AT&T Technical Journal
Vol. 64, No. 2, February 1985
Printed in U.S.A.

A Study on the Ability to Automatically Recognize Telephone-Quality Speech From Large Customer Populations

By J. G. WILPON*

(Manuscript received August 7, 1984)

To ascertain whether a speaker-independent word recognition system, using current technology, could function in normal telephone environments, it was necessary to conduct a study under such real-world conditions. Such an experiment was described by Wilpon and Rabiner (1983), in which telephone customers, speaking under ordinary telephone conditions, in Portland, Maine, were asked to speak their telephone number as a sequence of isolated digits. For each customer a maximum of four digits were obtained. The results from that study were very encouraging and led to further improvements in our recognition systems. To further study the feasibility of implementing speech recognition systems for general use over the telephone network, another field study was initiated. In this test, spoken seven-digit telephone numbers were obtained from a large number of telephone customers over a variety of transmission facilities in Baton Rouge, Louisiana. This paper presents the results of several recognition experiments performed on this database. Experiments were also carried out quantifying the robustness of template sets created in Portland, Baton Rouge, and under laboratory conditions in our Murray Hill, New Jersey, laboratory. Finally, a recognition system that incorporates syntactic information available in a seven-digit telephone is discussed. Our tests indicate a number of distinct real-world problems that must be considered when implementing a speech recognition system for widespread use. A discussion of the overall results and the implications for future research will be given.

*AT&T Bell Laboratories.

I. INTRODUCTION

The development of a speaker-independent speech recognition system that performs well over dialed-up telephone lines has been a goal of AT&T Bell Laboratories for close to a decade.[1–8] However, until recently all evaluations of our recognition systems have been based on laboratory recording conditions. These conditions typically consisted of cooperative subjects using local dialed-up lines over a Private Branch Exchange (PBX). Peak signal-to-average noise ratios under these conditions generally ranged from 40 to 60 dB. Using such local switched lines, the performance of the speech recognition algorithms tested was found to be quite good for a wide range of vocabulary sizes and complexities and for a wide range of talkers.

An earlier effort was made to test the viability of our speaker-independent, isolated word recognition systems on a very large telephone customer population.[8] The task was conducted under "real world" conditions, i.e., asking telephone customers to speak their telephone numbers in a home environment over randomly dialed-up lines in Portland, Maine (PO). Under these conditions, signal-to-noise ratios (s/n) of between 8 dB and 60 dB were encountered. The results of this study yielded a recognition accuracy of 93.1 percent. While this was not as high an accuracy as was achieved in the laboratory,[2] given the transmission medium and the problems associated with obtaining isolated digit strings over standard telephone lines, these results were extremely encouraging.

There were several other shortcomings associated with our previous study. First, the wide variety of transmission and switching conditions made it very difficult to detect the spoken words automatically. Second, for privacy, our database consisted of at most the last four digits of the customer's telephone number. Because of this, parts of the first digit recorded were sometimes deleted. (The digitization of the input speech had to be initiated by a site observer after the first three digits were spoken. In some cases the observer was not quick enough to start the recording procedure before the fourth digit was spoken.) Third, about 50 percent of the speech data available from recording was thrown away, either because it contained some connected digit strings or the background conditions were too severe.

Another problem that existed in our earlier study was getting casual telephone customers to speak their phone number as a sequence of isolated digits. This was related to human factors issues, that is, people do not normally speak in an isolated word format.

As a result of the problems that were encountered during our initial exercise in the "real world," we found that we were testing our recognition systems on only a small percentage of all the speech data to which we had access. The purpose of this paper is to describe a new

data collection exercise that was carried out to more accurately determine our speech recognition system's capabilities over randomly dialed-up telephone lines. Over a two-week period we recorded all customer information from approximately 7400 callers. No calls were eliminated, and all seven digits were recorded.

The database was collected over randomly dialed-up telephone lines at an AT&T centralized switching office in Baton Rouge, Louisiana (BR). The customers that participated would normally speak their telephone number to an operator. That is, the subjects were performing a task that they had done before, except that now input was to be given in an isolated fashion. Special-purpose hardware[9] was attached to one operator console, which automatically answered a call and asked the customer to speak his phone number as a series of isolated digits. The hardware also cataloged the caller's transaction, and digitized and stored the customer's speech on magnetic tape.

