Enhancement of Connected Words in an Extremely Noisy Environment

Yuval Cohen, Adoram Erell, *Member, IEEE,* and Yuval Bistritz, *Senior Member, IEEE*

***Abstract*— A speech enhancement algorithm that is based on a connected-word hidden Markov model (HMM) is developed. Speech is assumed to be highly degraded by statistically independent additive noise. The minimum mean square error estimator is derived for a connected-word HMM. Further, we derive an estimator based on a connected-word HMM with explicit state duration. Listening experiments performed with digit strings have shown an increase of intelligibility. The best results were achieved when subjects who listened to the enhanced speech were given the results of an automatic recognition system.**

***Index Terms*— Noise reduction, robustness in the presence of noise, speech recognition**

I. Introduction

SPEECH enhancement concerns the improvement of perceptual aspects for human listening. This includes improving the speech quality, its intelligibility, and degree of listener fatigue. Different approaches have been applied to enhance degraded speech signals. Among them are approaches that exploit perceptual aspects of speech such as the periodicity of speech, or an underlying model for speech production. Methods based on spectral subtraction have been widely used. Other systems operate on more than one input, exploiting the correlation of the noise. These approaches are well presented in [1] and [2]. The statistical-model-based approach assumes that the joint statistics of the signal and the noise is known; for instance, a hidden Markov model (HMM) can be assumed. An optimal solution, in the statistical sense, to the speech enhancement problem can be defined by minimizing the expected value of a given distortion measure between the clean and the estimated speech signals.

Speech enhancement methods that are based on HMM's have been recently studied by Ephraim for enhancing speech signals recorded via a single microphone [3]–[5]. In his work, Ephraim addresses the problem of enhancing continuous, speaker-independent speech, degraded by a statistically independent additive noise at input signal-to-noise ratio (SNR) of above 5 dB. Ephraim takes a statistical approach to develop a minimum mean square error (MMSE) estimator, and a maximum *a posteriori* (MAP) estimator based on an acoustic speech model. The model is expressed in terms of a Gaussian autoregressive (AR) HMM. Noise levels considered have been such that the noisy signal is mostly intelligible, addressing applications of improving the performance of speech communication systems in noisy environments.

In the current work, we consider the case where speech, available from a single source, is highly degraded (SNR of less than 0 dB). This very low input SNR transforms the problem from improving speech quality for a more convenient listening to the problem of making the recorded speech intelligible. Our goal is to maximize the total number of correctly recognized words, without posing demands for real time. Possible applications for a small data base are recognition of telephone numbers or bank account numbers from a noisy recorded speech by intelligence and police surveillance, investigation of disputed credit card numbers provided over the phone in a noisy environment, and other similar situations where the goal is to maximize the total number of correctly recognized words spoken in a noisy environment, without posing demands for real time.

We present here a speech enhancement algorithm that is optimized for the case of our interest. We utilize a connected-word model, thus exploiting phonetic and linguistic information of the given vocabulary. We develop an enhancement algorithm for a connected-word model that includes explicit state duration. The above statistical model can be used also for automatic speech recognition. Under the model assumptions, the resulting recognizer is optimally adapted to the noise. The possibility of enhancing human performance by being given the results of the recognizer is also examined.

In order to evaluate the performance of the speech enhancement and recognition algorithms, we considered the set of digits as a test vocabulary. Although it is a small vocabulary, the concept of whole-word recognition and estimation can be fully tested. Moreover, the problem of recognizing and enhancing connected-digit strings is of practical use.

The enhancement algorithm was evaluated by the following three criteria:

1) SNR;
2) Itakura–Saito distortion measure, which is known to be correlated to speech intelligibility;
3) human listening tests performed by untrained listeners.

The paper is organized as follows. Section II presents the speech models and enhancement algorithms. The three following models are considered:

Manuscript received April 23, 1994; revised March 3, 1996. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. John H. L. Hansen.

Y. Cohen is with the Electrical Engineering Group, RND Networks, Ltd., Tel Aviv 61131, Israel.

A. Erell is with the Speech Group, DSPC Israel, Ltd., Givat Shmuel 51905, Israel.

Y. Bistritz is with the Department of Electrical Engineering–Systems, Tel Aviv University, Tel-Aviv 69978, Israel (e-mail: bistritz@eng.tau.ac.il).

Publisher Item Identifier S 1063-6676(97)01946-9.

1) the acoustic HMM (used by Ephraim);
2) a connected-word HMM;
3) a connected-word HMM with explicit state duration model.

Section III describes implementation issues and the tests procedures. An enhancement procedure that involves an interaction of the listener, an automatic recognizer, and the estimator is described in Section IV. The tests results are given in Section V and discussed in Section VI.

II. MMSE Estimation

This section describes the MMSE estimation based on a connected-word HMM. First we review Ephraim's model and his estimation algorithm. Then we show how to extend his formulation to connected-words, utilizing linguistic information.

A. Acoustic HMM

This section summarizes the MMSE enhancer developed by Ephraim [4]. The clean speech signal was assumed to be the output of an HMM with M states and L mixtures of Gaussian AR processes. Thus, at each time frame this model represents speech as a vector generated by the Gaussian AR process associated with the particular state and mixture component chosen at that given time. An ergodic topology was chosen so that the resulting acoustic ergodic HMM (EHMM) model could serve as a speaker and as a context independent, continuous speech model. The noise was assumed to be a stationary process characterized by Gaussian AR vectors which are statistically independent and identically distributed (this is identical to a one-state, one-mixture component HMM).

