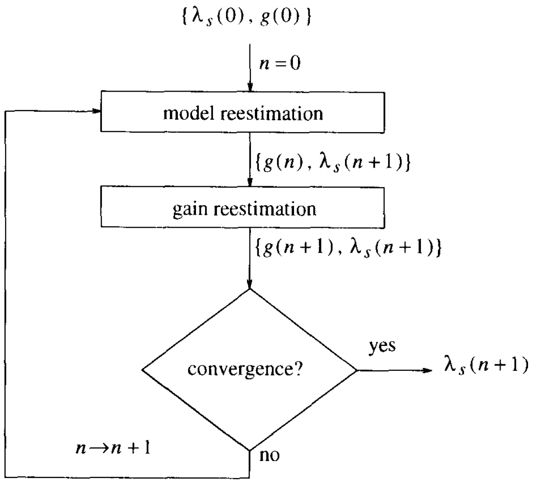


Fig. 10. Gain-adapted training. $\lambda_s(n)$ and $g(n) \triangleq \{g_t(n)\}_{t=0}^T$ denote, respectively, estimates of the HMM parameter set and the gain contour of the training signal at the nth iteration, where $\lambda_s(0)$ and $g(0)$ denote initially given estimates.

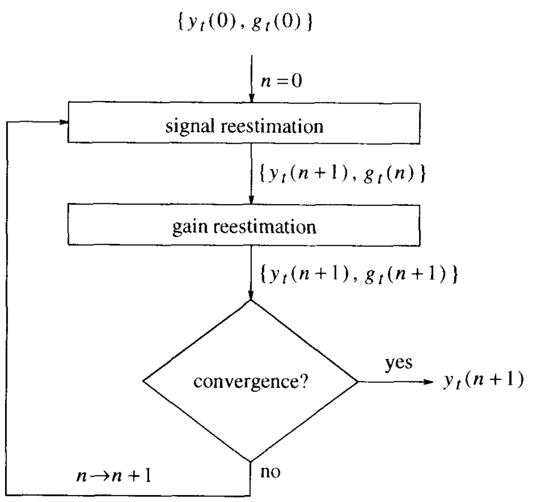


Fig. 11. Gain-adapted MAP signal estimation. $y_t(n)$ and $g_t(n)$ denote, respectively, estimates of the signal y_t and its gain at the nth iteration, where $y_t(0)$ and $g(0)$ denote initially given estimates.

For a given signal y_t and gain g_t at the nth iteration of gain-adapted MAP enhancement, a new gain estimate at the $(n+1)$th iteration is obtained from

$$g_t^2(n+1) = \sum_{\bar{x}_t} p_\lambda(\bar{x}_t|y_t, g_t, z_0^t)\sigma_{x_t}^2, \tag{61}$$

where $\sigma_{x_t}^2$ is defined in (60), and $p_\lambda(\bar{x}_t|y_t, g_t, z_0^t)$ is defined similarly to $p_\lambda(\bar{x}_t|y_t', z_0^\tau)$ in (52) with S_{x_t} being replaced by $g_t^2 S_{x_t}$ for every t using the assumed known gain contour g_0^{t-1} and the current estimate of the gain g_t. This gain reestimation procedure can be interpreted similarly to the gain reestimation procedure performed during training. Using the estimated gain g_t and the given signal estimate y_t, a

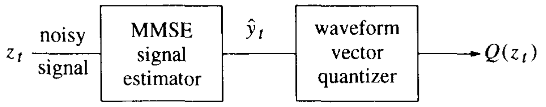


Fig. 12. MMSE waveform vector quantization.

new signal estimate can be obtained by using a reestimation formula similar to (54), with posterior state probabilities given by $p_\lambda(\bar{x}_t|y_t, g_t, z_0^t)$, and covariance matrices $\{S_{\bar{x}_t}\}$ replaced by $\{g_t^2 S_{\bar{x}_t}\}$. The initialization of the gain-adapted training and signal estimation algorithms, and efficient implementations of these algorithms in the frequency domain, were discussed in [29] and [82], respectively.

VII. Source Coding

The problem of encoding noisy signals using the modified distortion measure was formulated in subsection IV-B. In [62], sufficient conditions on the distortion measure and the PD's of the signal and noise were given which guarantee convergence of the generalized Lloyd algorithm [140] for designing vector quantizers. It can be verified that these conditions are satisfied by the squared error distortion measure and the HMM's with Gaussian subsources considered here.

For memoryless MMSE waveform VQ, this approach immediately shows that the optimal vector quantizer is a two-step encoder in which MMSE estimation of the signal is first performed and then MMSE VQ is applied to the estimated signal [64], [62]. A block diagram of this encoder is shown in Fig. 12. The MMSE estimator is a point estimator of y_t given z_t which can be derived similarly to (48) or simply by using the Bayes theorem,

$$\begin{aligned}\hat{y}_t &= E\{y_t|z_t\}\\ &= \sum_{\bar{x}_t} p_\lambda(\bar{x}_t|z_t)E\{y_t|z_t,\bar{x}_t\}, \end{aligned} \tag{62}$$

where $E\{y_t|z_t, \bar{x}_t\}$ is given by (32), and $p_\lambda(\bar{x}_t|z_t)$ can be calculated similarly to $p_\lambda(\bar{x}_t|z_0^t)$ in (33) using the "forward" formula. The MSE associated with this encoder equals the sum of the mean squared estimation and quantization errors,

$$E\{d(y_t, Q(z_t))\} = E\{d(y_t, \hat{y}_t)\} + E\{d(\hat{y}_t, Q(z_t))\}, \tag{63}$$

where the expression for the asymptotic (as $K \to \infty$) mean squared estimation error is given in (50), and the asymptotic mean squared quantization error can be calculated from the rate distortion theory [55].

The optimal AR model vector quantizer in the Itakura–Saito sense was derived in [62]. Similarly to the MMSE waveform vector quantizer, the AR model vector quantizer is a two-step encoder in which MMSE estimation of the sample spectrum of the clean signal is first performed, and then AR model VQ is applied to the estimated spectrum using the Itakura–Saito distortion measure [138]. A block diagram of this encoder is shown in Fig. 13. The average Itakura–Saito distortion associated with this encoder equals the sum of the average Itakura–Saito distortions resulting from the sample spectrum estimation and quantization:

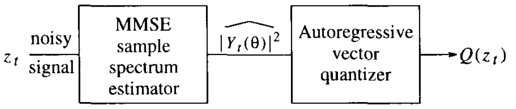


Fig. 13. Average Itakura-Saito autoregressive model vector quantization.

$$\begin{aligned}E\{d(|Y_t(\theta)|^2, Q(z_t))\} =&E\{d(|Y_t(\theta)|^2, |\widehat{Y_t(\theta)}|^2)\}\\ &+ E\{d(|\widehat{Y_t(\theta)}|^2, Q(z_t))\}. \end{aligned} \tag{64}$$

The MMSE estimator of the signal discrete sample spectrum $|Y_t(\theta)|^2$, $\theta = 2\pi k/K, k = 0, \cdots, K-1$, is given by

$$\begin{aligned}|\widehat{Y_t(k)}|^2 &= E\{|Y_t(k)|^2|z_t\}\\ &= \sum_{\bar{x}_t} p_\lambda(\bar{x}_t|z_t)E\{|Y_t(k)|^2|z_t,\bar{x}_t\}, \end{aligned} \tag{65}$$

where $Y_t(k)$ is the discrete Fourier transform (DFT) of y_t normalized by $K^{1/2}$. The conditional mean $E\{|Y_t(k)|^2|z_t, \bar{x}_t\}$ was calculated in [29] under the assumption that the covariance matrices of the subsources of the HMM's for the signal and noise are circulant [165]. In this case we have that

$$E(|Y_t(k)|^2|z_t, \bar{x}_t) = \boldsymbol{H}_{\bar{x}_t}(k)\boldsymbol{S}_{\bar{x}_t}(k) + \boldsymbol{H}^2_{\bar{x}_t}(k)|Z_t(k)|^2, \tag{66}$$

where $Z_t(k)$ is the DFT of z_t normalized by $K^{1/2}$, and $\boldsymbol{H}_{\bar{x}_t}(k)$ and $\boldsymbol{S}_{\bar{x}_t}(k)$ are defined similarly to $\boldsymbol{H}_{\bar{x}_t}(\theta)$ and $\boldsymbol{S}_{\bar{x}_t}(\theta)$ in (51), respectively, for $\theta = 2\pi k/K$. Note that under this assumption, the signal estimators (48) and (62) can also be implemented in the frequency domain [29]. In this case, the DFT of $E\{y_t|z_t, \bar{x}_t\}$ normalized by $K^{1/2}$ is obtained from $\{\boldsymbol{H}_{\bar{x}_t}(k)Z_t(k), k = 0, \cdots, K-1\}$. Similarly, the conditional probabilities $p_\lambda(\bar{x}_t|z_0^T)$ in (48), and $p_\lambda(\bar{x}_t|z_t)$ in (62) and (65), can be evaluated in the frequency domain using inverses and determinants of circulant matrices.

