Speech Recognition Using Randomized Relational Decision Trees

Yali Amit and Alejandro Murua

***Abstract*—We explore the possibility of recognizing speech signals using a large collection of coarse *acoustic events*, which describe temporal relations between a small number of local features of the spectrogram. The major issue of invariance to changes in duration of speech signal events is addressed by defining temporal relations in a rather coarse manner, allowing for a large degree of slack. The approach is greedy in that it does not offer an "explanation" of the entire signal as the hidden Markov models (HMMs) approach does; rather, it accesses small amounts of relational information to determine a speech unit or class. This implies that we recognize words as units, without recognizing their subcomponents. Multiple randomized decision trees are used to access the large pool of acoustic events in a systematic manner and are aggregated to produce the classifier.**

***Index Terms*—Classification, decision trees, labeled graphs, spectogram, speech recognition.**

I. Introduction

EXPERT human "observers" of bioacoustic signals, such as speech and bird songs, visualize the information carried in the acoustic waveform in two-dimensional images containing the time-frequency dynamics of the utterances. These images correspond to spectrograms or Log-spectrograms of the signals. The expert human observer (e.g., a phonetician) is somehow able to learn from a collection of spectrograms, acoustic invariants associated to specific units of vocalization (e.g., phonemes in speech, syllables in songs). These acoustic invariants allow the expert to identify the vocalizations present in the signals. Moreover, when learning to identify different vocalizations, human experts seem to focus on the global shape of the spectrograms, i.e., in the *temporal* relations among several local features in both time and frequency. In fact, properties of the gross shape of the spectrum such as the relation among energies at frequency peaks, and the change in energy distribution over time, are postulated to contain acoustic invariants for certain phonetic features of speech [8, p. 188].

In this paper, we attempt to address speech recognition from this point of view. In other words, we explore the possibility of recognizing speech signals using a large collection of coarse *acoustic events*, which describe temporal relations between a small number of local features of the spectrogram. The major issue of invariance to changes in duration of speech signal events is addressed by defining temporal relations in a rather coarse manner, allowing for a large degree of slack. The approach is greedy in that it does not offer an "explanation" of the entire signal as the hidden Markov models (HMMs) approach does; rather, it accesses small amounts of relational information to determine a speech unit or class. This, of course, implies that we recognize words as units, without recognizing their subcomponents.

This approach connects directly to ideas investigated in previous works on object and shape recognition; see [2] and [3]. There, discrimination is obtained through coarse global arrangements of local image tags in the plane. The tags are very stable in the sense that they occur with high probability in certain parts of the shape, even under rather severe deformations. The *spatial* relations among the image tags are defined in a very coarse manner in order to accommodate the required invariance to shape deformations which preserve shape class. These global arrangements provide a very rich family of representations of the gross shape of the different objects, and provide the tools for recursive relational quantization of the space of objects using multiple decision trees.

The basic ingredients are the following. A collection of local tags is defined together with a family of simple pairwise relations between them. For images the local tags are defined in terms of the data in a small neighborhood. An example could be oriented edge information. In acoustic data the local tags are defined in terms of the data in a small time/frequency interval. They are binary variables which detect the presence of a certain frequency in a certain range of energies. The ranges are determined after normalization so some degree of invariance to amplitude is obtained; they are moderate in size to accommodate for invariance to audio quality. Relations are specified by coarse constraints on locations between the tags. For example in image data the second tag could be constrained to lie in a wedge of some angle with respect to the first. In acoustic data the second tag could be constrained to lie within a certain interval of time relative to the first (e.g., between 100 to 300 ms after the occurrence of the first). Each arrangement is either present in the data or not, and hence defines a binary variable on the data.

An arrangement of tags is a labeled graph: each vertex of the graph corresponds to some tag type, and each edge, to some relation between the two vertices it connects. As labeled graphs, the tag arrangements have a natural partial ordering: each graph precedes any of its direct extensions, involving an additional tag and a relation. Proceeding along one of the paths of this ordering leads to a recursive relational partitioning of the sample space.

Manuscript received May 4, 1999; revised May 23, 2000. Y. Amit was supported in part by the Army Research Office under Grant DAAH04-96-1-0061 and MURI grant DAAH04-96-1-0445. A. Murua was supported in part by the University of Chicago Block Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. John Deller.

Y. Amit is with the Department of Statistics, University of Chicago, Chicago, IL 60637 USA.

A. Murua is with the Department of Statistics, University of Washington, Seattle, WA 98195-4322 USA (e-mail: murua@stat.washington.edu).

Publisher Item Identifier S 1063-6676(01)01327-X.

Even using a small number of tag types and relations the total number of arrangements (graphs) of say ten tags is huge, in particular since we are not imposing any particular ordering on the arrangements as in classical HMMs or dynamic time warping models. Although this rich family of binary variables may contain an enormous amount of information about the class of the data, it can never be computed in its entirety even for a single data point, let alone for a large training set. These variables can only be accessed incrementally in some order. Decision trees offer a very natural way to systematically explore the partial ordering of the arrangements: more complex arrangements are used as splitting rules as the tree grows deeper. The informative features are found at the same time the sample space is being recursively partitioned or quantized. The vast number of arrangements also allows us to use randomization to produce multiple decision trees which are conditionally weakly dependent, and which can be aggregated to produce powerful classifiers [2], [10].

In the context of bird song data, we were able to use these ideas to process both segmented and continuous data, and have achieved recognition rates similar to those of HMMs at a significant gain in computational cost, both in training and testing (see Section VI). For spoken digit data, we have achieved a correct recognition rate of 98.6% on the segmented TI/NIST data which, as we will show in Section V, is better than the rate achievable with HMMs, given the same amount of training data. Another advantage of our approach is the ability to visually interpret the outcome of the trees. See for example Fig. 6, and the related discussion in Section IV.

We do not, however, directly address the issue of segmentation and continuous speech as HMMs do, and we are still investigating ways to incorporate this classifier to analyze continuous speech. One encouraging point is the fact that the trees based on these loose relational arrangements are very robust to significant error in segmentation (see Section V). Also, it should be emphasized that the specific tags we have chosen may not be the optimal ones for the definition of the acoustic events. Ideas such as those presented in [11] and [12] may lead to tags with more information content, and a higher degree of invariance.