The model parameters were estimated using the segmental K-means algorithm [6]. This algorithm approximates the Baum algorithm with less computations. An initial model is obtained by applying the generalized Lloyd algorithm in conjunction with the Itakura–Saito distance measure to a subset of the training data to create an M entry code book. Then, all training set vectors are clustered using the estimated codebook and the data within each cluster is used for designing code words representing the L mixture components of that state. The mixture component probability vector c is also estimated during this procedure. A uniform initialization of the initial state probabilities π, and the probability transition matrix, A, completes an initial model estimation. This model is the starting point for an iterative procedure in which the state and mixture component paths are detected via the Viterbi algorithm, then the model parameters (π, A, c) along with the parameter set of the AR processes of the HMM are updated. Noise model training was performed simply by calculating the centroid of the noise model training data vectors.

Let y_t, n_t, z_t denote the time frame t vectors of K samples of clean speech, noise, and noisy speech, respectively. Let $Y_t(k)$, $Z_t(k)$ be the kth elements of the discrete Fourier transform (DFT) of the clean speech and of the noisy speech. It is assumed that the noise is additive and statistically independent of the clean speech, and that the speech and noise can be represented by the above models. Under these assumptions the MMSE estimator of the speech is a filter W that takes into account all the possible states and mixture components of the model. The estimation $\hat{Y}_t(k)$ is assumed to be a Gaussian vector with independent elements, thus its elements can be independently estimated by the following expression:

$$\hat{Y}_t(k) = W_t(k)Z_t(k) = \left[\sum_{\alpha=1}^{M}\sum_{\gamma=1}^{L} q_t^{\alpha,\gamma} W_t^{\alpha,\gamma}(k)\right] Z_t(k) \quad (1)$$

where $W_t^{\alpha,\gamma}(k)$ denotes the frequency domain Wiener filter, given that the state and mixture component are α and γ, and $q_t^{\alpha,\gamma}$ denotes the posterior probability of this particular state and mixture component given the noisy signal observations until time frame t. The estimated speech signal, $\hat{y}_t$, can be obtained by taking the inverse DFT of $\hat{Y}_t(k)$.

Let $S_y^{\alpha,\gamma}(k)$, $S_n(k)$ be the power spectra of the clean speech and noise code words (calculated by dividing the variance of the AR source innovation by the DFT of the auto correlated coefficients of the AR process), then

$$W_t^{\alpha,\gamma}(k) = \frac{S_y^{\alpha,\gamma}(k)}{S_y^{\alpha,\gamma}(k) + S_n(k)}. \quad (2)$$

The posterior probabilities for the noisy signal are equal to the normalized forward probabilities associated with the noisy signal, i.e.,

$$q_t^{\alpha,\gamma} = \overline{F}_t^{\alpha,\gamma} = \frac{F_t^{\alpha,\gamma}}{\sum_{\alpha=1}^{M}\sum_{\gamma=1}^{L} F_t^{\alpha,\gamma}} \quad (3)$$

where $F_t^{\alpha,\gamma}$ is the forward probability associated with the noisy signal.

The posterior probabilities are efficiently calculated [6], [7] by the following recursion:

$$q_t^{\alpha,\gamma} = \frac{\sum_{\nu=1}^{M}\sum_{\mu=1}^{L} q_{t-1}^{\nu,\mu} a_{\nu\alpha}\, c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma)}{\sum_{\alpha=1}^{M}\sum_{\gamma=1}^{L}\sum_{\nu=1}^{M}\sum_{\mu=1}^{L} q_{t-1}^{\nu,\mu}\, a_{\nu\alpha}\, c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma)} \qquad t>0$$

$$q_0^{\alpha,\gamma} = \frac{\pi_\alpha\, c_{\gamma|\alpha}\, f(z_0|\alpha,\gamma)}{\sum_{\alpha=1}^{M}\sum_{\gamma=1}^{L} \pi_\alpha\, c_{\gamma|\alpha}\, f(z_0|\alpha,\gamma)} \quad (4)$$

where $\alpha_{\nu\alpha}$ is the state transition probability from state ν to state α and $c_{\gamma|\alpha}$ is the mixture coefficient. Since a Gaussian AR model was assumed, the probability density function (pdf) of the noisy speech sample vector given the state and the mixture component, $f(z_t|\gamma,\alpha)$, can be shown to be Gaussian with covariance matrix which is equal to the sum of the clean speech and the noise covariance matrices.

B. Connected Digit HMM

This subsection presents a Gaussian AR connected-word estimator. It differs from the EHMM by the topology (the A matrix) and the training process. The connected word model incorporates linguistic information concerning the structure of individual words. Specifically, the A matrix contains all states from all words, having a form of a block matrix. Each block corresponds to a specific word probability transition matrix. Additional elements are added for the transition probabilities between words.

The connected word estimator was realized for the set of digits. Each digit was characterized by an HMM with a left–right Bakis topology [8] with a fixed number of states and L mixture components. Word transition probabilities were implemented in the last state of each digit allowing transitions to that state itself or to any first state of a digit. A special word in the vocabulary was designated for silence.

Similarly to the EHMM, the parameter set of our connected digit HMM (DHMM) model for the speech signal was estimated using the segmental K-means algorithm. However, in this case the Viterbi algorithm was supervised by the known digit sequence of each sentence in the training set. The initial model was constructed by combining separately created digit models. The initial digit model was created by applying the generalized Lloyd algorithm to a training data subset, manually segmented into words. Due to the nature of the vocabulary (digits only), the transition probabilities of the last state of each digit were forced to be equal. (A detailed description of the data base is given in Section III.)