A variety of different approaches for AR model quantization given noisy speech signals have been reported in the literature over the years. Since most of these approaches were proposed before VQ dominated the area of low-bit-rate speech coding, they attempt to first estimate the AR model for each vector of the clean signal and then apply scalar quantization to the parameters of the estimated model. The estimation of the AR model from the noisy signal is usually a difficult task, and it is not needed if the ultimate goal is to only estimate a finite number of AR models for the clean signal. Furthermore, the optimality of these approaches in any well-defined sense is not known. For completeness of the discussion, we briefly review these approaches.

In [3], and [66]–[68], different functions of the clean signal were estimated from the noisy signal, and they were later used for estimating the AR model for each vector of the clean signal. In [3], the sample spectrum of the clean signal was estimated using the spectral subtraction estimator

(2). In [66]–[68], the waveform, the autocorrelation, and the spectral magnitude of the clean signal were estimated, respectively. Since the sample spectrum estimation is required in the optimal scheme, the approach of [3] is probably the closest one to optimal.

In [70]–[72], the AR model for each vector of the clean signal is directly estimated from the given noisy signal using iterative estimation procedures. In [70], successive autocorrelation is applied to the noisy data for improving the SNR in the estimated model. In [71], the poles of the AR model in each iteration are enhanced using the estimate of the AR model from the previous iteration. In [72], an AR model estimate is obtained assuming some value for the noise variance, and this estimate is iteratively improved by improving the estimation of the noise variance. An algorithm for ML estimation of the AR model from noisy signals using the EM approach was developed in [69]. This approach seems most reasonable due to the asymptotic optimal properties of the ML estimation approach.

In [73]–[76], approaches for AR model estimation which are suitable when the noise is white were developed. These approaches rely on the fact that when an AR process is contaminated by white noise, the resulting noisy process becomes an autoregressive moving average (ARMA) process [77]. Hence, an ARMA model is estimated from the noisy signal, and the AR part of that model is attributed to the clean signal. In [74], an ARMA model for the noisy signal which approximates the sample correlation of that signal in the MSE sense was estimated. In [75], an ARMA model which approximates the sample spectrum of the noisy signal in the Itakura–Saito sense was estimated. In [76], the ARMA model for the noisy signal was estimated using an ML approach which was implemented using the Newton–Raphson method.

Another AR model estimation approach, which is different from those discussed above, was proposed in [78]. In this approach, the reflection coefficients of the AR model are compensated for the noise presence, and this estimate is embedded into the Levinson–Durbin algorithm [157] for recursive estimation of the AR model.

The variety of approaches for AR model estimation demonstrates the attention that this problem has received, and its importance in speech coding and AR modeling for, say, speech recognition applications (see Section VIII). The approach of Fig. 13 provides the exact statistic of the clean signal which must be estimated if the Itakura–Saito distortion measure is used, and in this approach AR estimation and quantization are simultaneously performed using VQ [130]. In Section IX we discuss the performance of the encoder in Fig. 13 and compare it with the most popular preprocessing approach, which uses the spectral subtraction sample spectrum estimator.

VIII. Signal Classification

The problem of classifying noisy speech signals was studied in [80]–[83] and [85]–[107]. Different approaches to this problem have been proposed. In [85]–[95], the clean signal, or some feature vectors of that signal, is first estimated from the noisy signal, and then recognition of the estimated signals is performed. Commonly used feature vectors include the AR model parameters and the derived cepstral coefficients [157] which correspond to each vector of the speech signal. In [96]–[99], the templates of the clean signal, normally used for recognition of those signals, are adapted to take into account the input noise. In [100]–[106], signal representations and distortion measures which are robust to noise are used in dynamic time warping [167] and HMM-based speech recognition systems. The approaches in [85]–[106] are fairly intuitive and are relatively easy to implement. It is difficult, however, to establish their optimality in any well-defined sense. In [107], a robust statistics estimation approach is applied for designing an asymptotically optimal recognition system in the minimum probability of error sense.

A different class of speech recognition approaches, one which attempts to minimize the probability of classification error, was proposed in [80]–[83]. In [83], an asymptotically optimal decision rule for classification of noisy sources with parametric models was proposed. In this approach, the parameters of the models are assumed unknown, and they are estimated during the recognition process from both the noisy test signal and the training data. This decision rule is given by

$$\max_{1\leq i\leq J} \frac{\max_{\lambda_s,\lambda_v} p_{\lambda_s,\lambda_v}(z,Y_i,V|W_i)}{\max_{\lambda_s,\lambda_v} p_{\lambda_s,\lambda_v}(Y_i,V|W_i)}, \qquad (67)$$

where $p_{\lambda_s,\lambda_v}(z,Y_i,V|W_i)$ is the joint pdf of the noisy test signal z, the training data Y_i from the ith word, and the training data V from the noise process, and all words are assumed equiprobable; i.e., $P(W_i) = 1/J$ for all $i = 1,\cdots,J$. The decision rule (67) can be intuitively interpreted by examining the value of $p_{\lambda_s,\lambda_v}(z,Y_i,V|W_i)$ for z emerging either from the ith word or from any other word in the vocabulary. If z emerges form the ith word then z and Y_i+V have similar statistics, while if z emerges from another word then z and $Y_i + V$ have different statistics. Hence, $\max_{\lambda_s,\lambda_v} p_{\lambda_s,\lambda_v}(z,Y_i,V|W_i)$ obtained in the first case should be larger than that obtained in the second case, and the decision rule (67) should correctly classify z as coming from the ith word. Note that the denominator of (67) is independent of z and it is used for obtaining the conditional pdf of z given Y_i, V, and W_i.

If the training sequences Y_i and V are significantly longer than the test signal z, then estimating $\lambda = (\lambda_s, \lambda_v)$ from either (z, Y_i, V) or (Y_i, V) should provide similar models. Since Y_i and V are statistically independent, then λ_s and λ_v can be estimated off line from Y_i and V, respectively, using the ML approach (see subsection V-D). Let these estimates be denoted by

$$\hat{\lambda}_{s_i} \triangleq \arg\max_{\lambda_s} p_{\lambda_s}(Y_i|W_i)$$

$$\hat{\lambda}_v \triangleq \arg\max_{\lambda_v} p_{\lambda_v}(V).$$

Combining this observation with the statistical independence of z, Y_i, and V when λ_s and λ_v are given, we find

$$\max_{\lambda_s,\lambda_v} p_{\lambda_s,\lambda_v}(z, Y_i, V|W_i) \approx p_{\hat{\lambda}_{s_i},\hat{\lambda}_v}(z, Y_i, V|W_i)$$
$$= p_{\hat{\lambda}_{s_i},\hat{\lambda}_v}(z|W_i) p_{\hat{\lambda}_{s_i},\hat{\lambda}_v}(Y_i, V|W_i). \quad (68)$$

Substituting (68) into (67), we find that the decision rule for this case becomes

$$\max_{1 \leq i \leq J} p_{\hat{\lambda}_{s_i},\hat{\lambda}_v}(z|W_i), \quad (69)$$

which is identical to the decision rule (20) for equiprobable words; i.e.,

$$\max_{1 \leq i \leq J} p_{\lambda}(z|W_i). \quad (70)$$

The difference between the decision rules (67) and (70) is that (70) is optimal when the parameter sets of the models for the clean signal and the noise process are assumed available from the off-line training procedure, while (67) is asymptotically optimal under the more realistic assumption that these parameters are not known. The implementation of (67) is substantially more complicated than the implementation of (70), since the former test requires an on-line maximization of the pdf $p_{\lambda_s,\lambda_v}(z, Y_i, V|W_i)$. Since (70) is more consistent with the main theme of this paper, we proceed with our discussion on signal classification using this decision rule only.