This paper is organized as follows. In Section II, we give a precise definition of the local tags we use; the relations between them are introduced in Section III. In Section IV, we describe the randomized tree growing procedure. In Section V, we present experimental results, on both the TI/NIST dataset and the CSLU Number Corpus dataset, and compare to the performance of HMMs using the commercial package [17]. We also outline a simple nearest neighbor method for boosting the classification rates. Finally, in Section VI, we argue why decision trees offer a gain in computational cost relative to HMMs.

II. Local Tags

The acoustic signal can be represented in several forms. A common goal of these representations is to make the time signal more amenable to further processing, thus entailing some kind of data-reduction and smoothing of the original signal.

Probably the simplest representation of speech is the *spectrogram*. This three-dimensional representation of the acoustic signal describes the frequency dynamics of an utterance over time. It offers a good visualization of the energy content of the frequencies. This is the representation we used in our experiments, but it is not the only possible choice (e.g., a wavelet transformation of the signal, such as the *waveletogram* in [6] could be used as well).

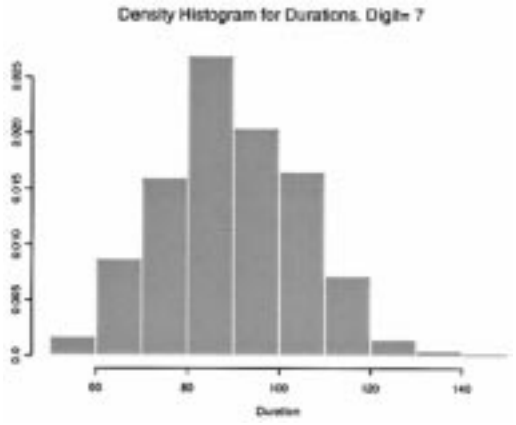

Fig. 1. Density histogram for durations for the digit *seven* in the TI-digits corpus. Duration is given in frame numbers; frames are overlapping and taken each 5 ms; frame window-size is 25.6 ms.

A. Spectrogram

The spectrogram can be thought of as a grayscale image, whose pixel intensities represent the energy content of the frequencies over time.

Our spectrograms are the output of a smoothing procedure over the time-frequency domain. The time axis is divided in consecutive overlapping frames. Within each frame the signal is weighted with a Hamming window; the resulting signal is then Fourier-transformed, giving rise to a vector of frequency energies (the frequency axis). This vector represents an estimate of the spectrum associated to the signal at this particular time frame. The resulting spectrogram can also be thought of as a matrix (X_{tf}), $t = 1, \cdots, T$, $f = 1, \cdots, F$, where t is the time in frames, and f, the frequency bin. T measures the duration or length of the utterances; it varies from utterance to utterance. A simple statistical analysis shows that T approximately follows a Poisson distribution within a specific speech unit (see Fig. 1); this observation will be used later in our experiments in Section V. F is the total number of frequency bins considered; it is kept fixed for all utterances (in our experiments, we set $F = 18$ or 14, according to the sampling rate at which the data were recorded). We use overlapping frames of length 25.6 or 32 ms (depending on the sampling rate at which the data were recorded), taken each 5 ms.

B. Frequency and Energy Quantization

First, in order to ensure invariance to global changes in amplitude we normalize spectrograms by their mean over all times and frequencies, i.e., we take $X_{tf}/\overline{X}_{tf}$. We believe that the information contained in the spectrogram is very redundant due, in part, to high local correlations, and that the energy content X_{tf} can be quantized rather coarsely in such a way so as to keep the relevant information for recognition intact. We experiment

with five different types of frequency binning: Bark scale with 18 frequency bins (B18) [18], Mel frequency spectrogram [17, ch. 5], with 18 frequency bins (M18), and 12, 21, and 36 uniform bins over the frequency range of 50 to 5400 Hz, denoted U12, U21, and U36, respectively.

The energy is either quantized uniformly into $2Q+1$ levels or using a log-normal model for the energy levels. In the latter case we use bins defined by an estimated standard deviation σ of the log-energy obtained from training data. Specifically, given σ, and the mean log-energy μ, we define the q'th quantile as

$$\tfrac{1}{2}(q-Q)\sigma \le \log X_{tf} - \mu < \tfrac{1}{2}(q-Q+1)\sigma$$

for $q = 0, 1, 2, \cdots, 2Q-1$, and the $2Q'$th quantile as

$$\log X_{tf} - \mu \ge \tfrac{1}{2}Q\sigma.$$

The idea here is that it may be useful to use a nonuniform quantization of the energy content, which is partial toward high values of energy content, and that at the same time allows for certain degree of slack on moderate values of the energy content. Notice that the probability of the quantiles is not constant; in fact, for $Q = 4$, the vector of probabilities $\{p_q\}_{q=0}^{8}$ is

$$\{0.044, 0.092, 0.150, 0.191, 0.191, 0.150, 0.092, 0.044, 0.023\}.$$

Also notice that small values of frequency energy are discarded from further processing. Note that few tags associated to high values of energy content are expected to be observed with this quantization scheme, and hence acoustic events involving these tags are rare and therefore likely to reveal important patterns for discrimination among different speech units.

The main conclusion is that the form of frequency binning and energy quantization does not have a major effect on the outcome of the procedure. The results are summarized in Table IV.

All of these quantization schemes differ from the usual vector quantization technique used with HMMs. In this latter framework, the whole F-dimensional vector space is quantized in about 10^3 regions. In our approach quantization is local in the frequency domain: each component X_{tf} is quantized separately. The number of resulting vector quantization regions is approximately $(2Q+1)^F$, which for moderate Q (e.g., $Q = 4, F = 18$), is huge. We never explicitly use multifrequency quantiles so this never creates a problem. Each of the $2Q+1$ levels at each of the F frequency bins is a "tag" (Fn, Qm), labeled by the energy quantile Qm, $m = 0, 1, \cdots, 2Q$, and the frequency bin Fn, $n = 1, \cdots, F$. It is a binary feature which is either present or not. Important information regarding co-occurrences of certain frequencies at the same time are represented through the *relations* between the tags, see Section III below.