Since the separate digit models were combined to one model by creating a global transition matrix in which all digit states participate, speech enhancement may proceed using the same (1)–(4) as in the EHMM case. For the current model, M is equal to the total number of states in all digits.

In the above causal approach, the forward probability is used for calculating the posterior probability for the state and mixture component given the noisy speech. A noncausal approach is obtained by using the forward-backward probabilities to compute the posterior probabilities given all observations of the noisy signal. We anticipate that it will be advantageous to use the DHMM with the forward-backward procedure in cases where the SNR at the end of the word is better than at its beginning. For this approach, the calculation of the posterior probabilities in the enhancement algorithm has to be changed as follows:

$$q_t^{\alpha,\gamma} = \frac{\overline{F}_t^{\alpha,\gamma}\overline{B}_t^{\alpha,\gamma}}{\sum_{\alpha=1}^{M}\sum_{\gamma=1}^{L}\overline{F}_t^{\alpha,\gamma}\overline{B}_t^{\alpha,\gamma}} \tag{5}$$

where $\overline{F}_t\alpha,\gamma$ is the normalized forward probability defined in (3) and $\overline{B}_t^{\alpha,\gamma}$ is the normalized backward variable recursively calculated as follows:

$$\overline{B}_t^{\alpha,\gamma} = \frac{\sum_{\nu=1}^{M}\sum_{\mu=1}^{L}\overline{B}_{t+1}^{\nu,\mu}a_{\alpha\nu}c_{\mu|\nu}f(z_{t+1}|\nu,\mu)}{\sum_{\alpha=1}^{M}\sum_{\gamma=1}^{L}\sum_{\nu=1}^{M}\sum_{\mu=1}^{L}\overline{B}_{t+1}^{\nu,\mu}a_{\alpha\nu}c_{\mu|\nu}f(z_{t+1}|\nu,\mu)}$$

$$\overline{B}_{T-1}^{\alpha,\gamma} = \frac{1}{\sum_{\alpha=1}^{M}\sum_{\gamma=1}^{L}} = \frac{1}{ML}. \qquad t<T-1 \tag{6}$$

C. Connected Digit HMM with State Duration

In this Section we develop an estimator based on a connected-digit HMM with explicit state duration (DDHMM). This part of the work was motivated by reported evidence that recognition rates for speaker-dependent systems have been improved by adding explicit duration model to the HMM [9].

The state duration probability $p_\alpha(d)$ was estimated by building a duration histogram for each state out of the state sequence revealed by the Viterbi algorithm during the training process. Following [10], the normalized histograms were associated to the particular states as the discrete probability functions of the state duration.

The transition matrix A was modified so that all state transitions from a state to itself ($a_{\alpha,\alpha}$) were set to zero.

Speech signal estimation was based on (1). The additional state duration affects the calculation of the posterior probabilities, as more possible paths can be taken to reach a specific state. A well-known recursion exists for the calculation of the forward probabilities, which includes state duration [11]. These forward probabilities, which are associated with the noisy signal, are defined by

$$F_t^{\alpha,\gamma} = P(z_0 z_1 \cdots z_t, \gamma, \alpha \text{ ends at } t) \tag{7}$$

the probability of being at state α and mixture component γ assuming that state α ends at time t. For speech recognition, the recursion is followed by selection of an optimal path so that all false "assumptions" that a particular state ends at time t will not be taken into account.

For the case of estimation, we introduce a new forward variable. Define

$$\mathcal{F}_t^{\alpha,\gamma} = P(z_0 z_1 \cdots z_t, \gamma, \alpha) \tag{8}$$

as an absolute (total) probability of being at state α. This value (normalized by the sum over all α,γ) then multiplies the Wiener filter. The absolute forward probability is calculated by adding to the forward probability the possibilities of the current state ending at time $t+1, t+2, \cdots, t+D-1$, where D is the maximum permissible duration. A detailed derivation of the absolute forward probability is given in the Appendix. It is shown there that the product of the state duration pdf is substituted by one minus the state duration probability distribution (cumulative) function, so that

$$\mathcal{F}_t^{\alpha,\gamma} = \sum_{\nu=1}^{M}\sum_{\mu=1}^{L}\sum_{d=1}^{D} F_{t-d}^{\nu,\mu}\, a_{\nu\alpha}\, M_\alpha(d) c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma) \cdot \prod_{s=t-d+1}^{t-1} f(z_s|\alpha) \tag{9}$$

where

$$M_\alpha(d) = \sum_{k=d}^{D} p_\alpha(k) = 1 - \sum_{k=1}^{d-1} p_\alpha(k) \tag{10}$$

is the state duration probability distribution (cumulative) function, $F_{t-d}^{\nu,\mu}$ are the forward probabilities, and $f(z_s|\alpha) = \sum_{\eta=1}^{L} c_{\eta|\alpha} f(z_s|\alpha,\eta)$ is the probability of observing z_s given a state α. (Products and sums assume null values of 1 and 0, respectively, when subindex exceeds upper index.)

