On the Application of Mixture AR Hidden Markov Models to Text Independent Speaker Recognition

Naftali Z. Tishby

Abstract—**Linear predictive hidden Markov models have proved to be an efficient way for statistically modeling speech signals. The possible application of such models to statistical characterization of the speaker himself is described and evaluated. The results show that even with a short sequence of only four isolated digits, a speaker can be verified with an average equal-error rate of less than 3%. These results are slightly better than the results obtained using speaker dependent vector quantizers, with comparable numbers of spectral vectors. The small improvement over the vector quantization approach indicates the weakness of the Markovian transition probabilities for characterizing speaker dependent transitional information.**

I. Introduction

The identity of a speaker is important for many applications such as access control, telephone shopping, automatic money transfer, access to central data banks, etc. It also has a more fundamental interest. Separation of speaker dependent features from the spoken speech should also help in speaker normalization for automatic speech recognition. There are two basic modes of speaker recognition: identification and verification. In speaker identification we want to select the correct speaker out of a given population, by comparing his test utterances with those of all the population. In the verification problem, however, the speaker's identity is assumed to be known, and only his reference model is compared with the test utterances. While the speaker identification error rate (false identification) is bound to increase with the population size, the verification error rate, in principle, approaches a constant, independent of the population size. Speaker verification is generally more important than speaker identification for most commercial applications.

Speaker recognition can be performed on a known text or "password." In this case, conventional speaker dependent speech recognition techniques, such as dynamic time warping, can be used to recognize the speaker together with his password. In that way the text and the speaker are inseparable in the overall distortion score. When the text is known, time order information is provided so that the system is capable of giving better verification robustness. On the other hand, fixed text methods may be less convenient for the user, since the password is part of the verification process and must be remembered. Text dependent methods cannot be used when the speaker is uncooperative or when verification is required during a normal conversation. To avoid such problems, text independent or text free speaker recognition methods must be developed. In these methods, the speaker uses a random sequence of utterances or simply converses in a normal manner, while the recognition is carried out. Another advantage of text independent verification is that it can also be done in a sequential way, until a desired significance level is reached, without the annoyance of repeating the password again and again.

Previous work on automatic speaker recognition suggests that the important features are the spectral envelope parameters, given by the short-term autocorrelation coefficients of the speech samples. These can be transformed into another set of parameters, like the LPC or the LPC-derived cepstral coefficients. The spectral envelope is directly related to the physical dimensions of the speaker's vocal organs, and as such is a reliable speaker characterization. A vector quantization (VQ) approach to the problem, was developed by Soong *et al.* using the LPC likelihood ratio distortion [1] and by Rosenberg and Soong, using a weighted LPC-derived L_2 cepstral distortion measure [2]. The VQ method gives very good results based on a small size (16–64) codebook, and is the main reference for the present experiment. These researchers also showed that transitional information, in the form of the temporal derivatives of the cepstral coefficients, improved the performance of the verification system.

A possible way to incorporate temporal correlations in the VQ model, is by a Markov source of information, or a hidden Markov model (HMM). In such a model, the states of a first-order Markov chain are identified as short-term stationary sources, while the slower variations of the signal are modeled by the Markovian transitions between such states. If the signal in each state is modeled by an autoregressive source, we have a special type of HMM, which has been called AR or linear predictive HMM [3].

Hidden Markov models have been used in various ways in automatic speech recognition [4]. This paper examines the possible application of such models to the problem of automatic speaker recognition. Pioneering work in this direction was reported by Poritz in 1982 [3], when AR-hidden Markov models were first introduced. In this work, Poritz used a 5-state ergodic model, i.e., all possible transitions between states were allowed, to characterize each speaker, and reasonable discrimination among 10 speakers was achieved. The present work follows Poritz's original idea. We extend his work to the richer class of mixture autoregressive HMM's as developed by Juang and Rabiner [5]. In these models, the states are described by a linear combination (mixture) of AR sources. It can be shown that mixture models are equivalent to a larger HMM with simple states, together with additional constraints on the possible transitions between states [6]. Furthermore, a VQ-based system can also be viewed as a special (degenerate) case of an HMM, with a single state and discrete observation symbols. The main part of the present experiment is in constructing an HMM using the VQ algorithm. In this way the vector quantization codebook

Manuscript received July 13, 1988; revised April 18, 1990.
The author is with AT&T Bell Laboratories, Murray Hill, NJ 07974.
IEEE Log Number 9041601.

1053-587X/91/0300-0563$01.00 © 1991 IEEE

is embedded into the states and mixtures of the hidden Markov model.

In the next section, we review the elements of a mixture AR-HMM and describe our method for speaker verification. The experimental results and the importance of the different components of the model are discussed in Section III.

II. Speaker Verification Using Hidden Markov Models

A. *The Elements of an Autoregressive HMM*

A hidden Markov source is a stochastic function of a Markov chain. As such it is composed of two elements: a Markov process and a set of stochastic functions, or output probabilities. We briefly introduce these elements, together with the notation used further in this work.