There are several very important issues that need to be studied before speech recognition can be made available to large telephone user populations. The most important issue is end-to-end system recognition accuracy. That is, if over time N calls are received by the system and must be handled, what is the percentage of the N calls that will be able to go through the system automatically without any failures? Such failures include the caller hanging up, word endpoint problems, isolated input problems, and the possibility that a human operator would have to intervene during the course of the transaction (e.g., if the customer misunderstood the instructions). These issues are examined in detail within the text of this paper.

Another issue that will be discussed is the robustness of speaker-independent templates created in one recording environment, using one set of talkers, and tested under different conditions with new sets of talkers. In past recognition studies, training data and testing data were collected under laboratory conditions in our Murray Hill, N.J., (MH) laboratory.[1-7] The subjects for these studies were all native speakers of American English mostly from the New York metropolitan area. In our Portland study, the speech data obtained were tested against a speaker-independent template set created from laboratory speech data. The results indicated that the MH template set was inadequate for recognizing speech from Portland customers. With the addition of the Baton Rouge database, more experiments were carried out using speech data from BR, PO and MH. All possible combinations of template sets and testing conditions were tried and the results show the Baton Rouge template set to be quite robust for a wide range of talkers and over a wide range of transmission mediums.

Although past research has shown that isolated word recognition systems perform adequately, the power of speech recognition lies in

its ability to perform a given task reliably, i.e., the word recognizer should be embedded within a larger system. The task can usually be specified as a set of simple rules that define the task syntax. The syntax is able to limit the possible recognition sequences at each point in the transaction. Several task-oriented systems have been described in early work, for example, a voice-controlled repertory dialer system[10] and a directory listing retrieval system.[11] For each of these systems the addition of syntactic constraints greatly increased recognition performance.

Since past studies have shown the additional syntactic information to be useful, a system was constructed that incorporated knowledge about our task, i.e., the speaking of a seven-digit telephone number as a series of isolated digits. A full description of the system syntax and results will be presented.

In Section II, we briefly review the results obtained from the Portland study. Section III gives a description of the recording procedure used to obtain data in Baton Rouge. Section IV discusses the composition of the BR database. In Section V, we review some recent advances in speech recognition, which apply to our study. In Section VI we present the results from a series of recognition experiments performed on the BR database. The issue of template robustness is discussed in Section VII. A discussion of the overall results and their implications is given in Section VIII.

II. REVIEW OF PORTLAND DATA COLLECTION EXPERIMENT

Recordings were made at an AT&T switching office in Portland, Maine.[8] A prerecorded spoken message (a prompt) was given to each customer requesting that he speak his telephone number as a sequence of isolated digits. For reasons of customer privacy we recorded only the last four digits of the telephone number. As each of the digits was spoken, a site observer entered the digits on a keyboard. The observer determined whether the digit sequence was spoken in an isolated format (i.e., spoken with sufficient pauses between words). If not, the observer initiated another prerecorded spoken message (a reprompt) requesting the user to repeat his number with a longer pause between digits. If the observer decided that the final speech was unacceptable (either because it was spoken in a connected manner or because of unacceptably poor telephone line conditions), a reject code was entered and the entire procedure was terminated for the current call.

The recordings were bandpass filtered from 100 Hz to 3200 Hz, sampled at a 6.67-kHz rate, and then digitally transmitted to our laboratory in Murray Hill, N.J., for analysis. The log energy of the waveform was displayed to another observer, along with the automatically determined sets of endpoints indicating where in the recording

interval the isolated words could be found. At this point the second observer had the option of modifying any or all sets of endpoints computed or eliminating any digit from the string. The segmented speech was then entered into a database for later examination. Using this procedure 11,035 digits from 3100 customers were recorded over a 23-day period.

Using a 3900 token subset of the PO speech data to train the recognizer, a 30-template-per-word reference set was created. (Several different template sets were tested in the PO study. The results presented here are for that template set that yielded the highest recognition accuracy.) When this set was tested against the full 11,035-digit database, a recognition accuracy of 93.1 percent was obtained.