To conclude, speech enhancement can be formulated as in (1),

$$\hat{Y}_t(k) = \left[\sum_{\alpha=1}^{M}\sum_{\gamma=1}^{L}\mathcal{Q}_t^{\alpha,\gamma}W_t^{\alpha,\gamma}(k)\right]Z_t(k) \tag{11}$$

where $\mathcal{Q}_t^{\alpha,\gamma}$ is defined as the absolute posterior probability given by the normalization

$$\mathcal{Q}_t^{\alpha,\gamma} = \frac{\mathcal{F}_t^{\alpha,\gamma}}{\sum_{\alpha=1}^{M}\sum_{\gamma=1}^{L}\mathcal{F}_t^{\alpha,\gamma}}. \tag{12}$$

D. Scaling

The recursion formulae for the forward probability and for the absolute forward probability (9) are prone to numerical instability. The stability of the procedure may be improved by introducing some appropriate scaling and as a result the recursion is performed on the posterior probabilities rather than on the forward probabilities [7], [11]. The procedure is a modification to the basic scaling as in (4), but takes in consideration that all previously calculated q_t must be normalized by the same factor. The following equation gives the resulting recursion for the scaled forward probability $F_t'^{\alpha,\gamma}$

$$\begin{aligned} F_t'^{\alpha,\gamma} &= \sum_{\nu=1}^{M}\sum_{\mu=1}^{L}\sum_{d=1}^{D} q_{t-d}^{\nu,\mu}\, a_{\nu\alpha}\, p_\alpha(d)\, c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma) \\ &\quad \cdot \prod_{s=t-d+1}^{t-1} f(z_s|\alpha) w_s \qquad t > D-1 \\ F_t'^{\alpha,\gamma} &= \pi_\alpha\, p_\alpha(t+1)\, c_{\gamma|\alpha} f(z_t|\alpha,\gamma) \prod_{s=0}^{t-1} f(z_s|\alpha) w_s \\ &\quad + c_{\gamma|\alpha} f(z_t|\alpha,\gamma) \sum_{\nu=1}^{M}\sum_{\mu=1}^{L}\sum_{d=1}^{t} q_{t-d}^{\nu,\mu}\, \alpha_{\nu\alpha}\, p_\alpha(d) \\ &\quad \cdot \prod_{s=t-d+1}^{t-1} f(z_s|\alpha) w_s \qquad 0 \le t \le D-1 \end{aligned} \tag{13}$$

where the scale factors w_t are given by

$$w_t = \left[\sum_{\alpha=1}^{M}\sum_{\gamma=1}^{L} F_t'^{\alpha,\gamma}\right]^{-1} \tag{14}$$

and the posterior probabilities q_t are given by

$$q_t = w_t F_t'^{\alpha,\gamma}. \tag{15}$$

However, we encountered cases where this procedure still led to an unstable behavior. This occurred when $F_t'^{\alpha,\gamma}$ was very small for all α, γ, causing the scale factor w_t to be very large. As w_t is involved in the calculations of $F_{t+1}'^{\alpha,\gamma}, F_{t+2}'^{\alpha,\gamma}, \cdots, F_{t+D}'^{\alpha,\gamma}$ having nothing to balance it, their values tended to grow beyond the computer's dynamic range. In fact, depending on the A matrix and the state duration pdf, cases where $F_t'^{\alpha,\gamma} = 0 \forall\, \alpha, \gamma$ may also be encountered and imply divisions by zero.

In order to overcome such difficulties we developed a new scheme for stable forward probabilities recursion. Define

$$f'(z_t|\alpha,\gamma) = k_t f(z_t|\alpha,\gamma) \tag{16}$$

where the scale factors k_t are given by

$$k_t = \frac{1}{\max_{\alpha,\gamma}[f(z_t|\alpha,\gamma)] \sum_{\alpha=1}^{M}\sum_{\gamma=1}^{L}\mathcal{F}_{t-1}'^{\alpha,\gamma}} \tag{17}$$

TABLE I
DATA BASE PARAMETERS

speaker	sentences in training set	sentences in testing set	digits in training set	digits in testing set	digits manually segmented
male	69	8	210	56	110
female	69	8	199	54	81

and perform recursion for $F_t'^{\alpha,\gamma}$ and $\mathcal{F}_t'^{\alpha,\gamma}$ as follows:

$$\begin{aligned} F_t'^{\alpha,\gamma} &= \sum_{\nu=1}^{M}\sum_{\mu=1}^{L}\sum_{d=1}^{D} F_{t-d}'^{\nu,\mu}\, a_{\nu\alpha}\, p_\alpha(d) c_{\gamma|\alpha}\, f'(z_t|\alpha,\gamma) \\ &\quad \cdot \prod_{s=t-d+1}^{t-1} f'(z_s|\alpha) \end{aligned} \tag{18}$$

$$\begin{aligned} \mathcal{F}_t'^{\alpha,\gamma} &= \sum_{\nu=1}^{M}\sum_{\mu=1}^{L}\sum_{d=1}^{D} F_{t-d}'^{\nu,\mu}\, a_{\nu\alpha}\, M_\alpha(d) c_{\gamma|\alpha}\, f'(z_t|\alpha,\gamma) \\ &\quad \cdot \prod_{s=t-d+1}^{t-1} f'(z_s|\alpha). \end{aligned} \tag{19}$$

It can easily be shown that

$$F_t'^{\alpha,\gamma} = k_0 k_1 \cdots k_t\, F_t^{\alpha,\gamma} \tag{20}$$

$$\mathcal{F}_t'^{\alpha,\gamma} = k_0 k_1 \cdots k_t\, \mathcal{F}_t^{\alpha,\gamma}. \tag{21}$$

Thus, all the introduced constants cancel out when the posterior probabilities q_t and $\mathcal{Q}_t$ are calculated. Note that the scale factor k_t is composed of two elements: The maximum function which brings $f(z_t|\alpha,\gamma)$ to the range [0, 1], and the sum of the previous scaled absolute probabilities. By definition this sum is always greater than the sum of the forward probabilities and is never equal to zero. Consequently divisions by very small numbers are avoided. The term $\sum_{\alpha=1}^{M}\sum_{\gamma=1}^{L}\mathcal{F}_{t-1}'^{\alpha,\gamma}$ acts as a feedback that keeps the values of $F_t'^{\alpha,\gamma}$ in the dynamic range. The feedback term itself is clipped to a lower and upper limit to prevent underflow or overflow.

