Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models

JAY G. WILPON, SENIOR MEMBER, IEEE, LAWRENCE R. RABINER, FELLOW, IEEE, CHIN-HUI LEE, MEMBER, IEEE, AND E. R. GOLDMAN

Abstract—**Speaker independent recognition of small vocabularies, spoken over the long-distance telephone network, has been demonstrated to be a viable technology. However, the algorithms tested and the tasks evaluated typically assume that user input be restricted to only a set of defined vocabulary words. Recently, a large scale trial of speaker independent isolated word speech recognition technology was carried out in Hayward, CA. The task chosen required that users speak, in isolation, one of five defined vocabulary words (*collect, calling card, person, third number,* and *operator*). Recognition results were obtained which showed that when users spoke the vocabulary words in an isolated fashion, the words were correctly recognized about 99% of the time. However, observations of customer responses during this trial showed that about 20% of the utterances had the desired vocabulary word along with extraneous input which ranged from nonspeech sounds (e.g., clicks and breath noises) to groups of nonvocabulary words (e.g., "I want to make a *collect* call please"). Most conventional recognition algorithms have not been designed to handle this type of input. As such, modification of the algorithms had to be made to recognize vocabulary words embedded in speech (i.e., a form of keyword spotting).**

This paper describes the modifications made to a connected word speech recognition algorithm based on hidden Markov models (HMM's) which allow it to recognize words from a predefined vocabulary list spoken in an unconstrained fashion. The novelty of our approach is that we create statistical models of both the actual vocabulary words and the extraneous speech and background. An HMM-based connected word recognition system is then used to find the best sequence of background, extraneous speech, and vocabulary word models for matching the actual input. Word recognition accuracy of 99.3% on purely isolated speech (i.e., only vocabulary items and background noise were present), and 95.1% when the vocabulary word was embedded in unconstrained extraneous speech, were obtained for the five word vocabulary using the proposed recognition algorithm.

I. Introduction

THE development of robust, speaker independent, speech recognition systems that perform well over dialed-up telephone lines has been a topic of interest for over a decade [1]–[7]. This work has progressed from systems that can recognize a small number of vocabulary items spoken in isolation [3], [5], [7], to systems that can recognize medium size vocabulary sets spoken fluently [4]. A basic assumption for most speech recognition systems is that the input to be recognized consist solely of words from the recognition vocabulary and background silence. However, previous studies on the recognition of a limited set of isolated command phrases for making "operator assisted calls" have shown that it is extremely difficult, if not impossible, to get real-world subscribers to such a service to speak only the allowable input words [1], [5], [6]. In a large scale trial of speaker independent, isolated word, speech recognition technology, carried out at an AT&T central office in Hayward, CA (i.e., the San Francisco Bay area) [1], [8], live telephone customer traffic was used to evaluate the call handling procedures being developed for a new generation of telephone switching equipment. Using these procedures, customers making operator assisted calls were given the option of verbally identifying the type of call they wished to make (i.e., *collect* for a collect call, *calling card* for a calling card call, *person* for a person-to-person call, *third number* for a bill-to-third-party call, and *operator* to get the operator). Each caller was requested to speak one of the five orally prompted commands in an isolated fashion. While 82% of the users actually spoke one of the command words, only 79% of these inputs were spoken in isolation (i.e., only 65% of all the callers followed the protocol). Monitoring of the customer's spoken responses indicated that 17% of all responses contained a valid vocabulary item along with extraneous speech input (e.g., *I want to make a* **collect** *call please*). Most conventional isolated word recognition algorithms have not been designed to recognize vocabulary items embedded in various carrier sentences. As such, modifications to the algorithms had to be made to allow for the recognition of words embedded in extraneous speech.

In this paper we discuss the problem of recognizing a small set of prescribed vocabulary words spoken in the context of unconstrained speech. In the general case, the recognition system is presented with continuous input and must decide whether or not any of the pre-defined vocabulary words is present anywhere in the speech. While much research has been performed on the general word-spotting task, very little of it has been published. Most of the techniques that have been described in the literature are template-based dynamic time-warping approaches (DTW) [9]–[11]. For example, in [9], Christiansen and Rushforth described a speaker trained keyword spotting system which uses an LPC representation of the speech signal without any syntactic or semantic information about the task. Using this approach they achieved good results on a vocabulary set of four words and the ten digits. Myers

Manuscript received May 11, 1989; revised November 30, 1989.
The authors are with AT&T Bell Laboratories, Murray Hill, NJ 07974.
IEEE Log Number 9038429

et al. [11] described an approach which used a local minimum DTW-based algorithm for the problem of word spotting. However the proposed system was not evaluated on any real task.

Higgins and Wohlford [10], also proposed a DTW-based system for keyword spotting. In their system, knowledge about the vocabulary and syntax of the input speech was used. A set of keyword templates and nonkeyword templates was created and compared against several pooled *filler* templates as to their ability to detect keywords in fluent speech. These filler templates were generated 1) using data from six "function" words, or 2) by clustering some nonvocabulary words into segments roughly equal to syllables using hand-marked data. Their results indicated that while explicit knowledge of the vocabulary may not be that important, the use of filler templates may be important. However, they found that the number of such filler templates greatly influenced their results. Additionally, they determined that the durations of the filler templates controlled the accuracy of their system. As the number of templates was increased and the duration of the average filler template was shortened, the system accuracy improved. Duration constraints are a major problem in any DTW-based template matching recognition system, since each template has a physical duration and the algorithms are forced to adhere to some local time duration constraints. An advantage of using hidden Markov models is that durations are statistically modeled as part of the training procedure.