Recognition approaches similar to that studied here were proposed in [80] and [81] and tested in recognition of real speech signals. In [81], however, estimation of λ from noisy data, rather than estimation of λ_s and λ_v from clean and noisy data, respectively, is performed. This appears to be a disadvantage, since the models for the noisy signals must be reestimated for each new noisy environment, whereas in the approach studied here, the models for the clean signal are estimated once only while the model for the noise is reestimated. If the noise model is simple, as in the case of wideband noise, then it is significantly easier to reestimate the noise model than to reestimate the entire set of models for the noisy signals. Note that, in the recognition approach (70), no *a priori* estimation of the clean signal or of any other function of the clean signal is needed, and the problem is essentially that of classifying clean signals.

For *continuous* time signals, it was shown in [84] that minimum probability of error classification can also be obtained by applying the MAP decision rule to the *causal* MMSE estimator of the clean signal, provided that the variance of the clean signal is integrable, and the noise is Gaussian, white, additive, and statistically independent of the signal. A block diagram of this two-step classification approach is shown in Fig. 14. This interesting theoretical result provides the intuitive basis for the popular preprocessing approach for recognition of noisy speech signals [85]–[95]. This approach can also be applied using the MMSE estimator developed in [41]. In the statistical framework of hidden Markov modeling, however, the direct recognition approach of (70) is significantly simpler, since the noisy signal from a given word is an HMM (see subsection V-B), and the pdf $p_{\lambda}(z|W_i)$ can be efficiently evaluated using the "forward" recursion as shown in (36).

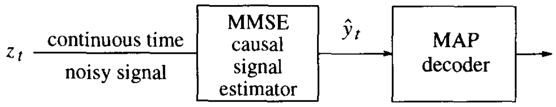


Fig. 14. Two-state minimum probability of error classification (Kailath [84]).

A gain-adapted MAP decision rule was developed in [82] for the particular case of noise sources that can be modeled by a single-state HMM with a Gaussian AR output process. The extension to noise sources with multiple states is straightforward. In the gain-adapted recognition approach, HMM's for gain-normalized clean signals are designed from the training data using (57). Recognition is performed by applying the MAP decision rule to the noisy signal, using the models for gain-normalized clean signals and ML estimates of the gain contour of the clean signal obtained from the noisy signal. Specifically, gain-adapted recognition of the utterance $z = z_0^T$ is performed by applying the MAP decision rule to the statistic

$$\max_{g} p_{\lambda}(z|W_i, g) = \max_{g} \int p_{\lambda_v}(z - y) p_{\lambda_s}(y|W_i, g) dy, \quad (71)$$

where $g = g_0^T$ is the gain contour of the clean signal $y = y_0^T$, $p_{\lambda}(z|W_i, g)$ is the pdf of the noisy signal from the ith word given the gain contour g and the HMM λ_s, and

$$p_{\lambda_s}(y|W_i, g) = \frac{p_{\lambda_s}(y_0/g_0, \cdots, y_T/g_T|W_i)}{g_0^K \cdots g_T^K} \quad (72)$$

is the pdf of the acoustic signal from the ith word given the gain contour g and the HMM λ_s for gain-normalized signals from that word.

Local maximization of the pdf in (71) was efficiently performed using the EM algorithm. The gain reestimation formula for this case is given by

$$g_t^2(n+1) = \sum_{\bar{x}_t} p_{\lambda}(\bar{x}_t|z, g(n)) \sigma_{\bar{x}_t}^2, \quad (73)$$

where

$$\sigma_{\bar{x}_t}^2 = \frac{1}{K} E\{y_t^{\#} S_{x_t}^{-1} y_t | z_t, \bar{x}_t, g_t(n)\}, \quad (74)$$

$g(n) = \{g_t(n), t = 0, \cdots, T\}$, and $p_{\lambda}(\bar{x}_t|z, g(n))$ is defined similarly to $p_{\lambda}(\bar{x}_t|z_0^{\tau})$ in (31) with S_{x_t} being replaced by $g_t^2(n) S_{x_t}$ for every t. A block diagram of the algorithm for calculating the gain-adapted pdf in (71) is shown in Fig. 15. Convergence of this algorithm is determined by examining the difference of the values of the likelihood function in two consecutive iterations. The algorithm is stopped if this difference becomes smaller than a given threshold. Note that (74) comprises the conditional mean of the gain estimate (60) used in gain-adapted signal estimation. This difference in the two gain reestimation procedures results from the fact that in (60) the clean signal is estimated at each iteration, while here no such estimation

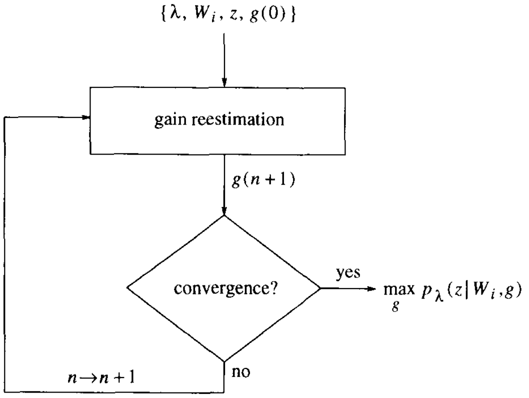


Fig. 15. Gain-adapted MAP recognition test. λ_s denotes the parameter set of an HMM for gain-normalized signals from the ith word in the vocabulary. λ_v denotes the parameter set of the HMM for the noise process. $g(n)\overset{\Delta}{=}\{g_t(n)\}_{t=0}^{T}$ denotes the estimate of the gain contour of the acoustic signal obtained at the nth iteration, where $g(0)$ denotes an initially given estimate.

is performed. For HMM's with Gaussian subsources (74) becomes

$$\begin{aligned}\sigma_{\bar{x}_t}^2 &= \frac{1}{K}\mathrm{tr}\{E[y_t y_t^{\#}|z_t, \bar{x}_t, g_t(n)]S_{x_t}^{-1}\}\\ &= \frac{1}{K}\mathrm{tr}\{[H_{\bar{x}_t}S_{\bar{x}_t} + (H_{\bar{x}_t}z_t)(H_{\bar{x}_t}z_t)^{\#}]S_{x_t}^{-1}\}, \quad (75)\end{aligned}$$

where $H_{\bar{x}_t}$ is defined similarly to (32) with S_{x_t} being replaced by $g_t^2(n)S_{x_t}$. The implementation of gain reestimation formula ((73) and (75)) in the frequency domain and the estimation of the initial gain estimate $g(0)$ were discussed in [82].

IX. Performance Evaluation

In this section we discuss the performance of the HMM-based speech enhancement approach in signal estimation, coding, and recognition, when only noisy signals are available. Specifically, we discuss the performance of the MMSE signal estimator (48) using the global gain-matching procedure (55); the causal MAP gain-adapted signal estimator (54), (61); the two-step AR model vector quantizer in Fig. 13 using the MMSE sample spectrum estimator (65) and the gain estimator (55); and the gain-adapted MAP decision rule (71) for recognition of noisy speech. We provide only a summary of the performance of these systems; the complete results can be found in [29], [37], [82], and [79].

These systems were exclusively tested using Gaussian white noise at input SNR greater than or equal to 5 dB. The SNR is defined as the ratio between the average power of the signal and the average power of the noise. White noise was used for the following reasons: First, white noise affects the entire frequency band of speech signals and is therefore considered one of the most perceptually harmful noise sources. Second, white noise is a good model for wideband noise sources which are often encountered in practice, e.g., thermal noise in communication systems. Third, white noise has been commonly used in studying the performance of speech enhancement systems and can therefore be seen as a "standard" test noise source. The Gaussian white noise is modeled by a single-state Gaussian AR HMM. The order of this model should be low as no resonants are expected with this type of noise. This order was chosen to be $N_v = 4$. The model for the noise was always estimated from an initial segment of the noisy signal (20 frames) which was known to contain noise only.