In cases where the same tag occurs in consecutive time frames we cluster to one tag at the first time of occurrence. The maximal duration of clustering is typically five time frames. In Figs. 4–6, we show the locations of the tags.

III. Acoustic Events

Our procedure is based on the assumption that there are acoustic events that either occur fairly often or rarely on most utterances representing a determined speech unit (class). The presence or absence of several of these acoustic events in a given utterance, gives strong hints for the identification of the speech unit represented by the utterance.

Statistically speaking, the probability of observing the presence or absence of several of these acoustic events is high, given that an utterance is a realization of certain determined speech unit; at the same time, this same probability is fairly small over all utterances (regardless of the speech unit they represent).

A. Relations and Labeled Graphs

We consider particular acoustic events defined by a collection of binary relations between tags in a spectrogram. We note that tags only carry information on energy content at particular frequency bins; hence, tags alone are too primitive features to be relevant for recognition. A tag ℓ_1 is related to another tag ℓ_2 by the time interval I, if the relative time between their occurrences in the spectrogram, $t(\ell_2) - t(\ell_1)$, is contained in the time interval I, i.e., $t(\ell_2) - t(\ell_1) \in I$.

An acoustic event is a *connected* graph. The vertices of the graph are labeled by tag types. The edges between pairs of these vertices are labeled by time intervals defining their relationship. To be precise, let $V = \{\ell_1, \ell_2, \cdots, \ell_k\}$, be a list of tags, and

$$E = \{(\ell_{i,1}, \ell_{i,2}, I_i), i = 1 \cdots, n, \ell_{i,j} \in V, j = 1, 2\}$$

be a list of ordered pairs of tags with a time interval. The associated acoustic event is

$$t(\ell_{i,2}) - t(\ell_{i,1}) \in I_i, \qquad i = 1, \cdots, n. \tag{1}$$

The value of n denotes the *depth* of the event. The condition that the graph is connected implies that there is a path of edges between any two vertices. In other words, for any ℓ_a, ℓ_b in V there is an integer m and a sequence of edges $(\ell_{i_j,1}, \ell_{i_j,2})$, $j = 1, \cdots, m$, in E such that $\ell_{i_1,1} = \ell_a$, $\ell_{i_m,2} = b$ and $\ell_{i_j,2} = \ell_{i_{j+1},1}$ for $j = 1, \cdots, m-1$

Temporal relations such as the ones given by (1), allow for certain slack in the timing of events on individual utterances; in this way, the sensitivity to the time-warping problem is controlled. We observed in our experiments, that a few nonoverlapping intervals suffice in order to obtain good classification rates. We have used the following five intervals (in time frames): (0, 20), (20, 40), (40, 70), (70, 100), (100, $+\infty$).

Note that on top of the time warping allowed for by the definition of the relations, the acoustic event corresponding to the graph is entirely translation invariant. One can think of particular realizations of these acoustic events as being aligned (warped) to ideal "templates" of the events; these templates correspond to our labeled graphs.

The top panels of Figs. 4 and 5 show one instance of such a graph on four different spectrograms of four different digit utterances. The variability in the possible instantiations of this graph is clearly manifest in the four images.

B. Information Content

One might ask how informative these acoustic events are. For example the distribution on class for those digits containing the acoustic event shown in the top panels of Figs. 4 and 5 is given

TABLE I
FIRST COLUMN: TOTAL PROBABILITY OF EVENT. SECOND COLUMN: CONDITIONAL ENTROPY ON CLASS GIVEN THE EVENT (BASE e). COLUMNS 3–13: PROBABILITIES ON CLASS GIVEN THE PRESENCE OF THE ACOUSTIC EVENT. FOR EACH EVENT WE SHOW THE NUMBERS ON THE TRAINING (tr) AND ON TEST (te) SETS. THE EVENTS CORRESPOND TO THE TOP TWO AND BOTTOM TWO PANELS OF FIGS. 4, 5, AND TO NODES "NO, YES, YES" AND "NO, YES, YES, YES, YES" IN THE TREE OF FIG. 2

	Prob	Ent	0	1	2	3	4	5	6	7	8	9	10
Ev. 1 (tr)	.08	1.33	.04	.05	0	.04	.14	0	.61	.07	.01	0	0.02
Ev. 1 (te)	.09	1.27	.05	.04	0	.02	.16	.01	.59	.13	0	0	0
Ev. 2 (tr)	.007	.77	0	0	0	.06	.75	0	.15	.03	0	0	0
Ev. 2 (te)	.007	.55	.11	0	0	.05	.83	0	0	0	0	0	0

in Table I for both the training set (4460) points and the test set (2486) points.

One immediately sees that even an event involving three tags and two relations may be very informative, and an additional two tags narrows the distribution effectively to only two classes. The entropy of the prior distribution on class is 2.4 (base e). Note also the strong similarity in the shape of the distributions between test and training even though the conditioning events are rather low probability events. This is an indirect indication that a large degree of invariance is accommodated through the relations and the tag definitions.

It is of interest to note the difference between the use of graphs representing the geometric arrangements in this context, and the use of random graphs for structural synthesis as in [15]. The idea there is that an instance of a random graph can be initially attached to every data point in the training set. The question is how to hierarchically *cluster* these graphs using an entropy criterion which determines a distance between them. In the current setting the graphs are used to hierarchically *split* the data, and there is no way to *a priori* attach a graph to a data point.

C. Partial Ordering

The number of distinct tags is $(2Q+1)\times F$, and the number of possible binary relations is $N_I\times((2Q+1)F((2Q+1)F-1))/2$, where N_I is the number of temporal intervals considered. Consequently, the total number of acoustic events of depth n with exactly k distinct tags is immense (e.g., if $n = k = 5$, and $N_I = 5, Q = 4$, and $F = 18$, on the order of 10^{19}). We therefore need an efficient procedure to explore the pool of acoustic events, with the goal of detecting those relevant for classification. Our approach is constructive, in the sense that events are elucidated vertex by vertex in a suboptimal fashion: an event of depth n is deepened to $n+1$ if the addition of a new vertex and edge, significantly improve the discrimination of the speech units in consideration. In other words, we exploit the partial ordering on the acoustic events inherited from the graphical descriptions. This is done using decision trees as described in the next section.