In Bossemeyer *et al.* [12], a DTW-based algorithm approach to the problem of finding keywords was described which matched the keyword templates to the unknown speech at each starting frame of the utterance. (Bossemeyer used the same keywords and test data that we use in this paper.) Penalties were added to account for voicing duration and energy level. This algorithm was tested on an independent data base and had a 90% recognition accuracy rate on utterances containing extraneous speech and 97.1% on utterances containing only the keyword. The algorithm we present here will be shown to perform significantly better on this same data base.

A significant amount of progress has been recently made in automatic speech recognition using hidden Markov modeling (HMM) [13]–[25]. Since the HMM approach uses a statistical characterization of the signal, it should contain more information about the signal than does the DTW-based approach. As such, we chose to develop an algorithm using HMM technology to attack the problem of recognizing vocabulary words in fluent speech.

In Wilpon *et al.* [26], a first attempt at using a hidden Markov model based recognition system for recognizing a limited set of vocabulary words spoken in unconstrained speech was described. The algorithm presented, similar to the one used in the template-based system described by Christiansen and Rushforth [9], can be thought of as sliding the input speech past each model in a continuous manner. However, the results achieved were comparable to those based on the template system described in [12].

In this paper we present a recognition system which uses hidden Markov modeling techniques to explicitly model both the actual vocabulary words as well as extraneous inputs. Evaluating this approach on a large speaker independent data base gave word accuracies of 99.3% when the vocabulary word was spoken in isolation, and 95.1% when the vocabulary word was embedded in extraneous speech.

In Section II we describe the HMM-based algorithm for recognizing vocabulary words in the context of unconstrained inputs. The structure of the HMM's used in our work is presented in Section III. Section IV describes the speech data base used to evaluate our algorithm. Finally, in Section V we present results from a series of recognition experiments.

II. HMM-Based Recognition Algorithm

Speech recognition systems which do not require explicit detection have been widely described in the literature [13]–[16]. In [16], Wilpon and Rabiner presented an HMM-based recognition algorithm, which showed that explicit endpoint detection of speech could be removed entirely from the recognition system while maintaining high recognition accuracy. To achieve this, the recognition system modeled the incoming signal as a sequence of *background signal* and *vocabulary words*. However, this work was limited in that the vocabulary words had to be spoken with no extraneous input.

In our current work, the ideas developed in [16] are extended to handle the case of vocabulary words spoken in the context of unconstrained speech. The approach that we have developed models the entire background environment, including *silence*, *transmission noises*, and, most importantly, *extraneous speech*. We represent a given input as an unconstrained sequence of background and extraneous speech followed by vocabulary words followed by another unconstrained sequence of background and extraneous speech. We do this by creating one or more hidden Markov models, which we call *garbage* models, representative of extraneous speech inputs. A grammar driven continuous word recognition system is then used to determine the best sequence of extraneous speech, background, and vocabulary words. Given this structure for the recognition system, the garbage models match the extraneous speech and the trained vocabulary word models match the actual vocabulary word that was spoken.

A comprehensive discussion of the complete HMM-based connected word system is given in [21]. In this section we present a brief overview of the recognition system. Fig. 1 shows a block diagram of the HMM-based recognition system. The key elements of the system are described in the following sections.

A. LPC and Cepstral Analysis

Speech is first filtered typically to a bandwidth of 100–3200 Hz and then digitized typically at a 6.67-kHz rate. The digitized speech is then preemphasized using a simple first-order digital filter with a preemphasis factor $a =$

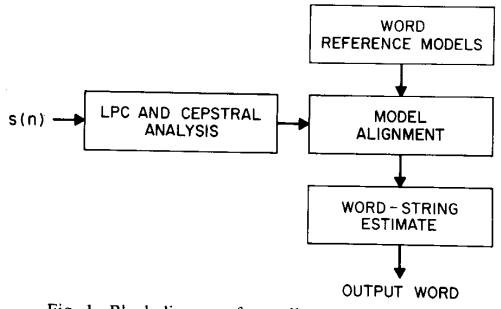

Fig. 1. Block diagram of overall recognition system.

0.95, and blocked into frames of 45 ms in length with a shift between frames of 15 ms. Each frame of speech is weighted by a Hamming window. A pth-order linear predictive coding (LPC) analysis is then performed on the data. Thus, for each frame, a set of $p + 1$ LPC coefficients is generated. The input signal is then reduced to a sequence of LPC frame vectors. There is no automatic endpoint detection performed on the data. The LPC derived cepstral vector is then computed up to the Qth component, where $Q > p$ (where $Q = 12$ in our implementation).

Each coefficient of the Q-coefficient cepstral vector, $c_l(m)$, at time frame l is weighted by a window $W_c(m)$ of the form

$$W_c(m) = \left[1 + \frac{Q}{2} \sin\left(\frac{\pi m}{Q}\right)\right], \quad 1 \le m \le Q \quad (1)$$

to give

$$\hat{c}_l(m) = c_l(m) \cdot W_c(m). \quad (2)$$

It has recently been shown that by extending the analysis vector to include spectral derivative (in time) information, performance of several standard speech recognizers improved significantly [17], [18], [22]. As such we include such spectral derivative information in our analysis vector as follows.