The HMM's used for signal estimation and coding were estimated using the VQ approach for reasons mentioned in subsection V-D. In this case, model estimation was performed by applying the generalized Lloyd algorithm to nonoverlapping vectors of the training data. In both cases, the models were estimated from a training data set which consisted of 5 min of conversational speech from six speakers, three males and three females. The PD's of the subsources of these models were assumed mixtures of Gaussian AR processes of order $N_s = 10$. For MMSE signal estimation, an eight-state with 64 mixture components per state HMM was used. For gain-adapted MAP signal estimation, an eight-state with two mixture components per state HMM was used. For signal coding at 1 bit/sample, an eight-state with 16 mixture components per state HMM was used. These values of the order of the HMM's were chosen since they provided best performance at 10 dB input SNR. A special state with a mixture of eight Gaussian AR processes was assigned to represent very low energy portions of the clean speech signals. The AR code words for this state were designed from vectors of the training data whose power was at least 30 dB below the average power of the entire training speech signal. Multiple mixture components were assigned to this state in order to ensure that perceptually important weak speech signals which were classified as silence will be well represented.

The speech enhancement systems were always tested on speech signals different from those used for training, and the speakers of the training and test speech material were not the same. For signal estimation, the test data consisted of nine sentences originally spoken by a male and a female. For signal coding, the test data consisted of 32 sentences spoken by four males and four females. In both cases, the test data were independently recorded from the training data, and thus gain mismatch exists between the two signals. The speech enhancement algorithms were applied to vectors of $K = 256$ speech samples obtained at an 8 kHz sampling rate. In signal estimation only, the vectors of the signal were premultiplied by a trapezoidal window with slope duration of eight samples, and reconstruction of the enhanced signal from the individually estimated vectors was performed using the overlap add synthesis approach [158]. The purpose of using such a window was to reduce block end-effects due to the fact that the signal is artificially treated as a fixed dimension vector source.

A. Signal Estimation

The MMSE waveform estimator provided a significant reduction in the level of the input noise when applied to noisy signals at 10 dB input SNR. Some of the words in some of the sentences, however, were accompanied by a low-level structured residual noise and usually these words sounded hoarse. This residual noise did not have the annoying nature of musical noise characteristic of signals estimated by the spectral subtraction estimators. In most of the examples studied here, consonants of the original speech were *not* suppressed by the filtering process. The reason for this perceptually important phenomenon is that the MMSE signal estimator is composed of filters which are matched to the different sounds of speech (including silence) as estimated from the training data. Thus consonants can be distinguished from noise, and can be enhanced using the appropriate matched filters. In general, the speech signals corresponding to the male speaker were better enhanced than the female speech signals. The difference was that more words in the female sentences were accompanied by the low-level structured noise mentioned above. A possible reason for this difference in performance is that the AR parameterization of the covariance matrices of the subsources of the HMM for the clean signal (see Section V) provides better modeling for male voice than for female voice. This fact has been observed in speech coding applications where AR models proved better for male speakers than for female speakers [157]. As the input SNR increases, the quality of the MMSE estimated signals was improved and the adverse phenomena mentioned above were significantly reduced.

The MAP signal estimator provided enhanced signals similar to those obtained using the MMSE signal estimator; however, these signals were accompanied by some additional muffled wideband noise at all input SNR. For this reason, the MAP signal estimator was judged inferior to the MMSE signal estimation. The intensity of this residual noise, however, decreased as the input SNR increased. At 10 dB input SNR, convergence of the MAP estimator was achieved using from four to ten iterations. Neither the MAP nor the MMSE signal estimator was found very effective at the low input SNR of 5 dB.

The MMSE and MAP estimators provided similar SNR improvement. For example, at 10 dB input SNR, the MMSE estimator provided signals with an SNR of 14.4–15.5 dB, while the SNR of the signals estimated by the MAP estimator was 14.0–15.0 dB.

B. Signal Coding

The performance of the two-step AR model vector quantizer using the MMSE sample spectrum estimator (65) was studied in [79]. This estimator was compared with a version of the two-state spectral subtraction estimator proposed in [24]. A comparison with another approach, in which the vector quantizer is designed for the clean signal and the input noise is simply ignored during encoding, was also performed.

The version of the two-state spectral subtraction estimator is given by

$$|\widehat{Y_t(k)}|^2 = \Pr(\text{speech present}\,|z_t)\max\{[|Z_t(k)|^2 - S_v(k)], \epsilon\}, \tag{76}$$

where $\Pr(\text{speech present}\,|z_t)$ is the probability of speech present in the current vector of noisy speech z_t, $|Z_t(k)|^2$ is the discrete sample spectrum of z_t, $S_v(k)$ is an estimate of the power spectral density of the noise process as obtained from the fourth-order Gaussian AR model, and ϵ is an arbitrarily chosen small positive number which prevents the estimated spectrum from being nonpositive. The probability $\Pr(\text{speech present}\,|z_t)$ was calculated similarly to $p_\lambda(\bar{x}_t|z_t)$ in (65), using a two-state HMM with Gaussian AR subsources for the speech signal. One state of this model represents speech presence and the other speech absence. The two-state HMM was estimated from the given training data of clean speech signals using the VQ approach as in subsection V-D. The AR models for the "silence" state and the "nonsilence" state were estimated as the centroids of vectors of the training data whose power was below and above a preset threshold, respectively. The threshold was set at 30 dB below the average power of the vectors in the training data.

Table 1 Average Itakura–Saito Distortion in AR model VQ of Noisy Speech Using the MMSE Estimator (65), the Spectral Subtraction (SPS) Estimator (76), and No Estimator (DIRECT)

	MMSE		SPS		DIRECT
SNR	train	test	train	test	test
5	1.97	2.24	3.14	2.79	4.06
10	1.73	1.84	2.45	2.20	3.09
15	1.43	1.67	1.88	1.72	2.46
20	1.12	1.31	1.43	1.37	1.92
∞	0.31	0.60	0.31	0.60	0.60

Table 1 shows the average of the Itakura–Saito distortion $d(|Y_t(\theta)|^2, Q(z_t))$ (18), obtained when 8-bit per frame (256 samples) AR model vector quantizers of order 10 were designed and tested. The distortion measured during both training and testing is shown.

The results in the table show that the MMSE estimator outperforms the spectral subtraction estimator at all input SNR's, and using either estimator is preferable to ignoring the input noise. A comparison between the MMSE estimation approach and the direct approach, in encoding the test signals, shows that the average distortion obtained by the former approach at 10 dB input SNR is similar to that obtained by the direct approach at 20 dB input SNR. Hence, the MMSE estimation approach achieves an improvement equivalent to a 10 dB increase in input SNR. A similar comparison between the spectral subtraction approach and the direct approach shows that here the equivalent improvement in input SNR is about 5 dB only. Hence, the average distortion achieved by the MMSE estimator at this practically important input SNR of 10 dB, is in fact significantly lower than that obtained by the spectral subtraction approach, even though an improved version of the spectral subtraction estimator has been used here. The improvement is in the sense that the parameters

Table 2 Average (ave.) Recognition Scores and the Corresponding Standard Deviations (sd.) of Multispeaker Recognition of the English Digits

	Nonadapted Gain		Adapted Gain	
SNR	ave.	sd.	ave.	sd.
5	87.00	8.71	84.00	10.90
10	90.75	4.47	92.75	5.17
15	93.00	2.69	96.25	3.75
20	94.25	4.61	98.25	2.75
30	94.75	4.53	99.75	0.75
∞	86.50	14.01	99.50	1.00

of the posterior probabilities of the two-state estimator (76) were systematically estimated from training data rather than experimentally chosen.

C. Signal Classification

The gain-adapted MAP decision rule was tested in [82] in recognition of noisy speech signals corresponding to the English digits. This decision rule was compared with the MAP decision rule (70) where no gain adaptation was applied. In that case, both the "shape" and the "gain" [139] of the AR models of the HMM's for the clean signals were exclusively estimated from the training data, and recognition was performed using M AR models per word. This is in contrast to the gain-adapted approach, in which only the "shape" of the AR models is estimated from the training data but the gain is estimated from the given noisy signal. Thus, the M AR models per word can theoretically be combined with an infinite number of gain factors.

Table 2 shows the recognition accuracy obtained in a multispeaker (two males and two females) recognition task, where the training and test speech signals were recorded under similar gain conditions. In this case, gain adaptation due only to quasi-stationarity of speech signals is performed. This implies that the results given here are conservative for any practical application, since usually gain mismatch between the test and training data exists. In this experiment, training and recognition were performed using 20 and 40 utterances from each word spoken by all four speakers, respectively. HMM's with ten states and a single mixture component per state were used for the clean signals. The PD's of the subsources of the HMM's were assumed Gaussian AR of order $N_s = 10$. The dimension of output vectors from these sources was $K = 256$ at a sampling rate of 8 kHz.