IV. DECISION TREES

The collection of possible arrangements is vast and cannot be pre-computed as a binary feature vector. Moreover many arrangements may be useless in terms of the classification problem at hand. Decision trees are used to systematically explore the collection of arrangements and find the most informative ones in terms of classification.

The trees are constructed as follows. At the root of the tree, a search through the collection $\mathcal{G}_0$ of graphs of two vertices is done. Each such graph corresponds to a binary query on the data: either an instance of the graph is present in the data ($G_0 = 1$) or not ($G_0 = 0$). Therefore this query produces a split in the training data. Let Y denote the class label of a data point, and assuming K classes, the conditional entropy on class given the query is

$$\begin{aligned} H(Y|G_0) = {} & P(G_0=1)\sum_{k=1}^{K} P(Y=k|G_0=1) \\ & \cdot \log P(Y=k|G_0=1) \\ & + P(G_0=0)\sum_{k=1}^{K} P(Y=k|G_0=0) \\ & \cdot \log P(Y=k|G_0=0). \end{aligned}$$

where P corresponds to the empirical distribution defined by the training data. The graph G_0 yielding the smallest conditional entropy on class is chosen. This is the standard *entropy reduction* splitting criterion for tree growing known as CART in the statistics literature (see [4]) and as ID3 and C4.5 in the machine learning literature (see [13]). Data points for which $G_0 = 0$ go to the "no" child node. In that node a new search through $\mathcal{G}_0$ is performed in order to find the best split, i.e., the one with lower conditional entropy on class using the empirical distribution defined by the training data at the "no" node. Data points for which $G_0 = 1$ go to the "yes" child node, and have one or more instances of the graph G_0, which is now called the pending graph. A search among minimal extensions of this graph is performed to choose the one which leads to the greatest reduction in conditional entropy on class, using the empirical distribution in the "yes" node. Minimal extensions involve an additional tag in relation to one of the tags in the existing graph. The "yes" and "no" nodes are subsequently split into children nodes and the procedure continues. Note that other criteria for choosing the optimal split are possible. The effect on classification rates is negligible, see [4]. Essentially one is attempting to increase the 'purity' of the nodes, i.e., make them more and more concentrated on one class.

As the tree is growing, there is a pending graph at each node, determined by all the "yes" answers on the path leading from the root to that node. The only possible splits entertained at the node are minimal extensions of this pending graph. For example, the graph of panel two in Fig. 5 is the pending graph at node 011 (No, Yes, Yes) and also at node 0110 (No, Yes, Yes, No). At node 01101, it is extended to the graph of panel four. Tree growing is stopped when not enough data is present at the node to determine a split, or when none of the possible splits yields a significant decrease in entropy. In Fig. 2, we show part of a decision tree traversed by the four data points of Figs. 4–6.

Finally, in Fig. 6 we show five instances of the word "one," that landed at the same terminal node of a tree, together with an instance of the acoustic event associated to that node. The event can be interpreted visually as capturing a trend of an increase in energy at a rather low frequency, as well as a faster increase in a "diagonal" direction of both time and frequency. In all five spectrograms, one sees a trend of the midrange energies (blue and green), moving from midrange frequencies to

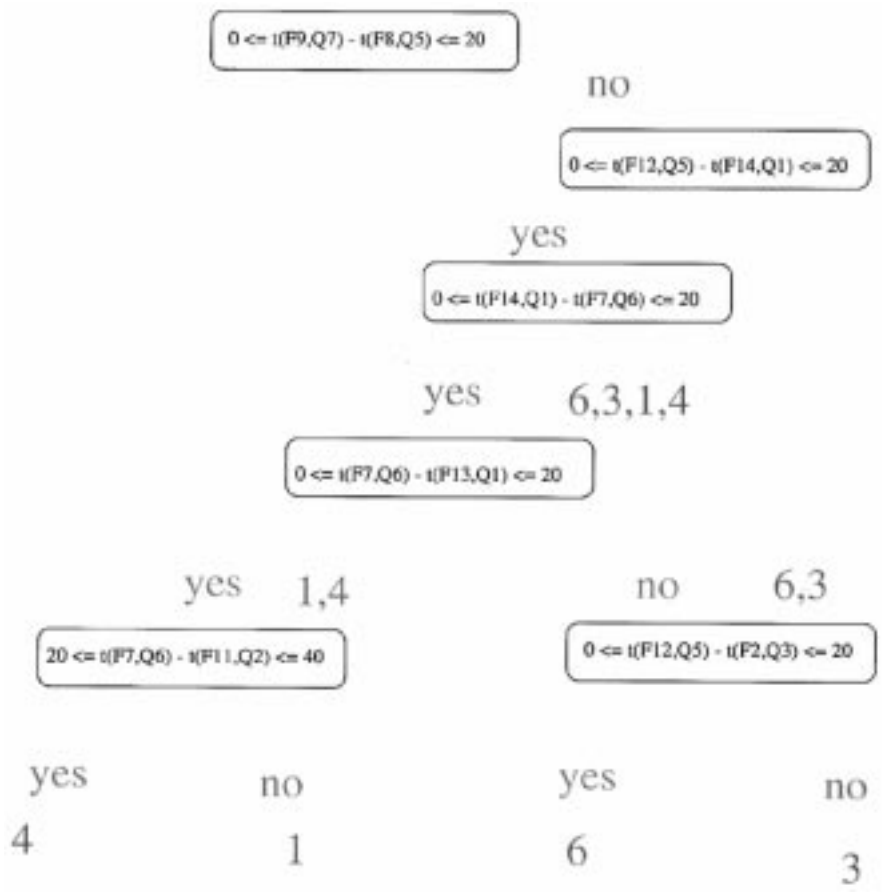

Fig. 2. Part of the decision tree traversed by the four data points of Figs. 4 and 5. The split used at each node is provided explicitly in terms of the tag types and the time interval.

lower frequencies. The low-energy part (red) is shrinking, and the high-energy part (yellow) is expanding. The graph is capturing this trend.

A. Multiple Trees

The size of the training set imposes a limit on the depth of a tree and only a very small number of tag arrangements is actually used. More information is accessed by growing randomized multiple trees, where instead of choosing the optimal split among *all* admissible splits (namely minimal extensions), one takes a small random sample of admissible splits and chooses the best among those. This yields somewhat less powerful trees, but the trees are substantially different, and offer complementary "points of view" on the data. Assume N trees are grown $T_1, \cdots, T_N$.