The time derivative of the sequence of weighted cepstral vectors is approximated by a first-order orthogonal polynomial over a finite length window of $(2K + 1)$ frames, centered around the current vector. ($K = 2$ in our implementation; hence the derivative is computed from a 5 frame (75 ms) window.) The cepstral derivative (i.e., the delta cepstrum vector) is computed as

$$\Delta\hat{c}_l(m) = \left[\sum_{k=-K}^{K} k\hat{c}_{l-k}(m)\right] \cdot G, \quad 1 \le m \le Q \quad (3)$$

where G is a gain term so that the variances of $\hat{c}_l(m)$ and $\Delta\hat{c}_l(m)$ are about the same. (For our system the value of G was 0.375.)

The overall observation vector $\boldsymbol{O}_l$ used for scoring the HMM's is the concatenation of the weighted cepstral vector, and the corresponding weighted delta cepstrum vector, i.e.,

$$\boldsymbol{O}_l = \left\{\hat{c}_l(m), \Delta\hat{c}_l(m)\right\} \quad (4)$$

and consists of 24 coefficients per vector.

B. Model Alignment Procedure

The sequence of spectral vectors of an unknown speech utterance is matched against a set of stored word-based hidden Markov models using a syntax-derived, frame-synchronous, network search algorithm (described in [19]). Word and state duration probabilities have been incorporated into the HMM scoring and network search. A finite state grammar, describing the set of valid sentence length inputs, is used to drive the recognition process. The recognition algorithm performs a maximum likelihood string decoding on a frame-by-frame basis, therefore making optimally decoded partial strings available at any time. The output of this process is a set of valid candidate strings.

C. Generating Word Reference Models

In order to generate one or more word models from a training data set of labeled speech, a segmental k-means training algorithm is used [20].[1] This word building algorithm (i.e., an estimation procedure for determining the parameters of the HMM's) is iterated for each model until convergence (i.e., until the difference in likelihood scores in consecutive iterations is sufficiently small).

To create multiple models per word an HMM-based clustering algorithm is used to split previously defined clusters [25]. This algorithm, based on the likelihoods obtained from the current set of HMM's, separates out from the set of training tokens those tokens whose likelihood scores fall below some fixed or relative threshold. That is, we separate out all the tokens with poor likelihood scores and create a new model out of these so-called outlier tokens. Once the tokens have been clustered, the segmental k-means training algorithm is again used to give a (locally optimal) set of parameters for each of the models. Further details of this algorithm can be found in [25].

III. Structure of Hidden Markov Models

Fig. 2 illustrates the structure of the HMM's used to characterize individual words as well as the background environment, and the extraneous speech [21]–[23]. The models are first-order, left-to-right, Markov models with N states. Each model is completley specified by the following:

1) A state transition matrix $A = a_{ij}$, with the constraint that

$$a_{ij} = 0 \quad j < i, j \ge i + 2 \quad (5)$$

(i.e., we allow transitions from state j only to itself, or to state $j + 1$).

2) A continuous mixture density matrix $B = b_j(\boldsymbol{x})$ of the form

$$b_j(\boldsymbol{x}) = \sum_{m=1}^{M} c_{mj} N[\boldsymbol{x}, \boldsymbol{\mu}_{mj}, \boldsymbol{U}_{mj}] \quad (6)$$

[1]The segmental k-means algorithm tries to optimize the likelihood of the observation sequence and the state sequence over all model parameters as opposed to the conventional Baum–Welch procedure which tries to optimize the likelihood of the observation sequence (over all possible state sequences) over all model parameters.

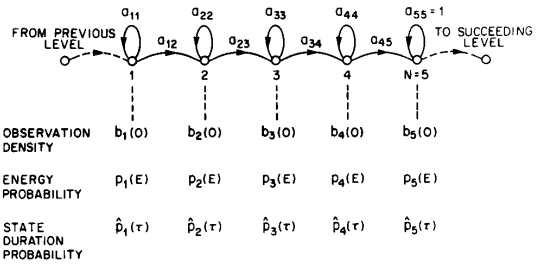

Fig. 2. Representation of HMM used for each word in the vocabulary and for the silence and extraneous speech models.

where $\boldsymbol{x}$ is the input cepstral vector, c_{mj} is the mixture weight for the mth component in state j, $\boldsymbol{\mu}_{mj}$ is the mean vector for mixture m in state j, and $\boldsymbol{U}_{mj}$ is the covariance for mixture m in state j. All evaluations described in this paper used diagonal covariance matrices. In our evaluations, the number of states per model was set to 10 and the number of mixture components per state M was set to nine. (Several other values for N and M were evaluated.)

3) A set of log energy densities $p_j(\varepsilon)$, where ε is the dynamically normalized frame energy, and p_j is an empirically measured discrete density of energy values in state j.

4) A set of state duration probabilities $\hat{p}_j(\tau)$, where τ is the number of frames spent in state j, and $\hat{p}_j$ is an empirically measured discrete density of duration values in state j. (Although $\hat{p}_j(\tau)$ is clearly not independent of the exponential duration density implied by the self-transition coefficient a_{jj}, in practice it has been found that one can assume independence and not have any serious effect on recognition performance [18].)

IV. Modeling of Background and Extraneous Speech

The grammar used in the recognition process allows for any number (or zero) of (extraneous speech) garbage models and background models followed by one or more (or none) of the vocabulary words to be recognized and followed by another unconstrained sequence of garbage and background models. In our tests we know *a priori* that only one vocabulary word appears in any utterance, hence we limited the grammar to find exactly one vocabulary word. This is shown graphically in Fig. 3, where node *0* is the starting node and node *1* is the terminal node.

The garbage models and background models are generated automatically, using the training procedures described in Sections II-C and VI.