The gain-adapted approach provided higher recognition accuracy than the nonadapted approach by 2%–13%, at input SNR's of from 10 dB to ∞, respectively. At the low SNR of 5 dB, the nonadapted approach outperforms the gain-adapted approach by 3%, apparently because at that low input SNR gain estimation from the noisy data is less reliable than the gain estimation performed *a priori* from the given training data. Another interesting result from Table 2 is that the nonadapted gain approach performs better at 30 dB input SNR than when the input signal is clean. This implies that better HMM's for the clean signal can be obtained if the diagonal values of covariance matrices of the models designed for the clean signal are slightly increased.

The same recognition task was also studied in [82] for training and test signals which were recorded under different gain conditions. It was shown that in this case, the gain-adapted MAP decision rule is significantly more robust than the nonadapted MAP decision rule. Specifically, the gain-adapted approach provided recognition accuracy similar to that shown in Table 2 where no recording gain mismatch existed, while the performance of the nonadapted approach dropped at all SNR's to a level similar to that obtained in Table 2 for an input SNR of 5 dB (i.e., 87.00%).

The typical number of iterations performed by the gain-adapted algorithm was 2 for recognition of clean signals, and 10–20 for recognition of noisy signals at 10 dB input SNR.

X. Discussion

The purpose of this paper was to integrate recent research on the model-based approach for speech enhancement, which has become the dominant approach in this area in recent years. The model-based approach was developed following two major breakthroughs in the areas of statistical modeling of speech signals, made in the mid 1970's and early 1980's, by the introduction of HMM's and VQ, respectively. These powerful statistical methods allowed the application of information theoretic approaches for signal estimation, coding, and classification given noisy signals. A unified framework for these problems was developed in this paper, and was shown to be consistent with more traditional approaches developed earlier from perceptual viewpoints. The latter approaches were well documented by Lim and Oppenheim in their 1979 tutorial paper [3]. The model-based approach provides a systematic way of implementing and expanding these approaches using statistics measured directly from training data from the signal and noise.

The selection of the particular type of HMM's for the signal and noise, and of the distortion measure, are the key to successful design of speech enhancement systems. We focused here on HMM's with Gaussian subsources which have proved useful for speech signals, and which are potentially useful for a broad class of noise sources. We concentrated on the MSE and MAP distortion measures for signal estimation, since they lead to estimators which are optimal for a wide class of distortion measures. Furthermore, the MMSE estimators are the optimal preprocessors for signal coding and recognition in the Itakura–Saito sense and the minimum probability of error sense, respectively. Very intuitive speech enhancement systems were obtained using this type of HMM and distortion measure. Furthermore, the resulting estimators are modular and are amenable to VLSI implementation. HMM's with either non-AR Gaussian subsources or with non-Gaussian non-AR subsources and different distortion measures (see, e.g., [133]) are possible in the model-based approach, but

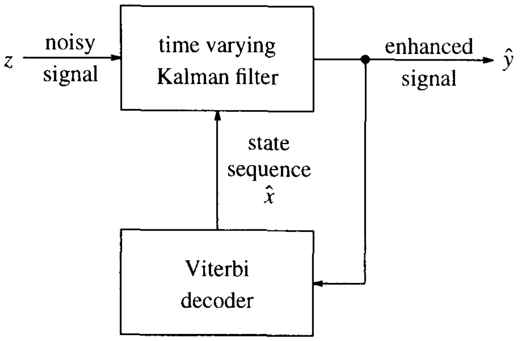


Fig. 16. Kalman filtering–Viterbi decoding enhancement.

they should lead to systems which are significantly more complicated than those studied here. The advantage of using HMM's with Gaussian subsources, and the MSE or MAP distortion measures, is that the resulting estimators are mixtures of linear estimators. If the number of states is sufficiently large, then the PD of the speech signal can be well represented, and effective filtering results.

The HMM-based speech enhancement approach can be extended in several directions. The first is by using more elaborate HMM's than those considered here, for example, HMM's in which the output vectors from a given sequence of states form a first-order vector Markov process [119], [120]. In these models, vectors generated from a given sequence of states are statistically dependent rather than independent, as is the case here. Unfortunately, the resulting speech enhancement systems using this model become significantly more complicated. A relatively simple case in this framework is that of noncausal MAP signal estimation in which the entire utterance of the clean signal is estimated from the entire utterance of the given noisy signal similarly to [37]. In this case, the iterative scheme of Fig. 9 is generalized to an iterative Viterbi decoding–Kalman filtering enhancement approach as shown in Fig. 16 [120].

A second possible extension of the HMM-based enhancement approach is to use models which have been adapted to the clean signal observed through the given noisy signal, for example, models whose spectral shapes are estimated from the training data but whose gain contours are estimated from the noisy signal. Two strategies for gain adaptation were discussed here. However, other approaches are possible, for example, Bayesian approaches for gain estimation. Adaptation of the spectral shapes of the models are also possible, for example, by supplementing the AR spectral estimate of each subsource by the pitch frequency pattern, as was done in adaptive transform coding of speech [168]–[170]. Preliminary experiments show that this approach improves the SNR of the enhanced signal by about 0.5 dB at an input SNR of 10 dB, but no noticeable perceptual improvement effects have yet been obtained with this approach. Another important adaptation aspect is that of using variable frame length analysis which is adapted to the different sounds of speech. Here a fixed frame length ($K = 256$) was used. This frame length may, however, be too long for certain speech sounds (e.g., plosives) and too short for other, more stationary sounds (e.g., vowels). Such adaptation can be elegantly accomplished by applying the HMM-based enhancement approach to the wavelet transform of the signal [171, ch. 8] rather than to the signal itself. This may provide more meaningful time–frequency representation of the signal and, hence, more meaningful modeling of that signal. Finally, the model for noise must be adapted if the noise is not strictly stationary. Alternatively, if different stationary noise sources are expected, then their models can be *a priori* designed and stored in the speech enhancement system.

A third possible extension is that of designing neural-network-model-based speech enhancement systems. Neural networks can be trained for realizing nonlinear transformations (filters) between the input noisy signal and the desired output clean signal [42]–[44]. Hence, such networks can be trained to realize the conditional mean estimator for each composite state of the noisy signal. With this approach, a nonparametric model for the conditional mean of the clean signal given the noisy signal is implicitly used. This approach for speech enhancement is similar to the hidden control neural network approach for speech recognition applications proposed in [172].

A fourth possible extension of the model based speech enhancement approach is to incorporate higher order statistics of the noisy signal [171, ch. 7], [173], [174]. Such statistics may naturally be needed if the state-dependent PD's are not assumed Gaussian. Similarly, higher order statistics may be useful if a nonlinear parametric model for the output signal from each state is used, whether this model is excited by Gaussian or non-Gaussian noise.

An important issue which has not yet been resolved is the estimation of the optimal order of the HMM's for the signal and noise. By the order of the HMM (say, for the signal) we mean the number of states M, the number of mixture components per state L, and the order of the AR model N_s (or of any other parametric model) for each subsource. In particular, it is not clear what is the optimal total number of subsources ($I \triangleq M \times L$) one should use, and what is the optimal factorization of I into M and L. In the systems reviewed here, HMM's with $I = 16 - 512$ subsources were used. Improved results should be expected if the optimal order of the models is found.

In summary, we have attempted to provide a common statistical framework for the three basic problems of speech enhancement; to specify the statistical knowledge about the signal, the noise, and the auditory system needed for optimal solution of these problems; to show relations between the solutions of the signal estimation, coding, and recognition problems; and to provide a systematic approach for the solution of these problems within the framework of Bayesian inference using HMM's. We strongly believe that the three problems of speech enhancement ought to have their solutions within the same statistical modeling framework. Solutions which work well for one of these problems but fail when applied to another problem may be based on models which do not capture the essence of the signal and noise. We hope that this paper will establish a first step toward such a unified solution and that it will encourage further research on this challenging, important problem.

Acknowledgment

The work summarized in this paper reflects research on the speech enhancement problem during the past ten years. Many of the author's colleagues have significantly contributed to his knowledge and understanding of this hard problem. He is especially grateful to D. Malah, I. Bar-David, S. Shamai (Shitz), J. Ziv, R. M. Gray, A. Dembo, R. J. McAulay, L. R. Rabiner, N. Merhav, and B.-H. Juang.

Special thanks go the perceptive anonymous referees who provided many helpful comments which significantly improved the presentation of this paper.