There were several problems that occurred during the recording phase. These were classified as being in one of two groups. The first group consisted of problems associated with the telephone transmission conditions, e.g., loud static noises—probably caused by atmospheric disturbances, pops and/or clicks (switching transients), loud tones (mostly carrier frequency tones at 2600 Hz), and loud broadband "humming" noises (probably caused by a missing ground connection somewhere in the transmission path). Resulting peak signal-to-noise ratios varied from as little as 8 dB to as much as 60 dB. The second group consisted of problems related to the talker and the environment in which he or she spoke. These included nonisolation of speech (i.e., the digits were connected) and the presence of extraneous background speech. Most of these failures were severe enough to warrant elimination of the customer's speech from the database. This occurred for 47 percent of all calls available for processing.

As a consequence of the Portland study, several areas for improving recognition performance were discovered. Subsequently, additional research in speech endpoint detection algorithms[12] and clustering algorithms[13] (i.e., template generation procedures) was carried out. Results from this research have been applied throughout our BR study (see Section V).

III. RECORDING PROCEDURE USED IN BATON ROUGE

Figure 1 shows a block diagram of the overall recording setup used in this study. All recordings were made at an AT&T switching office in Baton Rouge, La. To record customer data in an efficient manner special-purpose hardware and control software were required. The hardware included an MC68000 controller, a terminal, a 7-1/2 inch magnetic tape unit, A/D and D/A converters, a cartridge tape unit, and signal conditioning circuitry. This hardware was attached to a dedicated operator console, a full description of which is given in Pirz

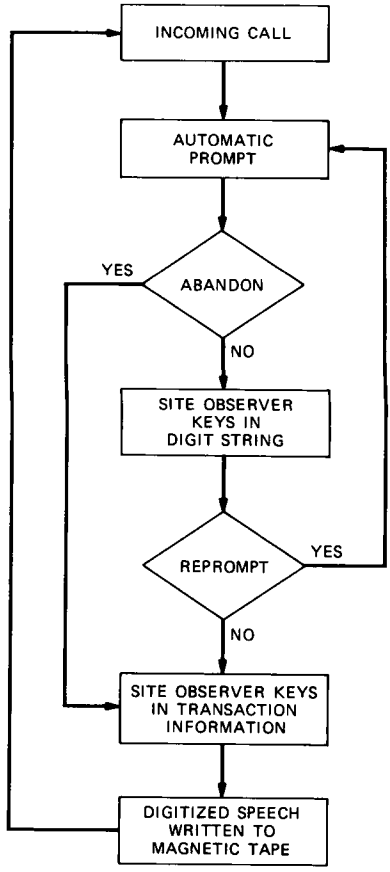

Fig. 1—Block diagram of overall recording system used in the Baton Rouge study.

and Bauer.[9] The sequence of events to record a single customer's speech input was as follows:

1. An incoming customer call was automatically answered by the Special-Purpose Hardware (SPH). A prerecorded prompt was then played to the customer requesting that he speak his seven-digit telephone number as a sequence of isolated digits. As the digits were spoken, a Site Observer (SO) keyed in the identity of the spoken digit string. Also, as the customer spoke, the SPH digitized the speech at a 6.67-kHz rate with appropriate filtering applied.

2. Once the customer finished speaking, the SO made a judgment as to whether the speech was spoken in an isolated format (i.e., with sufficient pauses between words). If not, the SO would initiate a reprompt requesting the customer to repeat his number with longer pauses between words.

3. After the customer completed his task, he or she was given a prerecorded "Thank you" message. The SO then entered an ASCII character string indicating any comments about the talker, such as sex, ability to follow instructions, etc.

4. After the above steps were completed, the digitized speech was written out to the magnetic tape unit. Appropriate header information containing the identity of the input speech string, date and time of utterance, and any comments entered by the SO was also recorded on tape. If a reprompt had to be made, both the original and reprompted speech were saved.

It should be noted that a significant number of customers abandoned their call without speaking their phone number. All abandoned calls were cataloged and will be discussed later.

Using this procedure we recorded data from 7373 subjects (on 33 magnetic tapes) over a two-week period. After the data collection was completed, the speech was read from the magnetic tapes into a Data General MV8000 minicomputer where all further analysis was performed.