V. Experimental Data Base

A speech data base, consisting of approximately 75 000 utterances, was collected during a large scale trial of speaker independent isolated word recognition technology, carried out at an AT&T central office in Hayward, CA [1], [8]. The five word vocabulary defined for this task was *collect*, *calling card*, *third number*, *person*, and *operator*. Each utterance was obtained from a telephone

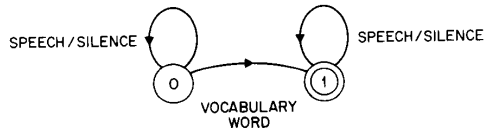

Fig. 3. The grammar used to recognize the vocabulary words in the context of extraneous speech and/or background silence.

customer during the course of a normal operator assistance call. Each caller was automatically prompted (by a voice response system) to speak one of the five keywords in an isolated fashion. During the trial about 17% of all customer responses contained a valid vocabulary word along with extraneous input which ranged from non-speech sounds, such as background music or door slams, to extraneous speech such as:

- ⟨silence⟩ *collect* call please ⟨silence⟩
- Um? Gee, okay, I'd like to place a *calling-card* call
- *Collect* from Tom ⟨silence⟩
- I want a *person* call
- ⟨silence⟩ Please give me the *operator*.

Of the 75 000 collected utterances, a random set of 7980 utterances was digitized and used for experimentation. Table I shows the distribution of the digitized utterances that were used in our recognition experiments. The distribution of utterances for each of the five vocabulary words varies widely and is proportional to the percent of operated assisted calls made of each type. Each customer's utterance contained only one of the prescribed vocabulary words. A trained listener labeled the 1483 customer utterances that contained extraneous speech, so as to create separate vocabulary word and extraneous speech data bases. Fig. 4 shows the segmentation output for a typical utterance. Fig. 4(a) shows the waveform and Fig. 4(b) shows the log energy contour. The dashed lines indicate the boundaries between words. The extraneous speech data base was used to train our garbage (extraneous speech) models. Additional examples are shown in Figs. 5 and 6. Roughly half the utterances in each category were used for training; the other half were used for testing.

Some interesting results can be found by examining the actual speech transcriptions. In addition to the five vocabulary words, there were 477 other unique words spoken. Table II shows the distribution of the three most frequent extraneous inputs as a function of the vocabulary word spoken. For example, when the vocabulary word *calling card* was spoken along with extraneous input, the word *um* was also spoken 53.8% of the time. Table III shows the distribution of the three most frequent words as a function of the vocabulary word spoken including whether the extra word was spoken *immediately* before or after the vocabulary word. If we were to optimize our algorithm to this specific vocabulary and task, this information could be very useful. We see in this table that the word *um* occurred immediately before a vocabulary word about 80% of the time (when the vocabulary word was not spoken in isolation).

TABLE I
HAYWARD SPEECH DATA BASE DISTRIBUTIONS

Vocabulary Word	Total	Words Spoken as Isolated Utterances		Words Spoken as Part of a Connected Phrase	
		Training	Testing	Training	Testing
Collect	4320	1703	1602	510	505
Calling-card	2121	910	955	132	123
Third-number	619	273	281	34	31
Person	170	77	67	14	12
Operator	751	320	309	65	57
Totals	7980	3283	3214	755	728

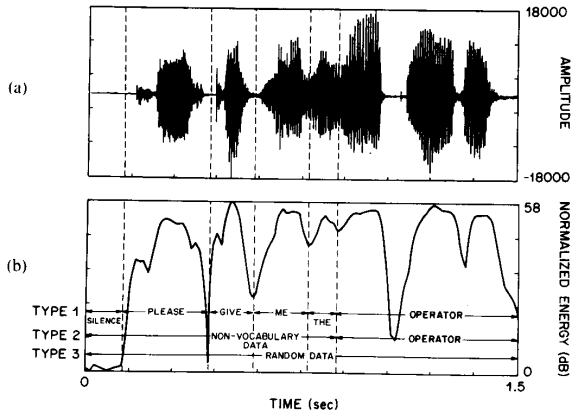

Fig. 4. (a) Linear amplitude and (b) normalized energy plots of the utterance, "Please give me the operator," showing 3 types of segmentation and labeling. Each of the 3 sets of segmentations and labels was used to train vocabulary word, background silence, and extraneous speech models.

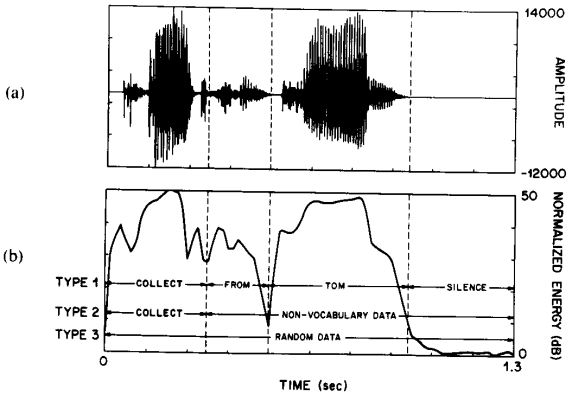

Fig. 5. (a) Linear amplitude and (b) normalized energy plots of the utterance, "Collect from Tom (silence)," and the respective segmentations and labels.

VI. RECOGNITION EXPERIMENTS

Several recognition experiments were carried out to determine the minimal amount of *a priori* knowledge of the data base needed to train the vocabulary word, background, and extraneous speech models. Additionally, we also examined the issue of number and size (in terms of number of states) of the extraneous speech models necessary to achieve the best performance.