1) The Markov Process: The Markov process composed of N symbolic states denoted by $S = \{S_1, S_2, \cdots, S_N\}$, which are supposed to represent the acoustic clusters of the speech signal, and transition probabilities between these states. We denote by q_t the actual state at time t, with $q_t \in S$.

The transitions between states are determined by the matrix $A \equiv (a_{ij})$, where

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i], \qquad 1 \leq i, j \leq N. \quad (2.1)$$

When the model is ergodic we generally mean that all transitions between states are possible. In a more restricted sense, we allow all the transitions in a single step, or $a_{ij} > 0$ for all i and j. An ergodic model is characterized by a unique stationary probability distribution of states, which can be used as the initial probabilities for the text independent modeling, i.e.,

$$\pi_j = P[q_1 = S_j] = P[q_t = S_j] \quad (2.2)$$

for a typical state sequence.

2) The Output Probabilities: To each of the states we attach a probability distribution for the observations, called the output probabilities and are the more important component of the HMM. The observation probability density at state j, is denoted by $b_j(O_t)$, where

$$b_j(O_t) = P[O_t | q_t = S_j]. \quad (2.3)$$

In the case of an AR-HMM, the output probabilities are determined by modeling the speech samples as a Gaussian autoregressive process. Consider a vector of K speech samples $O = (x_0, x_1, \cdots, x_{K-1})$. We assume that these samples are approximately related through the recursion

$$x_k = -\sum_{i=1}^{p} a_i x_{k-i} + e_k \quad (2.4)$$

where $\{e_k\}$ is the innovation sequence and $\{a_i\}$ are the p autoregression or linear prediction coefficients. If only the spectral envelope is considered we can use as the innovation $\{e_k\}$ Gaussian i.i.d.'s with zero mean and variance σ^2. The prediction coefficients $\{a_i\}$ are derived directly from the first $p + 1$ autocorrelations of the signal

$$r(i) = \sum_{n=0}^{K-i-1} x_n x_{n+i}, \qquad 0 \leq i \leq p. \quad (2.5)$$

Under these assumptions, and for large K, we can approximate [7] the probability density of the samples vector as

$$f_a(O) = (\sqrt{2\pi}\sigma)^{-K} \exp\left\{-\frac{1}{2\sigma^2}\delta(O; \boldsymbol{a})\right\} \quad (2.6)$$

where

$$\delta(O; \boldsymbol{a}) = r_a(0)\, r(0) + 2 \sum_{i=1}^{p} r_a(i)\, r(i) \quad (2.7)$$

and

$$r_a(i) = \sum_{n=0}^{p-i} a_n a_{n+i} \qquad (a_0 = 1) \qquad 0 \leq i \leq p.$$

We use the standard "residual energy normalization," i.e., in each frame we rescale the samples as

$$\hat{x}_n = \frac{x_n}{\sigma\sqrt{K}}. \quad (2.8)$$

In that case, the density (2.6) has the simpler form:

$$f_a(\hat{O}) = (\sqrt{2\pi})^{-K} \exp\left\{-\frac{K}{2}\delta(\hat{O}; \boldsymbol{a})\right\}. \quad (2.6a)$$

In practice we replace K with $\hat{K}$, an "effective" frame length, taking into account the overlap between adjacent frames. In our experiment K was 300 (45 ms), while $\hat{K}$ was only 100 samples, such that there was a 2/3 overlap between adjacent frames.

When autoregressive mixture densities are used, the observation density is given by a convex linear combination of M_j Gaussian autoregressive densities of the form (2.6). The output densities, $b_j(O)$, are thus specified by M_j densities, $b_{jm}(O)$, each of the form (2.6a), together with the mixture gain coefficients c_{jm}, such that

$$b_j(O) = \sum_{m=1}^{M_j} c_{jm} b_{jm}(O). \quad (2.9)$$

The overall spectral resolution of the model is determined by the total number of mixtures $M \equiv \Sigma_j M_j$.

The parameters of the model Λ are then given by the $N \times N$ transition matrix A, $M \times (p + 1)$ AR parameters, and M mixture gains $\{c_{jm}\}$. Denoting all the spectral shape parameters by $\boldsymbol{B}$, the model parameter set is written as $\Lambda \equiv (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$.

As in most other automatic pattern recognition systems, the operation consists of two phases: training and recognition. The quality of any model is determined by how well it fits the observed data during the training and the recognition phases. When the data is stochastic, or in the absence of complete understanding of the data when we want to treat it as stochastic, only a probabilistic quality measure can be given. The most natural measure in this case is the conditional probability of the observations given the model. The better the model describes the data the higher this probability, or likelihood, is. In the HMM case, the likelihood of the observations is explicitly given by:

$$P(\boldsymbol{O}|\Lambda) = \sum_{q_1, q_2, \cdots, q_T} \pi_{q_1} b_{q_1}(O_1) \cdot a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(O_T) \quad (2.10)$$

where $\boldsymbol{O} = O_1, O_2, \cdots, O_T$ are the sequences of overlapped vectors of speech samples. Both the training and the recognition phases involve calculation of this likelihood function. In the training phase the model is modified to maximize the likelihood of the training data, whereas in the recognition phase the model is given and the likelihood of the test data is calculated.