IV. COMPOSITION OF FINAL DATABASE

Recordings were made for an average of six hours a day, five days a week, for two weeks. Tables I and II show a detailed analysis of the final telephone customer database. Data were collected from a total of

Table I—Statistics on total number of calls handled in BR study

	Number of Callers	Percent
(1) Total callers	7373	100
(2) Abandoned calls	1468	20
(3) Net total calls (1–2)	5905	80
(4) Operator intervention	2301	31
(5) Unidentifiable calls	518	7
(6) > 7 digits spoken	269	4
(7) Processable calls (3–4–5–6)	2817	38

Table II—Sex makeup of processed calls from BR data

	Number of Calls	Percent
Net total calls	5905	100
Adult male	2137	36
Adult female	3524	60
Children	168	3
Unclassifiable	76	1

7373 callers. Of this total, 1468, or 20 percent were callers that abandoned their calls, and therefore did not enter any speech data. In these cases the caller hung up before beginning the recording task. This leaves a net total of 5905 calls that yielded some speech output. Of the remaining callers, 2137 (36 percent) were adult males, 3524 (60 percent) were adult females, 168 (3 percent) were children, and 76 (1 percent) were unclassifiable. Of the 5905 useful calls, 2301 or 31 percent required the telephone operator to cut in during the middle of the recording transaction. In these cases, the user got confused about the task he was to perform or simply did not want to cooperate. Since the caller had to supply his telephone number to complete his telephone call, the operator had to intervene. Generally, the user gave his phone number in a continuous fashion to the operator. Therefore, no useful isolated data could be extracted from these callers. There were several calls (518, or 7 percent) where the SO and later another observer could not understand one or more of the words that were spoken and therefore could not tag them correctly. These were caused either by very bad transmission noises or a very pronounced accent. For another 269 calls (4 percent), the customer spoke more than seven digits.

If we take into account all the calls that had some problems associated with them we would be left with a total of 2817 calls (38 percent) that were "processable," i.e., these calls contained only spoken digits. Therefore, an automatic procedure could be devised to first find the spoken words (endpointing) and second perform recognition on those words. All further discussion of the BR database will refer to this data set.

One problem that existed in our earlier recognition experiment of telephone-quality speech was the inability of the prompts to get the callers to speak in an isolated format.[8] This problem still existed in the BR study. Therefore, we decided to segment the database into two sets—one where all the digits were spoken in isolation, and another that contained those calls with any connected digits. Of the 2817 processable calls, 1837 contained only isolated digits, and 980 contained some connected digits.

The recording hardware had memory for at most a 15-second utterance. Some callers paused so long in between digits that they simply ran out of time. Therefore, we further classified the calls on the basis of whether all seven digits were present. The reason for this classification was that if it were known a priori that exactly seven digits were present, we could devise a procedure that recognizes them more accurately. Of the 1837 calls containing only isolated digits, 1634 (89 percent) consisted of all seven digits. Of the 980 calls containing some connected speech, only three calls had fewer than seven digits.

V. REVIEW OF RECOGNITION SYSTEM IMPROVEMENTS

5.1 Endpoint detector improvements

Before evaluating this telephone-quality speech database, several issues had to be addressed. One issue was the detection of speech within some time interval. In our previous study,[8] we used an endpointing algorithm developed by Lamel et al.[14] This technique had proved quite robust in detecting speech over local dialed-up telephone lines. However, endpoint detection becomes a much more difficult problem when the transmission system is corrupted by the many noises found on standard dialed-up telephone lines (such as those received at TSPS offices). Such factors as popping sounds, crackling noises, background speech, carrier frequency tones, and other nonstationary noises make it very hard to detect word boundaries accurately.

In Wilpon et al. it was shown that only 69 percent of all words were detected by the Lamel approach when tested on a large random subset of the PO database.[12] Among these, the recognizer accurately classified 85 percent. This yielded an overall recognition system accuracy of only 59 percent. Because of these results, it was decided to try and improve the endpoint algorithm before proceeding further. This led to the development of a new word detection algorithm, called a top-down design.[12] The new approach makes the assumption that if speech is present in some time interval its energy level will be above that of any noise also present. Simply put, the new algorithm searches for strong (vowel-like) peaks in the energy contour of a speech utterance and processes the speech around the peaks to find potential beginning and ending points. Several rules involving duration, onset, and decay times are then used to refine the endpoint estimates.