III. IMPLEMENTATION ISSUES

For applications such as speaker dependent systems or surveillance concerning speech enhancement, a large clean speech data base for the subject speaker is not likely to be available. Therefore, contra to the general belief that HMM's need a large data base for properly estimating the HMM parameters, we chose to implement our algorithm on a relatively small data base of a single speaker. Thus, we picked a data bases for a male and for a female speaker from the TI connected digit data base [12]. The vocabulary consists of 11 digits (Zero/O, One, Two, ..., Nine). Each data base was divided into two independent sets: A training set and a testing set, so that the sentences in the testing sets were not included in the training sets. A selected subset that included approximately half of the training set was manually segmented for initialization of the HMM parameter set for each digit. Table I summarizes the parameters concerning the data base.

In order to extract the speech features, i.e., the coefficients of the AR process, speech was divided into frames of $K = 256$ samples obtained at an 8 kHz sampling rate. Speech frames were weighted by a trapezoidal window with a slope duration of eight samples.

Two types of noise sources were used: a Gaussian white noise generated by the computer and noise recorded with a desktop microphone in a computer room. As opposed to white noise that covers all the speech bandwidth, room noise is more dominant in the lower frequencies. Hence, sentences degraded by room noise were usually more intelligible than those degraded by white noise. The input SNR in all tests was less or equal to 10 dB. The choice of the lower limit aimed to a degradation level for which human recognition error rate falls in the 40–60% range. This requirement guided us into choosing minimal input SNR's of -13 dB for the room noise and 0 dB for white noise.

The noisy speech was enhanced by the MMSE estimators described in Section II. The ergodic HMM (EHMM) had $M = 8$ states, and $L = 32$ mixture components of Gaussian AR processes of order $N_y = 10$. The connected-digit models had five states per digit (a total of $M = 60$ states including a word for silence), and $L =$ five mixture components per state. These parameters were chosen based on tests of a wide range of values, and several SNR levels. A special word was assigned to represent silence portions of the clean speech. The model for this word was trained by the silence segments of the training data set found at the beginning of the sentences. Explicit state duration was added to the connected-digit model. Each state was assigned a state duration probability with a maximum range of $D = 25$. The noise was assumed to be the output of a Gaussian AR pdf. The noise spectrum was estimated from an average of all available noise model training sequences. These sequences were available from noisy sentences segments in which speech was absent (usually at the beginning or end of a sentence).

A. Performance Evaluation

Comparison between models is given by means of i) SNR values for the enhanced signal, ii) the Itakura–Saito distance between the enhanced sentence and the clean sentence, and iii) human listening tests performed by untrained listeners. The Itakura–Saito distortion measure is known to be more correlated to speech intelligibility than SNR. In practice however, both methods become unreliable when speech is highly distorted. Therefore, for highly degraded speech, the evaluation was done only by human intelligibility tests. These tests were performed in two stages: First, the noisy sentence was played several times until the listener felt he extracted all possible information from the sentence. The digits he recognized were written down. In the second phase, a similar process was repeated with the enhanced sentence obtained from one of the competing models. In this phase, the listener could listen upon request to both the noisy and the enhanced sentences. Finally, the average error rates were computed for both stages.

The MMSE enhancement algorithm can be logically divided into two steps: First, finding the most likely sequence of states and mixture components and second, applying the appropriate Wiener filter. In the following tables a comparison to an MMSE estimator with an optimal state path (OSP) is provided as an attempt to give a theoretical performance bound for the MMSE estimators based on Gaussian AR models [6]. This estimator is calculated by choosing the Wiener filter for each time frame according to the state and mixture component path determined from the clean speech. This path was determined by a Viterbi decoder using the DDHMM.

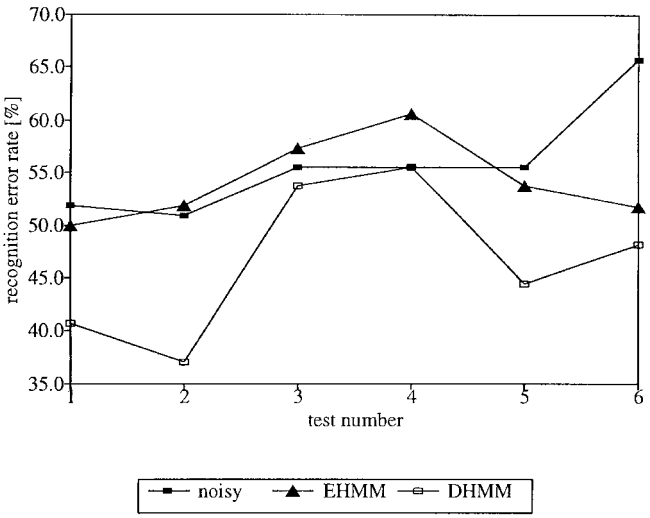

Fig. 1. Average error rates (%) obtained by six listeners for the recognition of noisy, EHMM enhanced sentences and DHMM enhanced sentences (female speaker, room noise).

IV. Man–Machine Interaction

In surveillance applications with very low input SNR's, the aim is to achieve a maximum recognition rate out of a given recorded speech with no constraint on time or computation effort. During preliminary tests, we noticed that human recognition rate was similar to that obtained by an automatic recognizer (a Viterbi recognizer adapted to noise [3]). However, the correctly recognized words were not always the same; i.e., the automatic recognizer succeeded in recognizing words that the listener failed to recognize, and vice versa. Motivated by these observations, we considered two interactive procedures to increase the overall recognition rate.