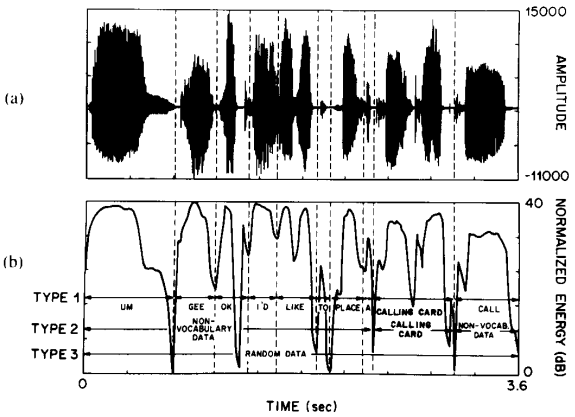

Fig. 6. (a) Linear amplitude and (b) normalized energy plots of the utterance, "Um gee OK I'd like to place a calling card call," and the respective segmentations and labels.

TABLE II
OCCURRENCE OF SELECTED NONVOCABULARY WORDS SPOKEN ALONG WITH VOCABULARY WORDS

Selected Non-keywords with their Frequency					
Non-Vocabulary	Collect	Calling-card	Third-number	Person	Operator
um	24.5%	53.8%	39.7%	45.8%	28.3%
please	25.0	19.6	20.6	16.7	44.2
call	34.1	2.3	0	8.3	0.9

A. Data Base Labeled as Vocabulary Words and Specific Nonvocabulary Words

In this first experiment, we assumed that we had a completely labeled speech data base, which consisted of both vocabulary words and specific nonvocabulary words. Several examples of this type of labeling can be seen in Figs. 4–6–indicated as *Type 1*. Based on the labeling of the data base described in Section V, 10-state, 9 mixture/state hidden Markov models were generated for each of the n most frequently spoken nonvocabulary words and noises, plus a single 10-state, 9 mixture/state model for the background. Table IV shows a list of the 13 most frequently occurring extraneous signals. The list contains mainly words, but also contains many nonword signals, for example, clicks and breath noises. Table V shows recognition word accuracy and the extraneous speech (signal) coverage as a function of n. We see that the recognition accuracy for the isolated data base is relatively insensitive to the number of extraneous speech models in the range of from 3 to 13 nonvocabulary word models. This implies that there is a large separation between the extraneous speech models and the specific vocabulary word models for this task. We also see that with $n = 1$, only 11.6% of all extraneous speech segments have been used to create the single garbage model. However, the

TABLE III
SELECTED NONVOCABULARY WORDS WITH THEIR FREQUENCY OF OCCURRENCE DIRECTLY BEFORE AND DIRECTLY AFTER THE KEYWORDS

Non-Vocabulary	Collect		Calling-card		Third-number		Person		Operator	
	% Before	% After	% Before	% After	% Before	% After	% Before	% After	% Before	% After
um	77.8	6.5	82.9	4.3	85.2	3.7	63.6	0	71.9	3.1
please	0	68.5	0	84.3	0	92.9	0	100	0	96.0
call	4.1	93.6	16.7	83.3	–	–	0	100	0	0

TABLE IV
MOST FREQUENTLY OCCURRING NONVOCABULARY ITEMS PRESENT IN THE DATA BASE

Frequency Number	Item
1	um
2	please
3	call
4	**clicks** (lip smacks and telephone lineclicks)
5	**breath noises**
6	from
7	**background noise** (eg. stereo)
8	**background speech** (eg. side conversation at public telephones)
9	is
10	this
11	to
12	my
13	a

TABLE V
RECOGNITION WORD ACCURACY AS A FUNCTION OF THE NUMBER OF NONVOCABULARY WORD SPECIFIC MODELS

# of Non-Vocabulary Word Specific Models	% Coverage of Extraneous Speech Segments	% Correct Isolated Data	% Correct Embedded Data	% Correct All Data
1	11.6	98.9	91.6	97.6
2	21.0	98.9	91.9	97.6
3	30.0	99.0	94.6	98.2
4	38.2	99.0	94.8	98.2
7	55.2	99.1	94.8	98.3
13	67.9	99.1	94.6	98.3

recognition accuracy is fairly high. As n varies from 2 to 13, we see that the system performance increases, until a coverage of 30% ($n = 3$) at which point it levels off. By creating models for the most frequently occurring words and noises (in addition to a background model), we were able to achieve a recognition rate of 94.8% when the vocabulary word was spoken along with extraneous input (*embedded*) and 99.1% when the vocabulary word was spoken without extraneous input (*isolated*).

B. Data Base Labeled as Vocabulary Word and Extraneous Speech Sequences

In this experiment, we relaxed the labeling requirements on the training data base. Here we made the assumption that the data base was either labeled as vocabulary words or extraneous speech sequences, without any detailed classification of the extraneous speech sequences. Examples of this type of data labeling are shown in Figs. 4–6 and are indicated as *Type 2*.

In this test, all the extraneous speech regions (including nonvocabulary words and noises) were combined together and a single hidden Markov model was trained as a universal extraneous speech model. A separate background model was also generated. For this experiment we examined the effects of varying the number of states used for the background model on the word recognition accuracy. The results of this experiment, as shown in Table VI, indicate that with a *single* extraneous speech model and a single 1-state background model, the word recognition accuracies (98.5% on isolated data and 94.0% on embedded data) were comparable to those obtained using 13 10-state HMM's generated from specific word tokens (as shown in Table V). Slight improvements were obtained using a single 10-state background model, for which the word recognition accuracies were 99.2% for isolated data and 94.2% for embedded data (for an overall accuracy of 98.3%). These recognition accuracies are significantly better (more than 5% improvement in word accuracy) than those reported by Bossemeyer *et al.* on the same data base and for the same task [12]. Tables VII and VIII show the recognition confusion matrix generated from this test for the isolated and embedded speech data. This table shows that the recognition accuracy is highest for the vocabulary words that had the most training data available (e.g., *collect* and *calling card*) and lowest for the words with the least training (e.g., *person* and *third number*).