B. *The Training Phase*

The training problem is thus to find the most probable model parameters, given the data. In the absence of any prior statistical knowledge on the parameters, this is equivalent to the maximum likelihood estimation criterion:

$$\max_{\Lambda} P(\boldsymbol{O} | \Lambda). \tag{2.11}$$

The first training step is to select the dimensions of the model, i.e., the number of states, mixtures, and spectral parameters. Once these are determined, there is an effective dynamic programming algorithm, the forward-backward procedure [4], [8] to evaluate $P(\boldsymbol{O} | \Lambda)$, with complexity linear in T. In order to maximize the likelihood (3.1), the estimation-maximization (EM) iterative method [9], via the Baum-Welch reestimation algorithm [8] is used. Given an initial model, the forward-backward procedure is carried out to estimate $P(\boldsymbol{O} | \Lambda)$ (the E step); the model parameters are then modified such that the resultant likelihood function is maximized (the M step), by calculating their expectation values with the given P. Due to the large number of local maxima of the likelihood function (3.1) in the model's parameter space, in order to converge to a good model iteratively, a carefully selected initial model is needed.

1) The Initial Model: In his original experiment, Poritz used a random initial model with 5 simple 2-pole AR states. In our experiment a richer spectral description was used. We found, that with richer models, random initial conditions did not converge to a good model, and better clustering of the training data was necessary. Two types of initial clustering were examined in our experiment: an automatic procedure based on a VQ codebook [10] and a manual clustering using the leader algorithm [11]. The VQ algorithm was done with the likelihood ratio distortion measure, as given by (2.7). The whole training data was first clustered into 8 states using VQ. Then the members of each state were clustered into 2, 4, or 8 mixtures using VQ again, yielding an HMM with spectral resolution of 16, 32, or 64 vectors, respectively. The initial transition probabilities were estimated by counting the transitions between states in the training data. Although such an algorithm is suboptimal compared to a full VQ, we found it to be an effective way to generate ergodic hidden Markov models. The disadvantage of the VQ clustering is that the dimensions of the model (number of states and mixtures) are arbitrarily predetermined for all speakers.

In order to select the dimensions of the model more carefully, as well as to gain some insight into the clustering procedure, an alternative manual method, based on the "leader algorithm," was used. In this algorithm, the first frame is taken as a leader of the first cluster. When $k \geq 1$ leaders are found, the next leader is the first frame with a symmetric distance greater than a given threshold θ_1, from all the previous leaders. The clustering is done using the symmetric distortion distance between frames:

$$d(O_s, O_t) \equiv \tfrac{1}{2}[\delta(\hat{O}_s; \boldsymbol{a}_t) + \delta(\hat{O}_t; \boldsymbol{a}_s)]. \tag{2.12}$$

In this algorithm, the number of clusters is determined by the data, and is not fixed arbitrarily. The members of each cluster are all the frames with distance less than another threshold, $\theta_2 < \theta_1$, from the cluster leader. By adjusting the two thresholds, θ_1, the minimal distance between cluster leaders, and θ_2, the maximal distance from the leader within the cluster, we were able to select the number of states, within some average distortion. The same procedure was used within each state, for the initial mixture clustering.

2) Reestimation of the Parameters: In the next step, for both types of initial models, reestimation of the parameters using the EM method is needed. This can be done with either the Baum-Welch procedure or with the simpler segmental k-means algorithm [12]. In the segmental k-means, given some model parameters, the most probable state sequence is constructed, using the Viterbi decoding algorithm [13]. When the state sequence is known, the model parameters are reestimated by averaging within the frames of each state. This double maximization process guarantees the increase of the likelihood, and turns out to be a very efficient estimation method. We can further increase the likelihood of the training data, through the Baum-Welch reestimation, but this increase turns out to be negligible. Similar results have also been found in other studies using HMM in speech recognition [4]. The meaning of this is that the states are not truly "hidden" in practical applications of hidden Markov models to speech, since the actual state sequence is essentially known. The reestimation is iterated, until the final convergence criterion on the log likelihood per frame is met. The "quality" of the model is determined by its overall log likelihood per training data frame.

The manual training scheme is shown in Fig. 1.