Applying this new endpoint algorithm to the same data set as was tested with the Lamel algorithm (i.e., a subset of the PO database) yielded a word detection rate of 98 percent and a recognition accuracy of 90 percent for an overall system accuracy of 89 percent. Clearly, from the results obtained, the top-down endpointing algorithm is superior to the Lamel approach. As a result of this research, the top-down algorithm was used in all studies of the BR database. Whereas in the PO database study over 50 percent of the database had to have manual corrections made to the endpoints (because of endpoint algorithm failures) *no* manual correction of endpoints was performed in the BR study.

5.2 Clustering analysis improvements

In Wilpon and Rabiner[13] a new clustering algorithm was presented that uses the best features of several previously used algorithms—i.e., ISODATA[2,15], K-means[2,15], and UWA[6]. This algorithm is called the

Modified K-Means (MKM) clustering algorithm. Its main advantage over other algorithms is that it is completely automatic, and requires no user input (other than a similarity matrix). This algorithm was tested extensively on the BR database and was shown to yield recognition results as good as previously used algorithms. In the experiments to follow, all template sets created from subsets of the BR speech data will have been created using the MKM algorithm.

VI. RECOGNITION RESULTS ON THE BR DATABASE

6.1 Isolated word recognition results

For all isolated word recognition experiments performed on the BR database, the isolated database was divided into two disjoint subsets, one to train the recognizer and the other to test the system. A total of 4783 tokens were used for training and another 7973 tokens for testing. Table III shows the distribution of the training and testing tokens among the ten digits.

In the PO study, template sets were created both from a random subset of the speech data and from the "cleanest" speech data (i.e., as judged by a human to be close to laboratory-quality data). Similar recognition results were obtained using both template sets. Therefore, in the BR study template sets were created using only a random subset of speech data.

The recognition system used in all evaluations was the Linear Predictive Coding (LPC)-based isolated word recognition system developed and tested extensively at AT&T Bell Laboratories.[1-7] As we stated in Section 5.2 the MKM clustering algorithm was used to create several sizes of template sets. Table IV shows the recognition results for seven different clustering configurations: 3, 6, 12, 20, 30, 50, and 75 clusters per word. Shown is the per-digit accuracy, average digit

Table III—Number of tokens for each digit used in training and testing for evaluating the BR database

	Training Set	Testing Set
0	271	312
1	259	405
2	675	1145
3	606	943
4	580	970
5	489	901
6	592	1070
7	443	750
8	454	834
9	414	643
Total	4783	7973

Table IV—Recognition results in percent using BR speech data for training and for testing

	Number of Templates per Word						
Digit	3	6	12	20	30	50	75
0	60.5	71.0	70.7	72.8	69.6	72.8	72.8
1	85.6	86.1	85.6	88.0	88.2	88.5	87.7
2	64.2	67.5	74.5	76.8	78.2	81.5	82.6
3	74.2	84.0	87.5	89.6	91.3	89.9	90.4
4	88.4	92.4	93.5	92.6	95.4	94.2	92.9
5	78.9	78.9	86.4	85.0	87.8	89.1	89.5
6	63.3	75.1	79.8	80.9	79.3	87.7	84.9
7	62.8	74.7	80.2	84.9	88.4	88.4	90.1
8	71.4	70.3	75.8	80.1	81.2	83.3	83.3
9	58.8	69.8	63.4	70.9	74.9	75.6	80.6
Average	71.0	77.0	80.5	82.7	84.4	86.1	86.3
String rate	16.3	24.2	29.5	34.1	36.0	42.2	43.5