References

Speech Signal Estimation

[1] J. S. Lim, Ed., *Speech Enhancement.* Englewood Cliffs, NJ: Prentice-Hall, 1983.

[2] J. Makhoul *et al.*, *Removal of Noise from Noise-Degraded Speech Signals.* Panel on removal of noise from a speech/noise signal, National Research Council. Washington, DC: National Academy Press, 1989.

[3] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.

[4] S. F. Boll, "Speech enhancement in the 1980's: Noise suppression with pattern matching," in *Advances in Speech Signal Processing,* S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992.

[5] D. O'Shaughnessy, "Enhancing speech degraded by additive noise or interfering speakers," *IEEE Commun. Mag.*, pp. 46–52, Feb. 1989.

[6] I. Pollack, "Speech communications at high noise levels: The role of a noise operated automatic gain control system and hearing protection," *J. Acoust. Soc. Amer.*, vol. 29, pp. 1324–1327, Dec. 1957.

[7] R. J. Niederjohn and J. H. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, pp. 202–207, Aug. 1976.

[8] R. J. Niederjohn and J. H. Grotelueschen, "Speech intelligibility enhancement in a power generating noise environment," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, pp. 208–210, Aug. 1978.

[9] I. B. Thomas and R. J. Niederjohn, "The intelligibility of filtered-clipped speech in noise," *J. Audio Eng. Soc.*, vol. 18, pp. 299–303, June 1970.

[10] I. B. Thomas and W. J. Ohley, "Intelligibility enhancement through spectral weighting," in *Proc. IEEE Conf. Speech, Commun., and Processing,* 1972, pp. 360–363.

[11] J. C. R. Licklider, "Effects of amplitude distortion upon the intelligibility of speech," *J. Acoust. Soc. Amer.*, vol. 29, pp. 429–434, Oct. 1946.

[12] K. D. Kryter, J. C. R. Licklider, and S. S. Stevens, "Premodulation clipping in AM voice communication," *J. Acoust. Soc. Amer.*, vol. 19, pp. 125–131, Jan. 1947.

[13] E. A. Kretsinger and N. B. Young, "The use of fast limiting to improve the intelligibility of speech in noise," *Speech Monographs,* vol. 27, no. 1, Mar. 1960.

[14] J. S. Lim and A. V. Oppenheim, "Reduction of quantization noise in PCM speech coding," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, pp. 107–110, Feb. 1980.

[15] Y. Ephraim and D. Malah, "Combined enhancement and adaptive transform coding of noisy speech," *Proc. Inst. Elec. Eng.*, vol. 133, pt. F, no. 1, pp. 81–86, Feb. 1986.

[16] I. B. Thomas and A. Ravindran, "Intelligibility enhancement of already noisy speech signals," *J. Audio Eng. Soc.*, vol. 22, pp. 234–236, May 1974.

[17] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.* vol. 27, pp. 113–120, Apr. 1979.

[18] B. Widrow *et al.*, "Adaptive noise cancellation: Principles and applications," *Proc. IEEE*, vol. 63, pp. 1692–1716, Dec. 1975.

[19] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, pp. 197–210, June 1978.

[20] M. R. Weiss, E. Aschkenasy, and T. W. Parsons, "Processing speech signals to attenuate interference," in *IEEE Symp. Speech Recognition* (Pittsburgh, PA), Apr. 1974, pp. 292–293.

[21] A. Feit, "Intelligibility enhancement of noisy speech signals," M.Sc. thesis, Department of Electrical Engineering, Technion-Israel Institute of Technology, July 1973 (in Hebrew).

[22] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Aprl 1979, pp. 208–211.

[23] J. S. Lim, "Evaluation of correlation subtraction method for enhancing speech degraded by additive white noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, pp. 471–472, Oct. 1978.

[24] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, pp. 137–145, Apr. 1980.

[25] M. M. Sondhi, C. E. Schmidt, and L. R. Rabiner, "Improving the quality of a noisy speech signal," *Bell Syst. Tech. J.*, vol. 60, no. 8, pp. 1847–1859, Oct. 1981.

[26] P. Vary, "On the enhancement of noisy speech," in *Signal Processing II: Theories and Applications.* New York: Elsevier, 1983, pp. 327–330.

[27] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, pp. 679–681, Aug. 1982.

[28] Y. Ephraim and D. Malah, ‘Speech enhancement using a minimum mean square error short time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, pp. 1109–1121, Dec. 1984; see also *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (Boston), Apr. 1983, pp. 1118–1121.

[29] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, pp. 725–735, Apr. 1992.

[30] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, pp. 443–445, Apr. 1985.

[31] H. Drucker, "Speech processing in a high ambient noise environment," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, no. 2, pp. 165–168, June 1968.

[32] J. H. L. Hansen and M. A. Clements, "Iterative speech enhancement with spectral constraints," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Apr. 1987, pp. 189–192.

[33] K. K. Paliwal, "Speech enhancement using multi-pulse excited linear prediction filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Apr. 1986, pp. 101–104.

[34] K. K. Paliwal, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Apr. 1987, pp. 177–180.

[35] B. Koo, J. D. Gibson, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* May 1989, pp. 349–352.

[36] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Mar. 1984, pp. 18A.2.1–18A.2.4.

[37] Y. Ephraim, D. Malah, and B.-H. Juang "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 1846–1856, Dec. 1989.

[38] D. O'Shaughnessy, "Speech enhancement using vector quantization and a formant distance measure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Apr. 1988, pp. 549–552.

[39] T. F. Quatieri and R. J. McAulay, "Noise reduction using a soft-decision sine-wave vector quantizer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Apr. 1990, pp. 821–824.

[40] Y. Ephraim, "A minimum mean square error approach for

speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (Albuquerque, NM), Apr. 1990, pp. 829–832.

[41] Y. Ephraim and N. Merhav, "Lower and upper bounds on the minimum mean square error in composite source signal estimation," *IEEE Trans. Inform. Theory*, vol. 38, Nov. 1992.

[42] S. Tamura and A. Waibel, "Noise reduction using connectionist models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1988, pp. 553–556.

[43] S. Tamura and M. Nakamura, "Improvement to the noise reduction neural network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1990, pp. 825–828.

[44] M. W. White, R. M. Holdaway, J. J. Paulos, and S. Tambwekar, "The goal of perceptual fidelity in speech enhancement," Tech. Rep. NETR-89-7, North Carolina State University.

[45] M. S. Ahmed, "Comparison of noisy speech enhancement algorithms in terms of LPC perturbation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 121–125, Jan. 1989.

[46] K. K. Paliwal, "Estimation of noise variance from the noisy AR signal and its application in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, pp. 292–294, Feb. 1988.

[47] A-Chuan Hsueh and C. K. Chuang, "A multipulse excited pole-zero filtering approach for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1988, pp. 545–548.

[48] M. R. Sambur, "Adaptive noise canceling for speech signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, pp. 419–423, Oct. 1978.

[49] J. W. Kim and C. K. Un, "Enhancement of noisy speech by forward/backward adaptive digital filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1986, pp. 89–92.

[50] Y. Ariki, K. Kajimoto, and T. Sakai, "Acoustic noise reduction by two dimensional spectral smoothing and spectral amplitude transformation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1986, pp. 97–100.

[51] D. E. Veeneman and B. Mazor, "A fully adaptive comb filter for enhancing block-coded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 995–957, June 1989.

[52] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Amer.*, vol. 60, pp. 911–918, Oct. 1976.

[53] N. D. Degan and C. Prati, "Performance of speech enhancement techniques for mobile radio terminal applications," in *Signal Processing III: Theories and Applications*. New York: Elsevier Publishers B.V. (North Holland), 1986, pp. 381–385.

Encoding of Noisy Sources

[54] G. S. Kang and L. J. Fransen, "Quality improvement of LPC processed noisy speech by using spectral subtraction," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 37, pp. 939–942, June 1989.

[55] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[56] D. B. Paul, "The spectral envelope estimation vocoder," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 29, pp. 786–794, Aug. 1981.

[57] J. D. Gibson, T. R. Fisher, and B. Koo, "Estimation and vector quantization of noisy speech," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, Apr. 1988, pp. 541–544.

[58] R. L. Dobrushin and B. S. Tsybakov, "Information transmission with additional noise," *IRE Trans. Inform. Theory*, vol. 18, pp. S293-S304, 1962.