TABLE VI
RECOGNITION WORD ACCURACY AS A FUNCTION OF THE NUMBER OF STATES IN A SINGLE SILENCE HMM

# of States in Silence Model	% Correct Isolated Data	% Correct Embedded Data	% Correct All Data
1	98.6	94.0	97.8
2	97.3	91.9	96.3
3	99.2	93.8	98.2
4	99.1	94.2	98.2
5	99.3	94.1	98.3
10	99.2	94.2	98.3

Given the encouraging performance scores of Table VI, we used the HMM clustering algorithm described above to build multiple HMM's from our large unlabeled data base of extraneous speech inputs. Table IX shows the word recognition accuracies as a function of the number of extraneous speech HMM's used (all models were 10-state, 9-mixture per state models). The overall performance (i.e., word recognition accuracy) remains about the same when using from one model to five models to represent the extraneous speech.

To improve performance further, we can use a rejection

TABLE VII
CONFUSION MATRIX FOR ISOLATED SPEECH RECOGNITION TEST (IN PERCENT)

		Recognized Word					
		Collect	Calling-Card	Third-Number	Person	Operator	Total # of Utterances
Actual Word	Collect	99.7	0.2	0	0	0.1	1602
	Calling-Card	0.9	98.9	0	0.1	0.1	955
	Third-Number	7.1	3.6	85.8	0	3.6	67
	Person	0	1.5	0	97.0	1.5	28
	Operator	1.6	0.6	0	0.3	97.4	309

TABLE VIII
CONFUSION MATRIX FOR EMBEDDED SPEECH RECOGNITION TEST (IN PERCENT)

		Recognized Word					
		Collect	Calling-Card	Third-Number	Person	Operator	Total # of Utterances
Actual Word	Collect	94.7	2.2	0.8	0.1	1.4	505
	Calling-Card	0.8	97.6	0.8	0.8	0	123
	Third-Number	6.5	0	90.3	0	3.2	31
	Person	8.3	0	0	83.3	8.3	12
	Operator	3.5	0	0	0	96.5	57

TABLE IX
RECOGNITION WORD ACCURACIES AS A FUNCTION OF THE NUMBER OF GARBAGE HMM'S

# of Garbage Models	% Correct Isolated Data	% Correct Embedded Data	% Correct All Data
1	99.2	94.2	98.3
2	99.3	93.8	98.3
3	99.2	94.9	98.4
4	99.3	95.1	98.5
5	99.3	94.5	98.4

criterion to defer recognition decisions. Fig. 8 shows a plot indicating the tradeoff between recognition error rate and rejection rate (i.e., no recognition decision is made) based on the recognizer that used five extraneous speech models. The dashed line shows the results for vocabulary words embedded in extraneous speech; the solid line shows the results for the purely isolated speech; the dotted line shows the results on the entire data base. These results were obtained by applying a threshold test on the output likelihood scores generated by the recognizer. The figure shows that to achieve greater than 98% correct on the embedded speech about 10% of the utterances would have to be rejected. For the isolated data to be recognized with a 99.7% accuracy, 4.0% of the utterances would have to be rejected. Finally, for the entire data base to be recognized at an accuracy greater than 99.5%, 5.8% of all utterances would have to be rejected.

C. Data Base Labeled for Vocabulary Words Only

In this last experiment, we removed all constraints on the labeling in the data base used to train the extraneous speech model. The only requirements were that we orthographically label (but not segment) the vocabulary words in each training utterance. For initialization of the background silence model and the universal extraneous speech model, we assume only that we can use the training data base as an unlabeled set. Examples of this type of labeling can be seen in Figs. 4–6–indicated as *Type 3*. Even though a vocabulary word is present in these examples, the *entire* utterence is used to train the extraneous speech model. Fig. 7 shows a flow diagram of the training process used to obtain the final vocabulary word and extraneous speech models. To initialize the training process, a set of HMM's is built from the isolated vocabulary words and the pool of random speech data. Given this set of bootstrap models and the set of training data that contains the vocabulary words, the segmental k-means training algorithm is used to segment the training strings into vocabulary words, background, and extraneous speech. New models are then created and the process iterates itself to convergence. Using this fully automatic training procedure to generate a single extraneous speech model, a word recognition accuracy of 99.4% was obtained when the vocabulary word was spoken in isolation and 94.5% when the vocabulary was embedded within extraneous input. These results, which are comparable to those based on the most detailed training, show that both background and extraneous speech models can be bootstrapped successfully without having a detailed transcription of the training files.

In a final experiment a single extraneous speech model was generated, using the training procedure just described, without generating a separate background model. Recognition results were comparable to the results presented above, namely 99.1% word accuracy for the isolated data base and 94.0% word accuracy for the embedded data base. This indicates that a single extraneous

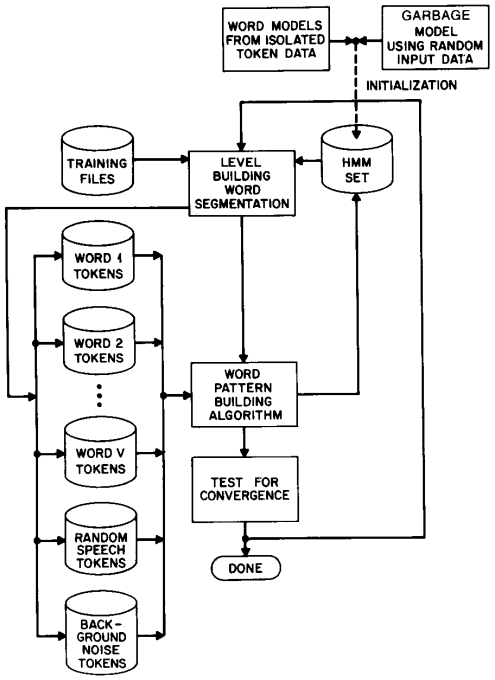

Fig. 7. Flow diagram of fully automatic training procedure.