A test data point is dropped down each tree by evaluating the top query Q_0; if the answer is 1 (Yes), the point proceeds to the "yes" node, and the query there is evaluated, and so on until the data point reaches a terminal node. With a slight abuse of notation, we will denote the terminal node reached by a point ω in tree n as $T_n(\omega)$. Consider K classes (speech units). Associated to each terminal node t of a tree, there is a (conditional) probability distribution over the classes, $\mu_t = (\mu_t(1), \cdots, \mu_t(K))$, which has been estimated from the training data. This is called the terminal distribution. When a test point is dropped down N trees, it encounters N such terminal distributions $\mu_{T_1(\omega)}, \cdots, \mu_{T_N(\omega)}$.

The simplest way of aggregating the information in a collection of N trees is to calculate the average distribution

$$\mu(\omega) = \frac{1}{N} \sum_{n=1}^{N} \mu_{T_n(\omega)} = (\mu(\omega, 1), \cdots, \mu(\omega, K)).$$

The *argmax* of this average distribution is taken to be the classification assigned to the point ω

$$C(\omega) = \underset{k=1,\cdots,K}{\text{argmax}} \{\mu(\omega, k)\}.$$

This classification rule will be referred to as the *Aggregate Distribution Rule* in the following section.

Earlier we indicated that the randomized trees access the data from different points of view. Statistically this translates into the trees being *pairwise weakly dependent* conditional on class. This weak dependence leads to a spectacular increase in classification rates when more and more trees are aggregated. A detailed discussion of the properties of this classification rule and of the multiple randomized tree procedure can be found in [1] and [2].

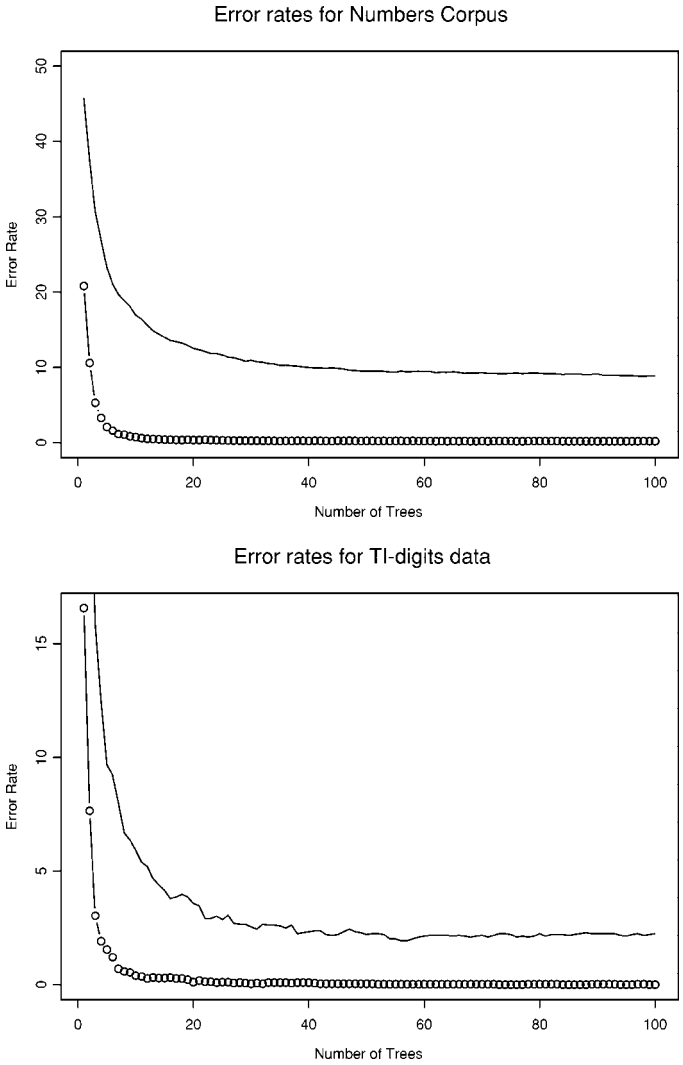

Fig. 3. Error rates for increasing number of trees. The solid line corresponds to the testing error rates, and the other line, to the training error rates.

V. Experiments

We applied our procedure to the recognition of isolated or segmented digits (*one*, *two*, *three*, …, *nine*, and *oh*). Our purpose was to find out how well our procedure does over a range of audio quality conditions. Hence, the data consisted of two speech corpora: one containing speech spoken over the telephone, and another one containing studio quality speech.

The telephone data were taken from the CSLU Number Corpus of the Center for Spoken Language Understanding of the Oregon Graduate Institute of Science and Technology. This corpus is a real world application containing "fluent numbers" spoken by thousands of people when saying numbers such as their street address numbers, zip codes, and telephone numbers. False starts, repetition, and background noise are very common in these data, and make the task difficult (see [5] for details). The corpus is divided into two sets of 8829 and 6171 speech files; the first one is reserved for training, and the second one for testing. We located and worked with all occurrences of the 11 digits in this corpus.

The studio-quality speech data were taken from the well-known TI/NIST Connected-Digits Recognition task (also known as TI-digits). This corpus does not include the segmentation of the utterances; hence we hand-segmented a small portion of them; specifically, we hand-segmented those

Fig. 4. 1) Graph on a digit "4" at node 011 in the tree used in Table I. 2) Graph on a digit "1" at same node. 3) Graph on the same digit "4'" at node 01111 in the same tree. 4) Graph on same digit "1" at node 01110 in the same tree. Each colored pixel denotes a tag (Fn, Qm), $n = 1, \cdots, 18$, $m = 0, \cdots, 8$. The frequency channel associated to a tag is read by the row number (from top to bottom) in the spectrogram. The energy quantile is coded by four colors. We have paired consecutive quantiles due to graphical constraints. Red: Q0,Q1, blue: Q2,Q3, green: Q4,Q5, yellow: Q6,Q7, cyan: Q8.