C. The Verification Phase

It is easy to see that log $P(\boldsymbol{O}_T | \Lambda)$ grows, on the average, linearly with T. The log likelihood per frame

$$\frac{1}{T} \log P(\boldsymbol{O}_T | \Lambda) \tag{2.13}$$

is thus the natural normalized measure between the model Λ and the data sequence $\boldsymbol{O}_T$ (see also Fig. 2). This measure plays the role of the average distortion in other methods, e.g., the VQ based approach. Using the Viterbi algorithm again, we can calculate the log likelihood of the most probable state sequence in an efficient way, linear in T. The verification is done by comparing the normalized score to a threshold ξ_Λ, dependent on the desired verification statistical significance:

$$\begin{cases} \text{if } \dfrac{1}{T} \log P(\boldsymbol{O}_T | \Lambda) > \xi_\Lambda & \textit{accept speaker} \\ \text{otherwise} & \textit{reject speaker.} \end{cases} \tag{2.14}$$

We can roughly characterize the speaker's model performance by using a Gaussian approximation for the distributions of the log likelihood per frame for a given model over the correct speaker and the impostors. The true experimental distributions are skewed and obviously not Gaussian. However, by dealing with only the first two moments of the distributions, the mean μ_Λ, and the variance σ_Λ^2, it is possible to roughly describe the overall performance of the model, in a much simpler way. Furthermore, the central limit theorem guarantees the convergence of such distributions to a Gaussian, under very general conditions. Any deviation from the normal distribution indicate, therefore, interesting anomalies, that should be clarified. Denoting the impostors moments by $\tilde{\mu}_\Lambda$ and $\tilde{\sigma}_\Lambda^2$, respectively, the error probabilities, per frame, are given under this Gaussian approximation by

$$\begin{cases} P[\textit{false reject}] \equiv \alpha = \operatorname{erf}\left[(\mu - \xi)/\sigma\right] \\ P[\textit{false acceptance}] \equiv \beta = 1 - \operatorname{erf}\left[(\xi - \tilde{\mu})/\tilde{\sigma}\right] \end{cases} \tag{2.15}$$

where erf (x) is the standard error function.

If the different utterances were truly statistically independent, we would expect the variance of the distributions to decrease

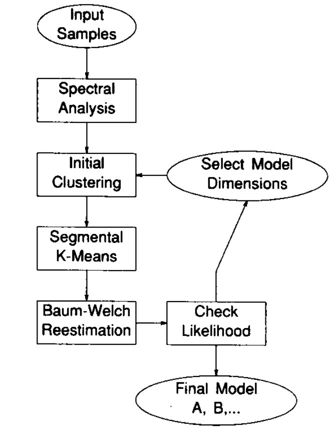


Fig. 1. The training phase of the HMM speaker recognizer.

$$S(Candidate \mid Model) = \frac{1}{T}\text{Log } P(O_T \mid \lambda)$$

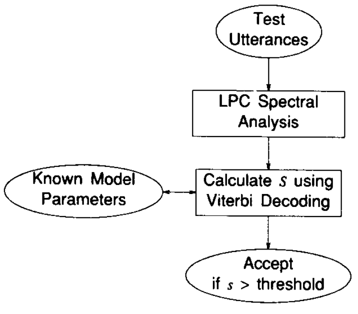

Test Utterances
LPC Spectral Analysis
Known Model Parameters
Calculate S using Viterbi Decoding
Accept if s > threshold

Fig. 2. The verification phase using the log likelihood score.

TABLE I
Log Likelihood per Frame Moments

	Interspeaker		Intraspeaker	
n	μ_n	σ_n	μ_n	σ_n
1	16.67	4.67	9.57	1.65
2	18.32	4.17	10.54	1.34
3	18.96	3.98	10.91	1.17
4	19.26	3.80	11.10	1.08
5	19.45	3.66	11.21	1.00
6	19.63	3.64	11.30	0.91
7	19.73	3.56	11.35	0.89
8	19.79	3.50	11.39	0.83
9	19.82	3.47	11.42	0.77
10	19.85	3.41	11.45	0.75

inversely with the test trial length, n; or:

$$\hat{\mu}_n = \mu, \qquad \hat{\sigma}_n = \sigma/\sqrt{n}. \tag{2.16}$$

The theoretical separation between the genuine speakers and the impostors populations should thus increase like $\sqrt{n}$ where n is the number of test utterances, as long as these are independent. In practice, we find that σ_n decays much slower than $1/\sqrt{n}$, so that the utterances are actually highly correlated (see Table I). This may indicate that with richer vocabulary (not just digits) the results may improve.

The standard way of evaluating a verification scheme is by adjusting the threshold ξ_Λ, so that the two error types have equal probability, called the "equal error rate."

III. Experimental Results and Discussion

A. The Experimental Setting

The data base we used consists of 20 000 isolated digit utterances, spoken by 100 speakers, 50 male and 50 female, over dialed-up local telephone lines. The recordings contain 20 repetitions of the 10 digits for each speaker, made in 5 sessions held over a period of up to two months. In each session, the speakers were prompted to utter 4 series of the 10 digits, in a randomized order. The analog speech, was bandpass filtered from 200 to 3200 Hz, and sampled at a 6.67 kHz rate. The speech samples were preemphasized by a first-order digital filter with the transfer function $1 - 0.95z^{-1}$. An eighth-order autocorrelation analysis was carried out every $\hat{K} = 100$ samples (15 ms) using overlapping $K = 300$ samples (45 ms) Hamming windows, in a single frame.