[59] T. Fine, "Optimum mean-square quantization of a noisy input," *IEEE Trans. Inform. Theory*, pp. 293–294, Apr. 1965.

[60] D. J. Sakrison, "Source encoding in the presence of random disturbance," *IEEE Trans. Inform. Theory*, pp. 165–167, Jan. 1968.

[61] H. S. Witsenhausen, "Indirect rate-distortion problems," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 518–521, Sept. 1980.

[62] Y. Ephraim and R. M. Gray, "A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization," *IEEE Trans. Inform. Theory*, vol. 34, pp. 826–834, July 1988.

[63] B.-H. Juang and L. R. Rabiner, "Signal restoration by spectral mapping," *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, Apr. 1987, pp. 2368–2371.

[64] J. K. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Inform. Theory*, vol. 16, pp. 406–411, July 1970.

[65] E. Ayanoglu, "On optimal quantization of noisy sources," *IEEE Trans. Inform. Theory*, vol. IT-36, pp. 1450–1452, Nov. 1990.

[66] M. R. Sambur, "A preprocessing filter for enhancing LPC analysis/synthesis of noisy speech," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, Apr. 1979, pp. 971–974.

[67] J. L. Melsa and J. D. Tomcik, "Linear predictive coding with additive noise for application to speech digitization," in *Proc. 14th Allerton Conf. Circuit and Systems Theory*, 1976, pp. 500–508.

[68] R. P. Preuss, "A frequency domain noise canceling preprocessor for narrowband speech communication systems," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, 1979, pp. 212–215.

[69] B. R. Musicus and J. S. Lim, "Maximum likelihood parameter estimation of noisy data," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, 1979, pp. 224–227.

[70] D. P. McGinn and D. H. Johnson, "Estimation of all-pole parameters from noise corrupted sequences," in *Proc. ASSP Spectrum Estimation Workshop II*, 1983, pp. 108–111.

[71] S. Kay, "Improvement of autoregressive spectral estimates in the presence of noise," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, 1978, pp. 357–360.

[72] L. Marple, "High resolution autoregressive spectrum analysis using noise power cancellation," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, 1978, pp. 345–348.

[73] S. M. Kay and S. L. Marple, Jr., "Spectrum analysis—A modern perspective," *Proc. IEEE*, vol. 69, pp. 1380–1419, Nov. 1981.

[74] V. K. Jain and B. S. Atal, "Robust LPC analysis of speech by extended correlation matching," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, 1985, pp. 473–476.

[75] R. D. Preuss and R. Yarlagadda, "Autoregressive spectral estimation in noise in the context of speech analysis," in *Proc. ASSP Spectrum Estimation Workshop II* (Tempa, FL), Nov. 1983, pp. 75–79.

[76] W. J. Done and C. K. Rushforth, "Estimating the parameters of a noisy all-pole process using pole-zero modeling," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, 1979, pp. 228–231.

[77] M. Pagano, "Estimation of models of autoregressive signal plus white noise," *Ann. Statist.*, vol. 2, pp. 99–108, 1974.

[78] S. M. Kay, "Noise compensation for autoregressive spectral estimates," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, pp. 292–303, June 1980.

[79] Y. Ephraim, "On minimum mean square error speech enhancement," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing* (Toronto), May 1991, pp. 997–1000.

Speech Recognition in Noise

[80] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1990, pp. 845–848.

[81] A. Nádas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 37, pp. 1495–1503, Oct. 1989.

[82] Y. Ephraim, "Gain-adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Trans. Signal Processing*,, vol. 40, pp. 1303–1316, June 1992.

[83] N. Merhav and Y. Ephraim, "A Bayesian classification approach with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 2157–2166, Oct. 1991.

[84] T. Kailath, "A general likelihood-ratio formula for random signals in Gaussian noise," *IEEE Trans. Inform. Theory*, vol. 15, pp. 350–361, May 1969.

[85] A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1990, pp. 849–852.

[86] A. Erell and M. Weintraub, "Estimation using log-spectral distance criterion for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1990, pp. 853–855.

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

[88] G. A. Neben, R. J. McAulay, and C. J. Weinstein, "Experiments in isolated word recognition using noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Apr. 1983, pp. 1156–1159.

[89] D. V. Compernolle, "Spectral estimation using a log-distance error criterion applied to speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* May 1989, pp. 258–261.

[90] D. Van Compernolle, "Noise adaptation in a hidden Markov model speech recognition system," *Computer Speech and Language,* vol. 3, no. 2, pp. 151–167, Apr. 1989.

[91] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 39, pp. 795–805, Apr. 1991.

[92] J. H. L. Hansen and M. A. Clements, "Stress compensation and noise reduction algorithms for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* May 1989, pp. 266–269.

[93] D. Mansour and B.-H. Juang, "The short-time modified coherence representation and noisy speech recognition," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 37, pp. 795–804, June 1989.

[94] I. Lecomte, M. Lever, J. Boudy, and A. Tassy, "Car noise processing for speech input," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* May 1989, pp. 512–515.

[95] A. Noll *et al.*, "Real-time connected-word recognition in a noisy environment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* May 1989, pp. 679–682.

[96] B. P. Landell, R. E. Wohlford, and L. G. Bahler, "Improved speech recognition in noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (Tokyo), 1986, pp. 14.14.1–14.14.3.

[97] J. N. Holmes and N. C. Sedgewick, "Noise compensation for speech recognition using probabilistic models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Apr. 1986, pp. 741–744.

[98] D. B. Roe, "Speech recognition with a noise-adapting codebook," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing,* Apr. 1987, pp. 1139–1142.

[99] A. Varga, R. Moore, J. Bridle, K. Ponting, and M. Russell, "Noise compensation algorithms for use with hidden Markov model based speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Apr. 1988, pp. 481–484.

[100] O. Ghitza, "Robustness against noise: the role of timing-synchrony measurements," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing,* Apr. 1987, pp. 2372–2375.

[101] M. J. Hunt and C. Lefebvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing* (Glasgow), May 1989, pp. 262–265.

[102] H. Matsumoto and H. Imai, "Comparative study of various spectrum matching measures on noise robustness," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Apr. 1986, pp. 769–772.

[103] B. A. Hanson and H. Wakita, "Spectral slope distance measures with linear prediction analysis for word recognition in noise," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 35, pp. 968–973, July 1987.

[104] F. Soong and M. M. Sondhi, "A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 36, pp. 41–48, Jan. 1988.

[105] J.-C. Junqua and H. Wakita, "A comparative study of cepstral lifters and distance measures for all pole models of speech in noise," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing* (Glasgow), May 1989, pp. 476–479.

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

[107] N. Merhav and C.-H. Lee, "A minimax classification approach with application to robust speech recognition," *IEEE Trans. Speech and Audio Processing*, to be published.

Hidden Markov Models

[108] J. D. Ferguson, Ed., *Proc. Symp. Applications of Hidden Markov Models to Text and Speech.* IDA-CRD, Princeton, NJ, 1980.

[109] A. B. Poritz, "Hidden Markov models: A guided tour," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing,* Apr. 1988, pp. 7–13.

[110] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, pp. 257–286, Feb. 1989.

[111] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.,* vol. 27, pp. 379–423, 623–656, 1948.

[112] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[113] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, vol. 37, pp. 1554–1563, 1966.

[114] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41, no. 1, pp. 164–171, 1970.

[115] L. E. Baum, "An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes," *Inequalities*, vol. 3, no. 1, pp. 1–8, 1972.

[116] L. A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, vol. 28, pp. 729–734, Sept. 1982.

[117] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of Markov process to automatic speech recognition," *Bell Syst. Tech. J.*, vol. 62, no. 4, pp. 1035–1074, Apr. 1983.

[118] C.-H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic modeling for large vocabulary," *Comput. Speech Language*, vol. 4, pp. 127–165, 1990.

[119] C. J. Wellekens, "Explicit time correlation in hidden Markov models for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Apr. 1987, pp. 384–386.

[120] Y. Ephraim, "Speech enhancement using state dependent dynamical system model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Mar. 1992, pp. 289–292.

[121] B.-H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Process.,* vol. 23, pp. 1404–1413, Dec. 1985.

[122] N. Merhav and Y. Ephraim, "Maximum likelihood hidden Markov modeling using a dominant sequence of states," *IEEE Trans. Signal Processing,* vol. 39, pp. 2111–2115, Sept. 1991.