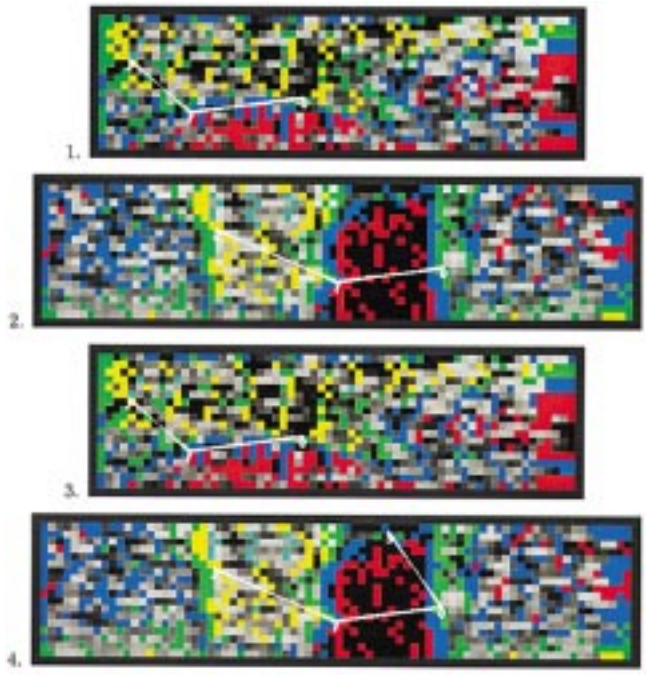

Fig. 5. 1) Graph on a digit "3" at node 011 in the tree used in Table I. 2) Graph on a digit "6" at same node. 3) Graph on the same digit "3" at node 01100 (same graph) in the same tree. 4) Graph on same digit "6" at node 01101 in the same tree. The color coding is the same as the one in Fig. 4.

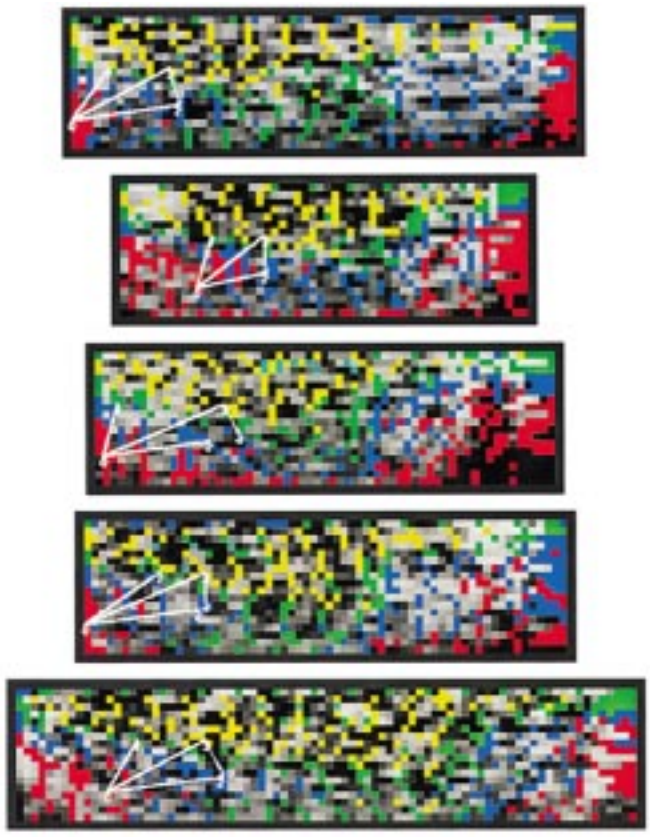

Fig. 6. Graph on five different "1" digits at node 0011011 in a tree. The color coding is the same as the one in Fig. 4. It appears that the acoustic event described in this node is identifying a certain pattern which coarsely represents the fact that the energy is increasing in time at a rather low frequency, but it is increasing faster in a "diagonal" direction of time and frequency.

utterances corresponding to digit sequences of at most two digits. Due to this limitation, our training data constituted a very small subset of the data available—4460 points in all. Therefore, our results are not directly comparable to previous results reported in the literature. Moreover, our testing data only consisted of 2486 isolated digits taken from the testing portion of the corpus.

The first column in Table II shows classification error rates for both corpora, using 100 trees, the Mel frequency spectrogram and log-normal quantization (see Section II-B). The second column shows the results with the additional boosting described below.

In general, recognition improves with the number of trees, but there appears to be a limit to the achievable error rates, as the asymptotes in Fig. 3 seem to indicate. Theoretical error rates can be obtained [2], [10] for these type of classifiers; these bounds depend on the average amount of dependency among the decision trees; it is conceivable that in practice due to the limited number of training data, it is impossible to create too many decision trees without building moderate correlation among them.

For comparison we experimented with HMMs using the exact same training and test data. We used a commercial HMM software [17]. The results are shown in Table III. It appears that using the same size training set, trees outperform HMMs. Nonetheless reported error rates for the TI-digits are much smaller than ours (less than 1%), of course making use of larger training sets.

TABLE II
CLASSIFICATION ERROR RATES USING 100 TREES WITH MEL FREQUENCY SPECTROGRAM AND LOG-NORMAL QUANTIZATION

Corpus	ADR	ADR+NN
TI-digits	1.41	1.41
Number	6.62	5.96

TABLE III
CLASSIFICATION ERROR RATES USING HMMS WITH THE [17] PACKAGE, AND GAUSSIAN MIXTURES. HMM-c STANDS FOR FOUR MIXTURES FOR "ONE," "TWO," AND "OH" AND SIX MIXTURES FOR THE REST. IT WAS NOT POSSIBLE TO USE MORE MIXTURES GIVEN THE SIZE OF THE TRAINING SET. COMPARABLE RESULTS ARE OBTAINED WITH 11 TREES. FOR RESULTS WITH 100 TREES SEE TABLE II

HMM 2 mixtures	HMM 4 mixtures	HMM c	11 trees
5.39	3.46	2.98	2.86

Finally, we show some error rates on trees using the different binning and quantization schemes previously mentioned in Section II-B. The main conclusion is that the outcome is not sensitive to the choice of quantization, except for U12 which seems to perform poorly.