The training sample size depends on the number of parameters in the model, the vocabulary size, and the consistency of the speaker. It is clear and is supported by the results of previous experiments that more than one session is needed for good training of the speaker's model. In order to be able to compare our results with previous results obtained with the same data base, we used the same conditions: half of the data (100 digits in 3 sessions) for the training, while the second half of the data was used for the testing. Note that in that case, 20% of the testing data was recorded at the same time as 20% of the training data. This was done in order to examine the effect of the time between testing and training.

Several experiments were conducted. The main set of experiments used an automatically trained, VQ-based, initial model with 8 states and 2, 4, and 8 mixtures in each state. The results were compared, in all these experiments, to a VQ based experiment, using the likelihood ratio distortion measure, with the same data. In the second experiment, manual clustering was done for the initial model, using the leader algorithm with a likelihood threshold criterion on the number of states and mixtures. Finally, an experiment with manually trained durational HMM was done. The purpose of this experiment was to check the importance of state transitions and durations to speaker recognition, independently of the spectral distortion part of the model.

B. The Genuine and Impostors Likelihood Distributions

The performance of the system is completely determined by the distributions of the log likelihood per frame (2.13) for the genuine speakers and the impostors. These distributions are plotted in Fig. 3, for various test trial lengths (in terms of the number of digits spoken). As can be seen, both distributions are unimodal and slightly skewed. As expected, we observe a decrease of the variance of both distributions, with the number digits in the test n, and therefore a better separation of the two populations. We see, however, that the experimental standard deviation σ_n for both distributions, decreases slower than $1/\sqrt{n}$. This indicates that for that type of modeling, the utterances are not statistically independent. The means and standard deviations for the intraspeaker and interspeaker distributions for the 8 mixtures case are presented in Table I.

Note that the means of the distributions μ_n increase monotonically with n, unlike the prediction in (2.15). This is due to the

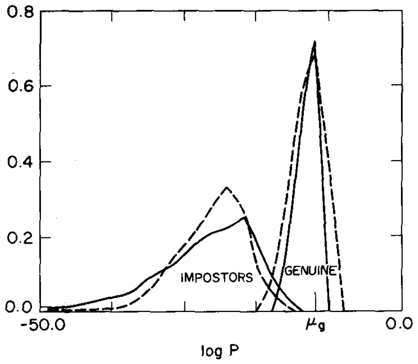


Fig. 3. Log likelihood per frame distributions over impostors and genuine speakers. The distributions are given for one (solid line) and two digits test trial.

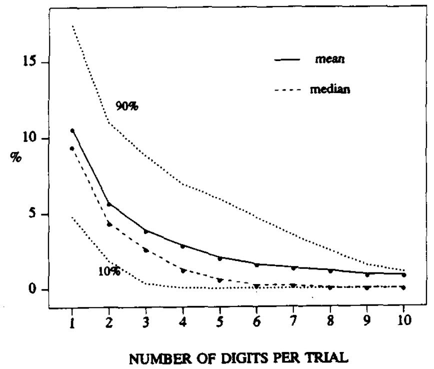


Fig. 4. Mean, median, ten, and ninety percentiles verification equal error rates, using AR-HMM with 8 states and 8 mixtures per states, as a function of the test trial length. The median error rate is lower than the mean, due to the existence of a few inconsistent speakers.

TABLE II
EQUAL ERROR RATES ON ALL TESTING DATA (%)

Test Length	Mean	10%	Median	90%
1	10.5	4.7	9.4	17.3
2	5.6	1.8	4.3	10.9
3	3.8	0.3	2.6	8.7
4	2.8	0.0	1.2	6.8
5	2.0	0.0	0.5	5.8
6	1.5	0.0	0.2	4.6
7	1.3	0.0	0.1	3.4
8	1.1	0.0	0.0	2.6
9	0.9	0.0	0.0	1.1
10	0.8	0.0	0.0	1.0

TABLE III
MEAN AND MEDIAN EQUAL ERROR RATES (%) FOR DIFFERENT MODEL SIZES

Mixtures:	2		4		8	
Length	Mean	Med.	Mean	Med.	Mean	Med.
1	15.8	14.7	11.6	11.1	9.6	8.0
2	10.9	9.8	7.0	5.5	5.1	4.0
3	8.9	8.8	5.0	3.1	3.7	2.4
4	6.5	6.2	4.4	2.9	3.3	0.7
5	5.0	4.0	3.2	0.7	2.5	0.2
6	4.8	2.9	3.0	0.4	2.4	0.2
7	3.9	1.6	2.6	0.3	2.0	0.1
8	3.5	0.7	2.4	0.2	1.8	0.0
9	2.5	0.5	2.3	0.0	1.7	0.0
10	2.4	0.2	2.3	0.0	1.6	0.0

fact that the short utterances fit the models better than the long ones and have higher likelihood per frame. This, however, has no effect on the error rate, which is related to the overlap between the distributions.