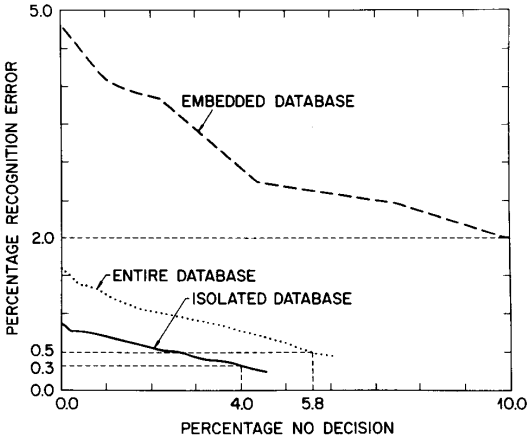

Fig. 8. Tradeoff in word recognition error rate versus rate of rejection (no decisions) for the embedded words, the isolated words, and the entire data base.

speech model can be generated which adequately characterizes both the extraneous speech and the background.

VII. Discussion

Automatic speech recognition can be used to greatly enhance current telephone network based services and to create a wide variety of new services. In addition, since Touch-Tone® penetration is only about 65% nationwide in residences, ASR capabilities are a must if new services are to be made available to the general public. One example, presented here, is automating operator services. Others include catalog order entry, credit card verification and repertory dialing. Unfortunately, since current recognition algorithms require that user input be restricted solely to a set of predefined vocabulary items (and sometimes grammatical structures), the human factors issue often become overwhelming. Hence, speech recognizers using even very small vocabulary sets (i.e., 10–20 words) have not been deployed for wide scale use.

In this paper we have presented an algorithm based on hidden Markov model technology, which was shown capable of accurately recognizing a predefined set of vocabulary items spoken in the context of fluent unconstrained speech.

A key point to note is that the assumptions made in this paper are the reverse of those made in previous work on creating a background silence model [16]; namely, that we expected we would have to create a large number of extraneous speech models to obtain high performance. By creating this large number of extraneous speech models we did get good performance on recognition of the vocabulary words in unconstrained speech. Our secondary goal became one of trying to find ways to reduce the number of extraneous speech models to a small value (e.g., one) and still maintain performance close to that of the system with a large number of extraneous speech models. We showed that indeed this was the case and presented two methods for achieving this goal. We never expected the system with a small number of universal extraneous speech models to outperform the system with a large number of such models—our goal was to maintain the performance of the system. We showed that, for a specified vocabulary of five words used for call type routing in operator assisted calls, we could correctly recognize 99.3% of purely isolated speech and 95.1% of the spoken words when they occurred in fluent speech spoken over the long-distance telephone network.

References

[1] J. G. Wilpon, D. M. DeMarco, and R. P. Mikkilineni, "Isolated word recognition over the DDD telephone network—results of two extensive field studies," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (New York, NY), Apr. 1988, 1S.1.10, pp. 55–57.

[2] C.-H. Lee, "Some techniques for creating robust stochastic models for speech recognition," *J. Acoust. Soc. Amer.*, suppl. 1, vol. 82, Fall 1987.

[3] F. Itakura, "Minimum prediction residual principle applied to speech applications," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, no. 1, pp. 67–72, Feb. 1975.

[4] L. R. Rabiner, and S. E. Levinson, "Isolated and connected word recognition—theory and selected applications," *IEEE Trans. Commun.*, vol. COM-29, no. 5, pp. 621–659, May 1981.

[5] J. G. Wilpon and L. R. Rabiner, "On the recognition of isolated digits from a large telephone customer population," *Bell Syst. Tech. J.*, vol. 61, no. 7, pp. 1977–2000, Sept. 1983.

[6] J. G. Wilpon, "A study on the ability to automatically recognize telephone quality speech from large customer populations," *AT&T Tech. J.*, vol. 64, no. 2, pp. 423–451, Feb. 1985.

[7] L. R. Rabiner, and J. G. Wilpon, "Some performance benchmarks for isolated word, speech recognition systems," *Comput. Speech, Language*, vol. 2, no. 3/4, pp. 343–358, Dec. 1987.

[8] R. Thanawala, B. H. Fetz, and R. J. Piereth, "Automatic speech recognition in the public switch network," in *Proc. 5th World Telecom Forum*, vol. 1, pt. 2, 1986, pp. 235–238.