A. Boosting the Classification Rates

We implemented a K-nearest neighbors (K-NN) rule over the space of aggregate distributions resulting from dropping data over the decision trees. The goal was to boost the classification rates with the hope of capturing useful information for recognition from the aggregate average distributions. In fact, we observed that the true digit (class) is among the top two modes of the aggregate average distributions on 96.8% of the data from the Number corpus, and on 99.7% of the data from the TI-digits corpus. This boosting procedure is based on the work in [14]; it consists of selecting a moderate size rejection set from the training data, so as to view the corresponding output aggregate distributions as centroids or prototypes of data that are likely to be rejected, namely data which the aggregate classifier is having trouble classifying. The rejection criterion is based on the ratio between the top two modes of the aggregate average distributions: if the ratio is smaller than certain *a priori* fixed threshold, then the data point is not recognized, but rejected, instead. About 25% of the training data is rejected in this way. This procedure is also applied to test data, but with a lower threshold, so as to reject at most 10% of the data. Each rejected data point in the test set is matched to its K-nearest neighbors in the training rejection set, according to the Kullback-Leibler distance between the two aggregate distributions. The data point is then recognized as a realization of the most frequent digit among its K-nearest neighbors. The column named ADR+NN of Table II shows the error rates for both corpora after applying the boosting procedure. We obtain an 10% reduction in the error rate for the Number corpus, and observe no improvement in the correct recognition rate for the TI-digits corpus. This indicates that this boosting procedure is more effective on data that are very difficult to discriminate, and hence it might work well on real world speech tasks.

TABLE IV
COMPARISON OF ERROR RATES FOR DIFFERENT FREQUENCY BINNING SCHEMES AND ENERGY QUANTIZATION SCHEMES

Frequency binning	Quantization scheme	
	Uniform	Log-normal
M18	2.13	1.57
B18	2.45	2.17
U12	4.39	4.10
U21	2.33	2.05
U36	2.66	2.29

TABLE V
CROSS-CLASSIFICATION ERROR RATES

Cross-Tested Corpus	ADR
TI-digits	13.50
Number	21.50

B. Cross-Testing

In order to assess how well our procedure generalizes to data recorded under different quality conditions, we cross-tested the testing portion of the corpora, i.e., the TI-digits were tested with trees trained with the Number corpus, and vice-versa. We observed that correct recognition rates decreased by slightly more than 10% when testing on different audio quality data. But they are still high enough to suggest that somehow our procedure is capturing acoustic events that are invariant across data sets. For comparison purposes, the testing portion of the TI-digits data set (2486 isolated digits) was cross-tested on HMMs trained on the Number Corpus, yielding an error rate of 12.8%, which is a very similar rate to the one obtained with our procedure (see Table V).

C. Sensitivity to Segmentation

To investigate how sensitive our procedure is to erroneous segmentations of word boundaries, we tested our procedure on data whose word (digit) boundaries were randomly marked. In order to randomize the boundaries, we modeled the duration (length) of each digit as a Poisson distribution with certain intensity λ (see Fig. 1), depending on the particular digit being considered. The intensities were estimated by the average duration of the digits in the training set. Table VI shows these estimates for both corpora.

Each utterance ω from the TI-digits testing data set was assigned a duration $d(\omega)$ chosen at random according to the Poisson distribution associated to the corresponding digit. If the random duration $d(\omega)$ was shorter than the actual duration $T(\omega)$ of the utterance, then the utterance was modified by shortening it to $d(\omega)$ time frames. The random segmentation was done by selecting at random from the collection of time frames $\{1, 2, \cdots, T(\omega) - d(\omega)\}$, a starting time (left boundary) for the modified utterance; the ending time (right boundary), was set so that the total duration of the modified utterance was $d(\omega)$. About half the utterances were modified in this manner. Only modified utterances were tested. This procedure was applied ten times to the available testing data set, yielding the error rates shown in Table VII.

TABLE VI
POISSON INTENSITIES (IN MILLISECONDS)

Corpus	0	1	2	3	4	5	6	7	8	9	oh
TI-digits	491	378	337	362	364	396	521	472	317	450	327
Number	430	300	319	327	360	433	446	440	297	379	211

TABLE VII
ERROR RATES FOR MODIFIED BOUNDARIES OF THE TI-DIGITS

Data	1	2	3	4	5	6	7	8	9	10	Mean	Std. Dev.
Error	1.9	2.1	1.9	2.2	2.5	2.0	2.0	1.9	2.2	2.0	2.0	0.2

The average error rate for the modified boundaries data is almost the same as the error rate of the correctly segmented data; this gives strong evidence that our procedure is not sensitive to small to moderate errors in word segmentation.

VI. COMPUTATION CONSIDERATIONS

It is important to note that our procedure not only produces good classification rates that are comparable to those yielded by HMMs, but also that it requires orders of much fewer calculations (computer operations) both during training and testing.

The following figures were measured on a PC with a 333 MHz Pentium II processor running Cygnus software over Windows NT. It takes 2.82 min to grow a tree of average depth 9.6 with the 4460 data points comprising the TI-digits corpus. We emphasize that the implementation of the algorithm has not been optimized, just as an example the data is reloaded into the program for every tree, both for training and testing. To obtain comparable performance to the HMMs we need 11 trees (see Table III), i.e., 31 min. The training time of the HMMs on the TI data is 123 min on a 250 MHz SunW Sparc Ultra-30. The two machines are comparable in speed. Testing takes about 0.003 seconds per data item per tree, and approximately .004/s for the HMM models.

Since it is hard to compare computation times between the algorithm we have programmed on our own and an optimized commercial software package, we present a more theoretical comparison between the order of calculations needed by our procedure and by HMMs to solve these tasks. We believe this offers some evidence on the computational gain conveyed by our approach.

1) HMMs: Let r denote the number of data points in the training set, and d, the average duration of the utterances in the training set. A HMM is characterized by its number of states s, and mixture components m. Each iteration of the Baum-Welch recursion formulas requires $O(s^2md \times r)$ operations (here $O(\cdot)$ stands for the order (magnitude) of the number of operations). On the other hand, ignoring the maximization step to find the best class, testing only requires $O(s^2md \times K)$ operations (recall that K is the number of classes).