C. Error Rates

The equal error rates, as a function of the test trial length, are plotted in Fig. 4. We observed rapid improvement in the error rate when the test trial length grows up to about 7 digits. The experimental equal error rates (for the 8 states and 8 mixtures case), are summarized in Table II.

As always, with speaker verification, the results are not uniform, i.e., there are speakers for which the error rate remains above 20% for any utterance length (see Fig. 4). This can be seen by comparing the mean and median error rates. Most of the errors, for the longer test length, come from the worst 10% of the population. The existence of such "inconsistent" speakers is well known ("goats"). However, the question of who are those speakers, may be a function of the verification algorithm.

D. The Effect of the Number of Mixtures

The spectral resolution of the model is the total number of mixtures in the model. This parameter determines the overall distortion and the performance of the system. We compare the mean and median error rates for 8-state models, with 2, 4, and 8 mixtures per state. The mean and median error rates for the different number of mixtures are given in Table III (results on the first 20 speakers, which is the reason for the differences compared with Table II).

We can see that the mean error rates are roughly halved, when we go from 2 mixtures per state (16 vectors) to 8 mixtures per state (64 vectors). The smaller models are therefore more effective, in terms of the error probability per parameter. The "goats" are less affected by increasing the size of the model, so that the improvement in the median error rate is more significant.

E. Manual Training

Using the manual clustering method, slightly more efficient models (in terms of error rate per parameter) can be trained. The manual training process is, however, very tedious and is not practical for a large number of speakers. The results obtained with manual training of a 5-state model, with about 5 mixtures per state (25 vectors) are similar to those with the 8 states and 4 mixtures (32 vectors) trained automatically using VQ. Thus, it seems that the automatic clustering method using VQ is sufficently good for the initial model.

F. The Effect of the Recording Session

As with any other speaker recognition method, the model needs to be updated from time to time, or better, an adaptive training method should be used. This can be clearly seen by looking at the error rates at the different recording sessions. Since the models are trained on the first 2.5 sessions, the third, fourth, and fifth sessions are used for this purpose. The verification error rates with the best models (8 states and 8 mixtures), along the different recording sessions, are presented in Table IV.

TABLE IV
VERIFICATION EQUAL ERROR RATES (%) AS A FUNCTION OF THE RECORDING SESSION

	Session 3:		Session 4:		Session 4:	
Length	Mean	Med.	Mean	Med.	Mean	Med.
1	7.5	6.5	9.9	8.4	11.3	10.3
2	2.7	0.9	4.5	3.7	5.8	4.1
3	0.9	0.1	2.5	0.6	3.7	1.3
4	0.5	0.0	1.6	0.3	2.5	0.4
5	0.1	0.0	1.0	0.1	1.8	0.2
6	0.0	0.0	0.7	0.0	1.4	0.0
7	0.0	0.0	0.5	0.0	1.1	0.0
8	0.0	0.0	0.4	0.0	1.1	0.0
9	0.0	0.0	0.3	0.0	0.8	0.0
10	0.0	0.0	0.2	0.0	0.7	0.0

TABLE V
VQ EQUAL ERROR RATES (%) (SESSION 3)

Test Length	Mean	10%	Median	90%
1	9.1	3.4	8.2	15.7
2	3.8	0.1	1.7	9.4
3	1.8	0.0	0.4	6.9
4	1.0	0.0	0.2	2.6
5	0.5	0.0	0.0	1.4
6	0.3	0.0	0.0	0.7
7	0.2	0.0	0.0	0.9
8	0.1	0.0	0.0	0.1
9	0.1	0.0	0.0	0.2
10	0.1	0.0	0.0	0.3

G. The Importance of the State Transitions and Durations

The essential difference between the vector quantization approach and hidden Markov modeling is in the "dynamic elements" in the form of state transitions and durations. A comparison of the results obtained with hidden Markov models to the VQ results with the same spectral resolution (see Appendix I) shows that the effect of the transition probabilities is very small. The common problem with the conventional way of Markov modeling is that the likelihood function is dominated by the "static" part, i.e., the spectral distortion measure. The dynamic range of the spectral likelihood, is so much larger than that of the transition probabilities that the exact value of the latter becomes almost irrelevant. When the dynamic part is given only through state transitions, the main contribution to the speaker verification ability is from "forbidden transition," i.e., transitions that never occur in the training data. This, however, can be very specific for the particular vocabulary used (isolated digits) and thus not in the spirit of text independent verification. The "dynamic" part of the model becomes much more important when state durations are taken into account. In the previously described HMM, the probability of remaining in the state i for duration τ frames, is completely determined by the transition matrix

$$P[S_i \textit{ has duration } \tau] = a_{ii}^{\tau-1}(1 - a_{ii}). \tag{3.1}$$

A much better way to specify the durational information is through the conditional probability that state i occurs after state j and has a duration of exactly τ frames

$$d_{ij}(\tau) = P[q_{t+1} = S_i, \cdots, q_{t+\tau} = S_i; q_{t+\tau+1} \neq S_i | q_t = S_j]. \tag{3.2}$$

These distributions, in particular their means, vary significantly among speakers and turn out to be useful for speaker characterization. There is a well-developed way to incorporate such state durations within the hidden Markov model framework [14].