The first procedure is as follows: The listener who had finished determining all digits in the string was supplied with the digit string output of the automatic recognizer (providing the listener this string at an earlier stage might have influenced his result). Now he began a new session in which he was again requested to recognize the words in the sentence.

In a second procedure, the listener was allowed to affect the enhancer by supplying it with a suggested digit string. A Viterbi recognizer adapted to noise and supervised by this digit string produced a state sequence which was used by the enhancer. Alternatively, the enhancer was designed to accept the suggested digit string, and perform the enhancement using a constrained model. The listener could hear the sentence that resulted, subject to the constraints he had given. He was allowed to change his input string repeatedly and test the results until satisfied.

V. Experimental Results

The MMSE enhancers were applied to both female and male testing sentences in the presence of white noise. Average Itakura–Saito distortion is given in Table II for the

TABLE II
AVERAGE ITAKURA–SAITO DISTANCES OF THE ENHANCED TESTING SENTENCES (FEMALE SPEAKER, WHITE NOISE) OBTAINED BY USING THE MMSE ESTIMATORS, THE MMSE ESTIMATOR WITH AN OPTIMAL STATE PATH OBTAINED FROM CLEAN SPEECH (OSP), AND NO ESTIMATOR (NOISY)

SNR	DDHMM	DHMM-FB	DHMM	EHMM	OSP	NOISY
10	1.15	1.09	1.19	1.42	0.88	4.23
5	1.67	1.61	1.68	1.85	1.16	7.52
0	2.63	2.68	2.65	2.49	1.33	13.47

TABLE III
SAME AS TABLE II FOR A MALE SPEAKER

SNR	DDHMM	DHMM-FB	DHMM	EHMM	OSP	NOISY
10	1.20	1.17	1.23	1.28	0.81	7.96
5	1.50	1.51	1.56	1.86	1.02	11.92
0	1.86	1.98	1.88	2.39	1.07	17.60

TABLE IV
AVERAGE ERROR RATES (%) AND THE TOTAL IMPROVEMENT OBTAINED IN HUMAN RECOGNITION TESTS (FEMALE SPEAKER, ROOM NOISE AT −13 DB SNR). ENHANCEMENT PERFORMED BY THE MMSE ESTIMATORS USING THE ERGODIC HMM (EHMM) AND THE CONNECTED-DIGIT MODEL (DHMM)

model	noisy	enhanced	improvement(%)
EHMM	56.2	54.2	3.5
DHMM	55.6	46.6	16.2

female speaker and in Table III for the male speaker. Careful listening to the enhanced sentences convinced us that the perceived quality was consistent with the results shown in Tables II and III. In most sentences the connected-digit models (DHMM, DHMM-FB, DDHMM) outperformed the ergodic model (EHMM). This was more noticeable with lower input SNR levels. All tested models obtained similar SNR improvement. The estimators provided enhanced sentences with an average SNR of 15.5, 11.5, and 8.0 dB for 10, 5, and 0 dB input SNR's, respectively. Sentences enhanced using the OSP model were found to be less distorted than the sentences enhanced using other models, and were completely intelligible even at an SNR level of −13 dB room noise.

Spoken digit strings at a 10 dB input SNR level were noisy enough to be inconvenient for the listener, but produced almost no problem of intelligibility especially when unlimited number of playbacks were allowed. In order to compare the relative intelligibility obtained by the various models, we performed intelligibility tests with untrained listeners for lower input SNR levels. Each listener participated in two independent tests, one for the EHMM enhanced sentences and one for the DHMM enhanced sentences, following the procedure described in Section III-A. The average error rates (%) for the noisy and enhanced sentences (female speaker, room noise at a −13 dB SNR) for EHMM and for DHMM were computed for six independent listeners. The results are shown in Fig. 1. As the error rate of the noisy sentences recognition was found to be similar for both tests, only the average was plotted. The average results for all listeners, summarized in Table IV, indicate superiority of the DHMM. Moreover, the listeners described the quality of speech in DHMM-enhanced sentences as better than with EHMM enhancement even when they were able to recognize digits correctly with both models.

In Section IV we described two procedures for man–machine interaction. For the first approach, a third phase was added to the human recognition tests. In this phase, after listening to the noisy sentence (first phase) and the enhanced sentence (second phase), we provided the listener with the output of the automatic recognizer. These tests were performed on female testing sentences degraded by room noise at a −13 dB SNR, which were enhanced and recognized with the DHMM. The average error rate for five listeners was computed, and is shown in Table V.

TABLE V
AVERAGE ERROR RATES (%) OBTAINED IN MAN-MACHINE INTERACTION TESTS. DHMM ESTIMATOR WAS APPLIED TO NOISY SENTENCES (FEMALE SPEAKER, ROOM NOISE, −13 DB). THE LISTENER WAS SUPPLIED WITH THE OUTPUT OF THE AUTOMATIC RECOGNIZER

noisy	enhanced	human+recognizer info	improvement(%)
56.7	47.8	45.6	23.3

The second approach proposed in Section IV was also tested. A fourth phase was added to the recognition test in which the listener could suggest to the enhancer a digit string. The output of this guided enhancer was played to the listener so he could confirm his assumption or alter it. This iterative procedure was repeated until the listener was satisfied. Such tests for a −13 dB room noise failed to improve the overall recognition rates. However, this procedure was found to be of value for higher SNR's in cases where the sentence contained only a single questionable digit. For example, in one case a listener was able to recognize a digit correctly after trying three alternatives. In another case, where a digit was completely unrecognized, he was able to reduce the uncertainty to one of two alternatives.