2) Decision Trees: As before, let N denote the total number of trees to be grown. The relevant quantities for a tree are the average depth n of the trees, and the number n_q of randomized queries entertained as admissible splitting rules at each nonterminal node. The selection of an optimal query at each nonterminal node requires $O((K+d) \times n_q)$ operations, since 1) evaluations of the form $t(\ell_1) - t(\ell_2) \in I$, for fixed tags ℓ_1, ℓ_2, over all

the locations of these two tags, require $O(d \times \text{length}(I))$ operations and 2) the average entropy of the children nodes requires $O(K)$ operations, since only a fixed number of data points are used in its computation. Since a tree of depth n has at most 2^n-1 nonterminal nodes, and at most 2^n leaves, the number of operations needed to grow N trees is of $O(N \times \{2^n(K+d)n_q + 2^n rK\})$. Again, testing only requires $O(ndN + KN)$ operations.

3) Computational Gain During Testing: From the above calculations, the testing step in both procedures requires about the same number of operations when the number of trees grown N is about $O(m \times s^2K/(n+K))$. Since usually n and s are rather small numbers, the balancing variable is the number of mixture components m. For large m, decision trees are much faster. In fact, assuming $s = 5$, and $K = 10$, a hundred trees of average depth ten, require about as many operations as ten HMMs with eight mixture components in each state.

4) Computational Gain During Training: There is no doubt of the enormous gain in computational cost, when using decision trees rather than HMMs. Indeed, from the above calculations, i iterations of the Baum-Welch algorithm for HMMs require $O(K(s^2mdr/K) \times i)$ operations, which is an extremely large number of operations, even for a moderate number of iterations, when compared to $O(N2^n(K+d)n_q + N2^nrK)$ operations required for N decision trees. In fact, if we set $s = 5$, $m = n = 10$, and $n_q = d = 100$, four iterations of the Baum-Welch algorithm require as many operations as growing 10 trees.

VII. Conclusion

Acoustic events based on tags localized in time and frequency, and simple coarse temporal relations, provide informative features for classification of acoustic signals. These events are defined in terms of labeled graphs and inherit a partial ordering. We employ multiple randomized decision trees to access the rich pool of acoustic events, in a systematic way, exploiting the partial ordering to proceed from coarse to fine representations. Time invariance is directly incorporated through the relations; invariance to audio quality is incorporated through the coarse definition of the tags. The learning stage for this approach is much more efficient than for HMMs. Recognition rates are better and recognition times are also faster than the more complex HMM models. On the other hand the issue of segmentation and continuous speech analysis is not addressed.

Acknowledgment

The authors would like to thank Prof. L. Atlas, the Director of the Interactive Systems Design Laboratory, University of Washington, for letting them use the lab to run speech recognition experiments with HMMs and J. Droppo for his help in introducing them to the HTK toolkit HMM software.

References

[1] Y. Amit, G. Blanchard, and K. Wilder, "Multiple randomized classifiers (MRCL)," Univ. Chicago, Chicago, IL, Tech. Rep. 496, 1999.

[2] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural Comput.*, vol. 9, pp. 1545–1588, 1997.

[3] Y. Amit, D. Geman, and K. Wilder, "Joint induction of shape features and tree classifiers," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, 1997.

[4] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[5] R. A. Cole, M. Fanty, and T. Lander, "Telephone speech corpus development at CSLU," in *Proc. Int. Conf. Spoken Language Processing*, 1994.

[6] B. Gidas and A. Murua, "Classification and clustering of stop consonants via nonparametric transformations and wavelets," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, 1995, pp. 872–875.

[7] J. Li and A. Murua, "A 2D extended HMM for speech recognition," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1999.

[8] P. Lieberman and S. Blumstein, *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge, U.K.: Cambridge Univ. Press, 1988.

[9] K. W. Ma, "Applying large vocabulary hybrid HMM-MLP methods to telephone recognition of digits and natural numbers," Int. Comput. Sci. Inst., 1995.

[10] A. Murua, "Upper bounds for error rates associated to linear combination of classifiers," Dept. Statistics, Univ. Washington, Seattle, Tech. Rep. 354, 2000.

[11] P. Niyogi and P. Ramesh, "Incorporating voice onset time to improve letter recognition accuracies," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1998.

[12] P. Niyogi, C. Burges, and P. Ramesh, "Distinctive features detection using support vector machines," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1999.

[13] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, pp. 81–106, 1986.

[14] K. Wilder, "Decision trees for shape recognition," Ph.D. dissertation, Univ. Mass., Boston, 1998.

[15] A. K. C. Wong and M. You, "Entropy and distance of random graphs with application to structural pattern recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 7, pp. 599–609, 1985.

[16] Y. Yan, M. Fanty, and R. Cole, "Speech recognition using neural networks with backward-forward probability generated targets," in *Int. Conf. Acoustics, Speech, Signal Processing*, 1997.

[17] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book, Version 2.2*: Entropic, 1999.

[18] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," *J. Acoust. Soc. Amer.*, vol. 33, p. 248, 1961.

Yali Amit received the Ph.D. degree from the Weizmann Institute, Rehovoth, Israel in 1988.

During his graduate studies, he was a Software Engineer with Architecture and Computer Aids, developing 3-D CADD algorithms for architectural and product design. In 1988, he joined the Probability and Pattern Theory Group, Division of Applied Mathematics, Brown University, Providence, RI. His research has centered around the theory and applications of deformable templates as a method for high-level analysis of images. He has also worked on rates of convergence of certain Monte-Carlo algorithms and on extensions of errorless coding theory to Gibbs random fields.

Dr. Amit received the Weizmann Fellowship for Post-Doctoral research in 1988.

Alejandro Murua received the Mathematical Civil Engineering Diploma from the School of Engineering, University of Chile, in 1988, and the Ph.D. degree in applied mathematics from Brown University, Providence, RI, in 1994.

He was a Visiting Assistant Professor with the Division of Applied Mathematics, Brown University, during 1993–1994, and an Assistant Professor of statistics with the University of Chicago, Chicago, IL, from 1994 to 1998. In 1998, he joined the faculty of the Department of Statistics, University of Washington, Seattle, where he currently serves as an Assistant Professor. His main research interests focus on applications of statistics and probability to problems dealing with machine learning, speech and object recognition, signal processing, and data mining.