In order to examine the net effect of the dynamic part of the models, we carried out a simple speaker identification experiment, based only on state transitions and durations. This was done using the trained speaker models including the state transition duration densities (3.2), while ignoring the spectral (output densities) part of the likelihood function, during the recognition phase. By doing this we measure the discrimination ability of the temporal part of the model, the only information not contained in the VQ approach.

The detailed results of this experiment are given in Appendix II. These results indicate that some information on the speaker is contained in the state durations, but its resolving power is inferior to that of the spectral distortion part of the model. Clearly, this indicates that additional dynamical information, such as state conditional state durations, should improve the speaker verification ability.

IV. CONCLUSIONS

Autoregressive hidden Markov models, when trained properly, can be used for statistically characterizing speakers, in a text independent manner. The general trends found in the VQ based system, in particular the variability in performance from speaker to speaker and the need to update the model from time to time, were found again in this experiment. However, the improvement in performance over the simpler vector quantization approach does not justify the effort in training such models for the sole purpose of speaker recognition. On the other hand, the mathematical structure of the HMM may be used to generate better speaker models, by using different modeling principles and information theoretic techniques [15], [16]. The results encourage further research on the ways to factorize the text and the speaker within the model, and the role of the durational information. As in the VQ case, adaptive modeling is expected to improve the results, but the correct way of adaptation of an HMM is still an open problem.

APPENDIX I
THE VECTOR QUANTIZATION RESULTS

As a reference, all the vector quantization results using likelihood ratio distortion were rederived, using the same data and the same methods of calculating the error rates. The results are obtained with VQ of 64 vectors for all the 100 speakers. The differences from the results reported in [1] are mainly due to the fact that our results are based on the distortion per frame and not per utterance. Table V shows VQ equal error rates for session 3, Table VI for session 4, Table VII for session 5.

APPENDIX II
SPEAKER RECOGNITION USING STATE DURATIONS AND TRANSITIONS

Table VIII shows a speaker identification experiment, using state transitions and durations likelihood only, on a 10 digit sequence. The log likelihood per frame of the conditional durational densities matrix is presented. The maximal value in each row is on the diagonal element (100% identification).

Table IX shows the results of the same experiment, using the full HMM log likelihood scores.

TABLE VI
VQ Equal Error Rates (%) (Session 4)

Test Length	Mean	10%	Median	90%
1	10.0	4.3	8.0	18.5
2	4.9	0.4	4.0	10.4
3	2.8	0.0	0.6	7.2
4	1.7	0.0	0.3	6.6
5	1.1	0.0	0.0	3.1
6	0.8	0.0	0.0	1.4
7	0.6	0.0	0.0	1.1
8	0.5	0.0	0.0	0.6
9	0.4	0.0	0.0	0.7
10	0.4	0.0	0.0	0.3

TABLE VII
VQ Equal Error Rates (%) (Session 5)

Test Length	Mean	10%	Median	90%
1	11.6	5.0	10.2	18.9
2	6.5	0.7	4.8	12.3
3	3.8	0.1	1.3	9.4
4	2.7	0.0	0.7	7.6
5	1.8	0.0	0.2	3.5
6	1.4	0.0	0.0	2.6
7	1.2	0.0	0.0	1.9
8	1.2	0.0	0.0	1.4
9	0.8	0.0	0.0	0.8
10	0.7	0.0	0.0	0.6

TABLE VIII

	Test Speaker Number									
Model	1	2	3	4	5	6	7	8	9	10
1	−10.58	−16.39	−13.75	−16.12	−16.85	−14.63	−16.86	−14.75	−18.35	−16.90
2	−14.92	−10.52	−16.63	−12.32	−12.89	−15.82	−17.82	−16.82	−14.32	−14.65
3	−19.43	−15.23	−8.53	−15.58	−16.46	−15.86	−12.75	−19.83	−12.29	−16.93
4	−18.19	−15.21	−20.93	−9.88	−11.16	−10.61	−14.19	−13.45	−16.55	−10.29
5	−15.11	−14.24	−12.78	−14.48	−9.45	−18.32	−15.51	−14.78	−12.36	−14.70
6	−19.82	−18.00	−15.73	−14.61	−19.96	−14.39	−20.82	−17.34	−14.42	−21.52
7	−9.93	−15.88	−17.33	−15.25	−13.25	−12.85	−9.45	−12.04	−12.70	−11.23
8	−14.04	−12.21	−13.62	−14.35	−12.64	−13.36	−14.35	−9.55	−15.02	−14.80
9	−11.70	−14.32	−13.97	−13.81	−13.75	−13.76	−11.19	−18.58	−10.46	−11.54
10	−10.29	−13.73	−13.48	−13.78	−11.16	−13.79	−13.14	−13.46	−13.67	−9.46