[9] R. W. Christiansen, and C. K. Rushforth, "Detecting and locating key words in continuous speech using linear predictive coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, no. 5, pp. 361-367, Oct. 1977.
[10] A. L. Higgins, and R. E. Wohlford, "Keyword recognition using template concatenation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (Tampa, FL), Mar. 1985, pp. 1233-1236.
[11] C. S. Myers, L. R. Rabineer, and A. E. Rosenberg, "An investigation of the use of dynamic time warping for word spotting and connected word recognition," in *Proc. Conf. Acoust., Speech, Signal Processing* (Denver, CO), Apr. 1980, pp. 173-177.
[12] R. W. Bossemeyer, J. G. Wilpon, C. H. Lee, and L. R. Rabiner, "Automatic speech recognition of small vocabularies within the context of unconstraint input," *J. Acoust. Soc. Amer.*, suppl. 1, vol. 84, Nov. 1988.
[13] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532-556, 1976.
[14] J. Spohrer, P. Brown, P. Hochschild, and J. Baker, "Partial traceback in continuous speech recognition," in *Proc. IEEE Int. Conf. Cybern. Soc.* (Boston, MA), Apr. 1980.
[15] L. Bahl, R. Bakis, P. Cohen, A. Cole, F. Jelinek, B. Lewis, and R. Mercer, "Speech recognition of a natural text read as isolated words," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (Atlanta, GA), Apr. 1981, pp. 1168-1169.
[16] J. G. Wilpon and L. R. Rabiner, "Application of hidden Markov models to automatic speech endpoint detection," *Comput. Speech Language*, vol. 2, no. 3/4, pp. 321-341, Dec. 1987.
[17] A. B. Poritz and A. Richter, "On hidden Markov models in isolated word recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (Tokyo, Japan), Apr. 1986, pp. 705-708.
[18] L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "High performance connected digit recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 6, pp. 1214-1225, Aug. 1989.
[19] C. H. Lee and L. R. Rabiner, "A frame-synchronous network search algorithm for connected word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 11, pp. 1649-1658, Nov. 1989.
[20] L. R. Rabiner, J. G. Wilpon, and B. H. Juang, "A segmental k-means training procedure for connected word recognition based on whole word reference patterns," *AT&T Tech. J.*, vol. 65, no. 3, pp. 21-31, May 1986.
[21] 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, Language*, vol. 1, no. 2, pp. 167-197, Dec. 1986.
[22] K.-F. Lee, "Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system," Ph.D. dissertation, Carnegie-Mellon Univ., Pittsburgh, PA, 1988.
[23] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532-556, 1976.
[24] L. R. Rabiner, B. H. Juang, S. E. Levinson, and M. M. Sondhi, "Recognition of isolated digits using hidden Markov models with continuous mixture densities," *AT&T Tech. J.*, vol. 64, pp. 1211-1234, 1985.
[25] L. R. Rabiner, C. H. Lee, B. H. Juang, and J. G. Wilpon, "HMM clustering for connected word recognition system," in *Proc. ICASSP '89* (Glasgow, Scotland), May 1989, pp. 405-408.
[26] J. G. Wilpon, C. H. Lee, and L. R. Rabiner, "Application of hidden Markov models for recognition of a limited set of vocabulary words in unconstrained speech," in *Proc. ICASSP '89* (Glasgow, Scotland), May 1989, pp. 254-257.

Jay G. Wilpon (M'84-SM'87) was born in Newark, NJ, on February 28, 1955. He received the B.S. and A.B. degrees (*cum laude*) in mathematics and economics, respectively, from Lafayette College, Easton, PA, in 1977, and the M.S. degree in electrical engineering/computer science from Stevens Institute of Technology, Hoboken, NJ, in 1982.

Since June 1977 he has been with the Speech Research Department at AT&T Bell Laboratories, Murray Hill, NJ, where he is a Member of the Technical Staff. He has been engaged in speech communications research and is presently concentrating on problems in isolated and connected word speech recognition. He has written extensively in this field and has been awarded several patents. His current interests lie in training procedures for both speaker dependent and speaker independent recognition systems, keyword spotting algorithms, speech detection algorithms, and determining the viability of implementing speech recognition systems for general usage over the telephone network.

Mr. Wilpon received the 1987 IEEE Acoustics, Speech, and Signal Processing Society's Paper Award for his work on clustering algorithms for use in training automatic speech recognition systems.

Lawrence R. Rabiner (S'62-M'67-SM'75-F'75) was born in Brooklyn, NY, on September 28, 1943. He received the S.B. and S.M. degrees simultaneously in June 1964, and the Ph.D. degree in electrical engineering in June 1967, all from the Massachusetts Institute of Technology, Cambridge.

From 1962 through 1964 he participated in the cooperative plan in electrical engineering at Bell Laboratories, Whippany, NJ, and Murray Hill, NJ. He worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently he is engaged in research on speech recognition and digital signal processing techniques at Bell Laboratories, Murray Hill. He is coauthor of the books *Theory and Application of Digital Signal Processing* (Prentice-Hall, 1975), *Digital Processing of Speech Signals* (Prentice-Hall, 1978), and *Multirate Digital Signal Processing* (Prentice-Hall, 1983).

Dr. Rabiner is a member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, the National Academy of Engineering, the National Academy of Science, and is a Fellow of the Acoustical Society of America.

Chin-Hui Lee (S'78-M'82) was born in July 1951. He received the B.S. degree from National Taiwan University, Taipei, in 1973, the M.S. degree from Yale University, New Haven, CT, in 1977, and the Ph.D. degree from the University of Washington, Seattle, in 1981, all in electrical engineering.

In 1981, he joined Verbex Corporation, Bedford, MA, where he was involved in research work on connected word recognition. In 1984, he became affiliated with Digital Sound Corporation, Santa Barbara, CA, where he was engaged in research work in speech coding, speech recognition, and signal processing for the development of the DSC-2000 Voice Server. Since 1986, he has been with AT&T Bell Laboratories, Murray Hill, NJ. His current research interests include speech modeling, speech recognition, and signal processing.

E. R. Goldman, photograph and biography not available at the time of publication.