Although the computation effort is irrelevant to the applications we address, we note here that enhancing a 3 s sentence on a Sparc10 workstation took 10, 14, 17, and 130 s for EHMM, DHMM, DHMM-FB, and DDHMM, respectively.

VI. CONCLUSIONS AND DISCUSSION

The purpose of the research presented in this paper was to develop and examine methods for achieving maximum human recognition for extremely noisy sentences. The basic model presented by Ephraim [4] was modified to utilize linguistic information by using a connected-word model with finite vocabulary. A further refinement of the model was the introduction of explicit state duration model. We examined two methods for achieving maximum recognition. First, MMSE estimators based on these models were derived. Second, two procedures for integrating the human and computer recognition abilities were studied.

The results in Tables II and III show that the DHMM out performs the EHMM, suggesting that linguistic information is indeed instrumental in the enhancement procedure. The explicit duration model did not produce any significant improvement in the perceptual experiments, but did improve automatic recognition. We hypothesize that the recognition rate improvement is due to a better selection of the state and mixture component path via the Viterbi algorithm. In the enhancement process, on the other hand, there is no selection of a unique path, but rather a summation over all states. Thus, as noise increases and the Viterbi path

becomes less dominant, the correlation between recognition and enhancement performances weakens.

The most significant result of our study is that processing the noisy signal with the DHMM estimator increased the intelligibility (Table IV). It is possible that the intelligibility increase is an outcome of more convenience in listening to the processed speech, which reduces the listener's fatigue, compared to the nonprocessed speech. The noise reduction encouraged the listener to hear the sentence many times, while the inconvenience of the noisy sentence usually led the listener to use fewer playbacks and possibly caused worse performance. The EHMM estimator also increased the SNR but the enhanced sentences suffered more distortion and the intelligibility did not improve.

The best intelligibility was achieved by combining the human and the machine recognition capabilities. The recognition error rates obtained for the automatic recognizer and for the listener were found to be at similar levels on independent tests. Making the recognizer's output available to the listener increased the overall intelligibility.

The second interactive approach, in which the estimator was guided by a digit string supplied by the listener, failed for very low input SNR levels. For these SNR levels the recognition error rate is approximately 50%, where almost half of the errors are due to deletions. Guiding the estimator along a partially mistaken path, with no indication of deleted words, led to poor results. For higher SNR levels, in which most of the string is correctly recognized, this approach is expected to show improvement. Preliminary tests with strings having only one questionable digit, have shown an increase in recognition rates.

In our work we considered very low SNR levels which would be unacceptable in most communication systems. Our setting and results are considered to be of practical interest in applications such as surveillance, where very low SNR levels are commonly encountered. A relatively small set of experiments was performed to evaluate the method. More experiments may be useful to obtain more reliable subjective assessments for the obtained results. The reported results are also limited to a very small vocabulary of 11 digits. It remains to be seen that the method developed here yields improvement over the EHMM with also a larger vocabulary. A possible extension to an unlimited vocabulary would be to use a phoneme-based HMM, which is less restrictive than a finite vocabulary but more restrictive than an ergodic HMM.

APPENDIX

A. Absolute Forward Probability Calculation

The absolute forward probability $\mathcal{F}_T^{\alpha,\gamma}$ as defined in (8) can be derived from the forward probability by adding it to the probability of the current state extending beyond time t. The recursion for the forward probability is given by

$$F_t^{\alpha,\gamma} = \sum_{\nu=1}^{M}\sum_{\mu=1}^{L}\sum_{d=1}^{D} F_{t-d}^{\nu,\mu}\, a_{\nu\alpha}\, p_\alpha(d)\, c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma) \cdot \prod_{s=t-d+1}^{t-1} f(z_s|\alpha) \qquad t > D-1$$

$$F_t^{\alpha,\gamma} = \pi_\alpha p_\alpha(t+1) c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma) \prod_{s=0}^{t-1} f(z_s|\alpha) + \sum_{\nu=1}^{M}\sum_{\mu=1}^{L}\sum_{d=1}^{t} F_{t-d}^{\nu,\mu}\, a_{\nu\alpha}\, p_\alpha(d)\, c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma) \cdot \prod_{s=t-d+1}^{t-1} f(z_s|\alpha) \qquad 0 \le t \le D-1. \tag{22}$$

Consider a current state at time frame t which has begun at time $t-d$. This state may last until time $t+1$ or $t+2$ etc., until time $t-d+D$. Therefore, an additional probability, $V_t^{\alpha,\gamma}(d)$ must be accounted for, as follows:

$$V_t^{\alpha,\gamma}(d) = \sum_{\nu=1}^{M}\sum_{\mu=1}^{L}\sum_{k=d+1}^{D} F_{t-d}^{\nu,\mu} a_{\nu\alpha} p_\alpha(k) c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma) \cdot \prod_{s=t-d+1}^{t-1} f(z_s|\alpha). \tag{23}$$