accuracy over all digits, and string accuracy, where a string is nominally seven digits long. We see that for all template sizes the digits zero (or oh) and nine have the highest error rates. The major confusion for the digit zero (oh) was the digit four. In the BR and PO studies about one-half of the talkers pronounced that word four as /foe/ rather than /fawr/ and used the word oh instead of zero more than 75 percent of the time. A possible explanation for the confusion could be that endpoint detector included too much background noise when determining the beginning point for some of the pronunciations of the word oh, thereby making the word oh like a /foe/ and misrecognizing it. Alternatively, since the frication at the beginning of the word four closely resembles typical background noise encountered in our testing environment, it would be easy for the speech endpoint detector to misplace the beginning marker for this word, thus totally eliminating the fricative sound. Since templates for the word four are created from this type of data, a spoken digit oh could be misrecognized. Low accuracy for the digit nine was also obtained. A possible reason for such low accuracy is that the nasal sound is being masked by the various noises on the telephone line. Figure 2 shows a plot of digit recognition accuracy as a function of the number of templates (or clusters) created for each word. We can see that as the number of templates per word increases, the recognition accuracy increases asymptotically, with the best accuracy (86.3 percent) occurring with a 75-template-per-word set.

The average string length was seven digits. Therefore, theoretically the average string accuracy is the average per-digit accuracy raised to the seventh power (since all single digit recognitions are independent of one another). However, for all template set sizes the actual string accuracy was greater than the theoretical result, that is, the error rate

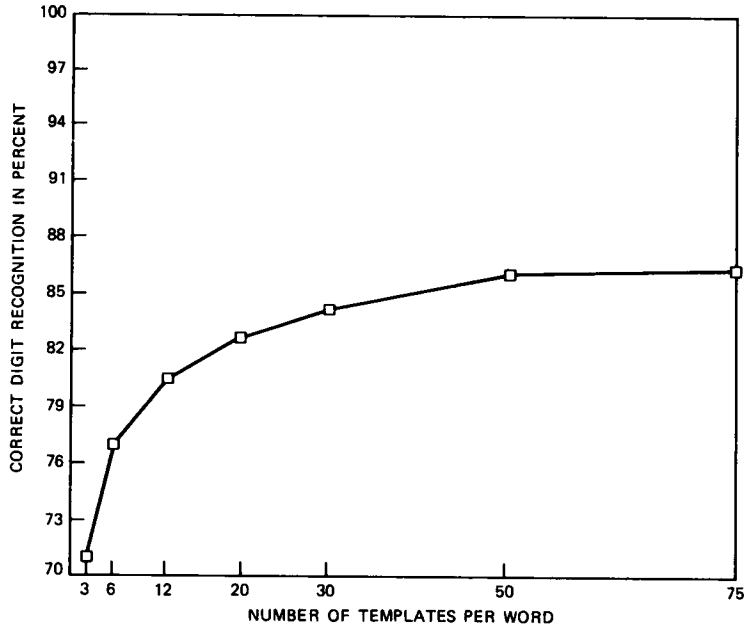

Fig. 2—Recognition accuracy for training and testing on BR data as a function of the number of templates used per word.

was not uniform over all talkers nor independent of the talker. A similar result on another speech database was obtained by Rosenberg and Shipley.[16]

Figure 3 demonstrates the effects of imposing a rejection threshold on the recognition system. A recognition distance score above the threshold would result in a no decision choice by the recognizer. Shown in this plot is the percent of no decisions versus the percent of error rate. We see that if only a 1-percent error rate could be tolerated by a task using this recognizer under these recording conditions, then a 60-percent no decision rate must also be accepted. However, a 10-percent probability of error was attained with only a 9-percent no decision rate.

If we compare these results (using a 30-template-per-word solution for comparison) with those obtained from the PO database study, the results from the BR study seem to be worse (84.4 percent for BR versus 93.1 percent for PO). However, in the PO study 50 percent of the speech data available for testing was eliminated from the database because of noise conditions, connected rather than isolated input, and hardware failures. Also, the automatic endpoint detector[14] was overruled by human intervention about 50 percent of the time.[8] In contrast, in the BR study all the data available were used and automatically