[123] N. Merhav and Y. Ephraim, "Hidden Markov modeling using a dominant state sequence with application to speech recognition," *Comput. Speech Language*, vol. 5, pp. 327–339, 1991.

[124] A. B. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proc. of the Symposium on the applications of hidden Markov models to text and speech* (Princeton, NJ), 1980, pp. 88–142; summarized in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing,* Apr. 1982, pp. 1291–1294.

[125] Y. Ephraim, A. Dembo, and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1001–1013, Sept. 1989.

[126] Y. Ephraim and L. R. Rabiner, "On the relations between modeling approaches for speech recognition," *IEEE Trans. Inform. Theory.* vol. 36, pp. 372–380, Mar. 1990.

[127] L. R. Rabiner, J. G. Wilpon, and B.-H. Juang, "A segmental k-means training procedure for connected word recognition," *AT&T Tech. J.*, pp. 21–40, May-June 1986.

[128] B.-H. Juang and L. R. Rabiner, "The segmental k-means algorithm for estimating parameters of hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, pp. 1639–1641, Sept. 1990.

Distortion Measures

[129] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and format frequencies," *Electron. Commun. Japan*, vol. 53-A, no. 1, pp. 36–43, 1970.

[130] R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, pp. 367–376, Aug. 1980; vol. 34, pp. 1033–1040, Sept. 1988.

[131] K. Dzhaparidze, *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series.* New York: Springer-Verlag, 1986.

[132] Y. Ephraim, H. Lev-Ari, and R. M. Gray, "Asymptotic minimum discrimination information measure for asymptotically

weakly stationary processes," *IEEE Trans. Inform. Theory,* vol. 34, pp. 1033–1040, Sept. 1988.

[133] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing,* vol. 18, no. 4, pp. 349–369, Dec. 1989.

Vector Quantization

[134] A. Gersho and V. Cuperman, "Vector quantization: A pattern matching technique for speech coding," *IEEE Commun. Mag.*, vol. 21, pp. 15–21, Dec. 1983.

[135] R. M. Gray, "Vector quantization," *IEEE ASSP Mag.,* pp. 4–29, Apr. 1984.

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

[137] A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, pp. 562–574, Oct. 1980.

[138] R. M. Gray, A. H. Gray, Jr., G. Rebolledo, and J. E. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," *IEEE Trans. Inform. Theory*, vol. 27, pp. 708–721, Nov. 1981.

[139] M. J. Sabin and R. M. Gray, "Product code vector quantizers for waveform and voice coding," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 32, pp. 474–488, June 1984.

[140] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Boston: Kluwer Academic 1991.

[141] J. Foster, R. M. Gray, and M. O. Dunham, "Finite-state vector quantization for waveform coding," *IEEE Trans. Inform. Theory*, vol. 31, pp. 348–359, May 1985.

[142] M. O. Dunham and R. M. Gray, "An algorithm for the design of labeled-transition finite-state vector quantizers," *IEEE Trans. Commun.*, vol. 33, pp. 83–89, 1985.

[143] L. C. Stewart, R. M. Gray, and Y. Linde, "The design of trellis waveform coders," *IEEE Trans. Commun.*, vol. pp. 702–710, Apr. 1982.

[144] M. J. Sabin and R. M. Gray, "Global convergence and empirical consistency of the generalized Lloyd algorithm," *IEEE Trans. Inform. Theory*, vol. 32, pp. 148–155, Mar. 1986.

[145] R. M. Gray and E. Karnin, "Multiple local optima in vector quantizers," *IEEE Trans. Inform. Theory*, vol. 28, pp. 256–261, Mar. 1982.

[146] R. M. Gray, J. C. Kieffer, and Y. Linde, "Locally optimal block quantizer design," *Information and Control*, pp. 178–198, May 1980.

EM Algorithm

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

[148] C. F. J. Wu, "On the convergence properties of the *EM* algorithm," Ann. Statist., vol. 11, no. 1, pp. 95–103, 1983.

MMSE Estimation in Control Theory

[149] D. T. Magill, "Optimal adaptive estimation of sampled stochastic processes," *IEEE Trans. Automat. Contr.*, vol. 10, pp. 434–439, Oct. 1965; cf. author's reply, *IEEE Trans. Automat. Contr.*, vol. 14, pp. 216–218, Apr. 1969.

[150] F. L. Sims and D. G. Lainiotis, "Recursive algorithm for calculation of the adaptive Kalman filter weighing coefficients," *IEEE Trans. Automat. Contr.*, vol. 14, pp. 215–217, Apr. 1969.

[151] D. G. Lainiotis, "Optimal adaptive estimation: Structure and parameter adaptation," *IEEE Trans. Automat. Contr.*, vol. 16, pp. 160–170, Apr. 1971.

[152] L. A. Liporace, "Variance of Bayes estimators," *IEEE Trans. Inform. Theory*, vol. 17, pp. 665–669, Nov. 1971.

[153] D. Kazakos, "New convergence bounds for Bayes estimators," *IEEE Trans. Inform. Theory*, vol. 27, pp. 97–104, Jan. 1981.

[154] L. Merakos and D. Kazakos, "Comments and corrections to new convergence bounds for Bayes estimators," *IEEE Trans. Inform. Theory*, vol. 29, pp. 318–320, Mar. 1983.

[155] G. Saridis, *Self Organizing Control of Stochastic System.* New York: Marcel Dekker, 1977.

Miscellaneous

[156] H. L. Van-Trees, *Detection, Estimation and Modulation Theory,* part I. New York: Wiley, 1968.

[157] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech.* New York: Springer-Verlag, 1976.

[158] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.

[159] T. Kailath *Lectures on Linear Least-Square Estimation.* New York: Springer-Verlag, 1976.

[160] R. M. Gray, *Probability, Random Processes, and Ergodic Properties.* New York: Springer-Verlag, 1988.

[161] T. M. Cover, "A hierarchy of probability density function estimates," in *Frontiers of Pattern Recognition*, S. Watanabe, Ed. New York: Academic Press, 1972.

[162] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review,* vol. 26, no. 2, pp. 195–239, Apr. 1984.

[163] T. F. Quatieri and R. J. McAulay, "Phase coherence in speech reconstruction for enhancement and coding applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* May 1989, pp. 207–210.

[164] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products.* New York: Academic, 1980.

[165] R. M. Gray, "Toeplitz and circulant matrices: II," Tech. Rep. 6504–1, Stanford Electron. Lab., Apr. 1977.

[166] U. Grenander and G. Szego, *Toeplitz Forms and Their Applications.* New York: Chelsea Publishing, 1984.

[167] L. R. Rabiner and S. E. Levinson, "Isolated and connected word recognition theory and selected applications," *IEEE Trans. Commun.*, vol. 29, pp. 621–659, May 1981.

[168] J. M. Tribolet and R. E. Crochiere, "Frequency domain coding of speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 512–530, Oct. 1979.

[169] J. M. Tribolet and R. E. Crochiere, "A modified adaptive transform coding scheme with post-processing-enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Apr. 1980, pp. 336–339.

[170] J. S. Lim, A. V. Oppenheim, and L. D. Braida "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, pp. 354–358, Aug. 1978.

[171] S. Haykin, Ed., *Advances in Spectrum Analysis and Array Processing,*, vol. I. Englewood Cliffs, NJ: Prentice-Hall, 1991.

[172] E. Levin, "Word recognition using hidden control neural architecture," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (Albuquerque, NM), May 1990, pp. 433–436.

[173] C. L. Nikias and M. R. Raghuveer, "Bispectrum estimation: A digital signal processing framework," *Proc. IEEE*, pp. 869–891, July 1987.

[174] J. M. Mendel, "Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications," *Proc. IEEE*, pp. 278–305, Mar. 1991.

[175] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE*, vol. 61, pp. 268–278, Mar. 1973.

Yariv Ephraim (Senior Member, IEEE) received the D.Sc. degree in electrical engineering in 1984 from the Technion, Israel Institute of Technology, Haifa, Israel. He was a Rothschild Post-Doctoral Fellow at the Information Systems Laboratory of Stanford University, Stanford, CA, in 1984 and 1985.

Since 1985 he has been a Member of Technical Staff in the Speech Research Department of AT&T Bell Laboratories, Murray Hill, NJ. Since 1991 he has also been a Visiting Research Associate Professor at the C^{3}I Research Center of George Mason University, Fairfax, VA.