A total expression accounting for all possible starting points of the current state is given by

$$V_t^{\alpha,\gamma} = \sum_{d=1}^{D} V_t^{\alpha,\gamma}(d) = \sum_{\nu=1}^{M}\sum_{\mu=1}^{L}\sum_{d=1}^{D}\sum_{k=d+1}^{D} F_{t-d}^{\nu,\mu} a_{\nu\alpha} p_\alpha(k) c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma) \cdot \prod_{s=t-d+1}^{t-1} f(z_s|\alpha) \qquad t > D-1$$

$$V_t^{\alpha,\gamma} = \sum_{d=1}^{t} V_t^{\alpha,\gamma}(d) + \sum_{k=t+2}^{D} \pi_\alpha p_\alpha(k) c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma) \cdot \prod_{s=0}^{t-1} f(z_s|\alpha) = \sum_{\nu=1}^{M}\sum_{\mu=1}^{L}\sum_{d=1}^{t}\sum_{k=d+1}^{D} F_{t-d}^{\nu,\mu} a_{\nu\alpha} p_\alpha(k) c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma) \cdot \prod_{s=t-d+1}^{t-1} f(z_s|\alpha) + \sum_{k=t+2}^{D} \pi_\alpha p_\alpha(k) c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma) \cdot \prod_{s=0}^{t-1} f(z_s|\alpha) \qquad 0 \le t \le D-1. \tag{24}$$

We define the function $M_\alpha(d)$ to be one minus the state duration probability distribution (cumulative) function

$$M_\alpha(d) \triangleq \sum_{k=d}^{D} p_\alpha(k) = 1 - \sum_{k=1}^{d-1} p_\alpha(k). \tag{25}$$

Using the above definition and the fact that

$$\mathcal{F}_t^{\alpha,\gamma} = F_t^{\alpha,\gamma} + V_t^{\alpha,\gamma} \tag{26}$$

we obtain

$$\begin{aligned}\mathcal{F}_t^{\alpha,\gamma} &= \sum_{\nu=1}^{M}\sum_{\mu=1}^{L}\sum_{d=1}^{D} F_{t-d}^{\nu,\mu}\, a_{\nu\alpha}\, M_\alpha(d)\, c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma) \\ &\quad \cdot \prod_{s=t-d+1}^{t-1} f(z_s|\alpha) \qquad t > D-1 \\ \mathcal{F}_t^{\alpha,\gamma} &= \pi_\alpha M_\alpha(t+1) c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma) \prod_{s=0}^{t-1} f(z_s|\alpha) \\ &\quad + \sum_{\nu=1}^{M}\sum_{\mu=1}^{L}\sum_{d=1}^{t} F_{t-d}^{\nu,\mu}\, a_{\nu\alpha}\, M_\alpha(d)\, c_{\gamma|\alpha}\, f(z_t|\alpha,\gamma) \\ &\quad \cdot \prod_{s=t-d+1}^{t-1} f(z_s|\alpha) \qquad 0 \le t \le D-1. \end{aligned} \tag{27}$$

This result shows that the defined absolute probability is similar to the forward probability except that the duration pdf, $p_\alpha(d)$, is substituted for one minus the duration distribution function, $M_\alpha(d)$.

REFERENCES

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

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

[3] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, pp. 1526–1555, Oct. 1992.

[4] ——, "A minimum mean square error approach for speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Apr. 1990, pp. 829–832.

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

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

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

[8] L. R. Rabiner, J. G. Wilpon, and B.-H. Juang, "A model-based connected-digit recognition system using either hidden Markov models or templates," *Comput. Speech Lang.*, pp. 167–197, 1986.

[9] A. Ljolje and S. E. Levinson, "Development of an acoustic-phonetic hidden Markov model for continuous speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 29–39, Jan. 1991.

[10] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice-Hall, 1993.

[11] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Comput. Speech Lang.*, pp. 29–45, 1986.

[12] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Apr. 1984, pp. 42.11.1–42.11.4.

Yuval Cohen received the B.S.c. and M.S.c. degrees in electrical engineering (with distinction) from Tel-Aviv University, Israel, in 1987 and 1994, respectively.

From 1987 until 1992, he was a design engineer at the Israeli Defense Force. Since 1993, he has held several management positions in leading Israeli industrial companies. Currently, he is Manager of the Electrical Engineering Group at RND Networks Ltd., Tel Aviv, Israel.

Adoram Erell (M'97) received the B.Sc, M.Sc, and Ph.D. (cum laude) degrees in physics from Tel Aviv University, Israel.

From 1985 to 1991, he was a post-doctoral fellow and then a faculty member with the Department of Electrical Engineering, Tel-Aviv University, working on auditory modeling and application of speech perception to speech coding. During 1988 and 1989, he was an International Fellow at the Speech Research Laboratory, S.R.I., Menlo Park, CA, working on noise-robust speech recognition. From 1992 to 1994, he joined the algorithm-development group at M.B.T., Israeli Air Craft Industry, working on signal processing of radar signals. Since 1994, he has been with the Speech Group, DSPC Israel, Ltd., Givat Shmuel, Israel, leading research and development on various speech processing topics.

Yuval Bistritz (S'79–M'81–SM'87) received the B.S.c. degree in physics, the M.Sc. degree (summa cum laude), and the Ph.D. in electrical engineering, all from Tel Aviv University, Israel, in 1973, 1978, and 1983, respectively.

He served in the Israeli Defense Force from 1972 to 1975 and held a research engineer industrial position from 1975 to 1979. From 1979 to 1984 he held various assistant and teaching positions in the Department of Electrical Engineering–Systems, Tel-Aviv University. From 1984 to 1986 he was a Research Scholar in the Information System Laboratory at Stanford University, CA, doing research in fast signal processing algorithms. From 1986 to 1987, he was with AT&T Bell Laboratories, Murray Hill, NJ, working as a consultant in speech processing. Since 1987 he has been a faculty member in the Department of Electrical Engineering–Systems, Tel Aviv University, where he is currently Associate Professor. His research interests include speech processing and more general topics in signal processing and system theory.