TABLE IX

	Test Speaker Number									
Model	1	2	3	4	5	6	7	8	9	10
1	−11.07	−23.95	−20.13	−20.83	−25.73	−31.22	−26.97	−29.96	−19.15	−29.47
2	−19.00	−12.08	−26.30	−16.57	−27.52	−24.70	−30.27	−23.35	−24.80	−22.89
3	−18.68	−25.70	−13.60	−24.77	−21.84	−35.17	−22.70	−28.40	−16.73	−31.97
4	−21.85	−18.43	−26.70	−12.34	−29.65	−27.87	−28.47	−27.35	−24.70	−22.15
5	−18.69	−25.47	−20.64	−23.41	−14.10	−32.26	−25.58	−24.41	−20.11	−28.35
6	−20.53	−17.68	−27.86	−15.56	−26.74	−13.96	−28.49	−24.22	−26.11	−22.42
7	−20.47	−25.15	−18.53	−23.50	−21.36	−34.34	−13.01	−28.19	−17.30	−30.93
8	−19.16	−17.42	−24.53	−18.57	−23.73	−26.25	−28.86	−14.16	−21.25	−24.10
9	−18.86	−25.98	−17.75	−21.64	−23.49	−35.95	−22.02	−28.71	−12.78	−31.79
10	−18.69	−18.31	−25.04	−15.68	−24.87	−24.04	−26.36	−24.00	−21.55	−14.87

ACKNOWLEDGMENT

The author is grateful to F. Juang, F. Soong, and M. Sondhi for their software. A great debt is owed to L. Rabiner, for pointing out the right way of using hidden Markov models, and to F. Soong and A. Rosenberg for critically reading the original manuscript and making very important comments. Useful remarks by S. Levinson, F. Juang, Y. Ephraim, and D. Mansour are gratefully acknowledged.

REFERENCES

[1] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, "A vector quantization approach to speaker recognition," in *Proc. ICASSP 85, IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1985, pp. 387-390; also *AT&T Tech. J.*, vol. 66, no. 2, pp. 14-26, 1987.

[2] A. E. Rosenberg and F. K. Soong, in *Proc. ICASSP 86, IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1986, pp. 873-876.

[3] Alan B. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proc. ICASSP 82* (Paris, France), May 1982, pp. 1291-1294.

[4] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[5] B.-H. Juang and L. R. Rabiner, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 6, p. 1404, 1985.

[6] B. H. Juang, S. E. Levinson, and M. Sondhi, "Maximum likelihood estimation for multivariate mixture observation of Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-32, no. 2, pp. 307-309, 1986.

[7] B.-H. Juang, *AT&T Tech. J.*, vol. 63, no. 7, pp. 1213-1243, 1984.

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

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, vol. 39, no. 1, pp. 1-38, 1977.

[10] R. M. Gray, "Vector quantization," *IEEE ASSP Mag*, pp. 4-29, 1984.

[11] J. A. Hartigan, *Clustering Algorithms*. New York: Wiley, 1975.

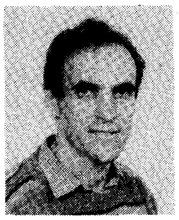
[12] L. R. Rabiner, J. G. Wilpon, and B. H. Juang, "A segmental k-means training procedure for connected word recognition," *AT&T Tech. J.*, vol. 65, no. 3, pp. 21-31, 1986.

[13] G. D. Forney, "The Viterbi algorithm," *Proc. IEEE*, vol. 16, pp. 268-278, 1973.

[14] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Comput. Speech, Language*, vol. 1, pp. 29-45, 1986.

[15] Y. Ephraim, A. Dembo, and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," in *ICASSP 87, IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1987, pp. 25-28.

[16] N. Tishby, "Speech modeling with prior information and its application to speaker recognition," in *ICASSP 1988, IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1988, pp. 87-90.



Naftali Z. Tishby was born in Jerusalem, Israel. He received the B.Sc. degree (*cum laude*) in physics and mathematics from the Hebrew University of Jerusalem in 1974, the M.Sc. degree (*cum laude*) in physics from Tel-Aviv University in 1980, and the Ph.D. degree in theoretical physics from the Hebrew University in 1985.

From 1974 to 1980 he was with the Israel Defense Forces (IDF) where he established and headed a research group in signal and speech processing. During 1984-1985 he served as a Vice President of Research in Sesame Systems Ltd., developing speech and speaker recognition systems, and in 1985-1986 he was a postdoctoral fellow at the Massachusetts Institute of Technology, working on chaotic Hamiltonian dynamics. Since 1987 he has been a Member of the Technical Staff (Information Principles Laboratory) at AT&T Bell Laboratories, Murray Hill, NJ. His current research focuses on nonlinear dynamics and its applications to speech processing, stochastic processes, learning theory, and statistical mechanics of neural networks.

Dr. Tishby received the Eliyahu Golomb Israel Security Award in 1980 and the Chaim Weizmann fellowship in physics in 1985.