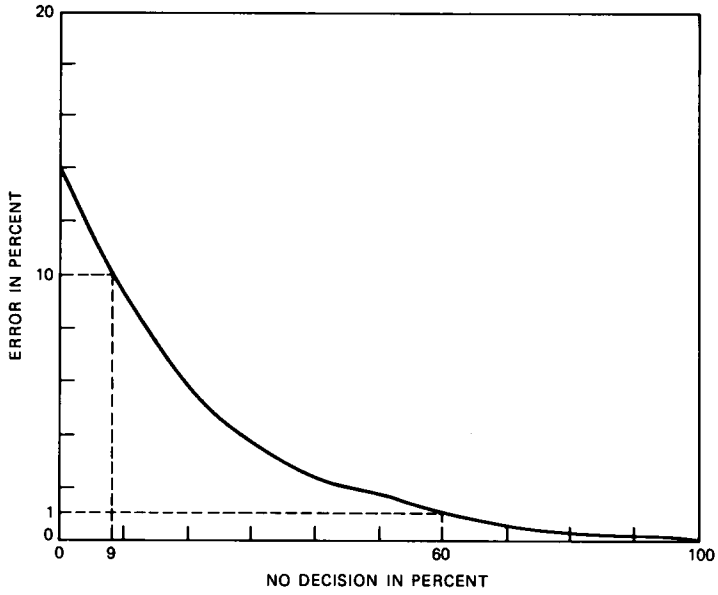

Fig. 3—Plot showing recognition error rate versus no decision choice—training and testing data from BR.

detected before recognition was performed.[12] For this reason it is felt that the BR results are very encouraging.

6.2 Connected word recognition results

Recognition was performed on the 980 call subset of the processable calls, which contained some connected digit sequences using the level-building Dynamic Time Warp (DTW) algorithm of Myers et al.[17] Testing was carried out with and without augmenting the Itakura log likelihood distance with an energy distance as described in Rabiner.[18] The template set used was the 30-template-per-word set created from isolated tokens from the BR database. No embedded training, as described in Rabiner et al.[19], was used.

Since there was not an abundance of connected digit strings within the 980 strings (only 1790), I will just present the results and not make any categorical remarks on connected digits input over noisy channels. Table V shows the results of these experiments. The results indicate that using energy information in the distance computation improved the string accuracy for all string lengths from 45.8 to 61.0 percent if the string length is unknown, and from 63.6 to 66.8 percent if the string length is known. It is expected that the use of embedded training would greatly improve these results.

Table V—Recognition results in percent from connected digit sequences from BR speech data

No. of Digits in String	No. of Occurrences	Percent Correct Recognition Without Energy		With Energy	
		Known Length	Unknown Length	Known Length	Unknown Length
2	1441	68.5	49.4	71.1	65.3
3	277	46.6	33.6	51.6	45.9
4	67	32.8	22.4	38.8	32.8
5	5	20.0	0.0	60.0	40.0
Total	1790	63.6	45.8	66.8	61.0

6.3 A syntax-directed recognition system based on isolated digit input

The results described previously assume that all single digit recognitions are independent of each other. But in fact that is not the case for this database, as customers were asked to speak their seven-digit telephone number. For this well-defined task there is some syntactic information that can be used to help guide the recognition system. For example, the first three digits of the seven-digit input define the local exchange. In general, there are significantly fewer than the 1000 exchanges within an area-code region. However, the last four digits are usually distributed uniformly over the 10,000 possible sequences.

A recognition system was assembled to make use of the syntactic structure of telephone numbers. First, the database was searched to find all valid exchanges. This yielded a total of 86 valid exchanges out of a possible 1000. The recognition system was then programmed to do the following task. For each customer, digit recognition was performed on the exchange. The output of the recognition system was a set of similarity scores[1] for each digit, for all digits in the exchange. Next, the customer's actual utterances for the exchange were tagged as being that valid exchange with the lowest total distance (i.e., the sum of the individual digit scores). The utterances were then converted into speaker-dependent templates and added to the previously created speaker-independent template set. This new template set was then used to recognize the last four digits in the telephone number. This procedure was done on a per-talker basis. If in the last four digits the customer spoke any of the digits that were in the exchange, having a template of those words created by the user should increase the probability that the recognizer would correctly recognize those words.

Whereas the recognition accuracy (for a 30-template-per-word reference set) yielded a digit accuracy of 84.4 percent when the above system was implemented, the digit accuracy increased to 87.2 percent. The string accuracy without syntax, as tested on 984 seven-digit

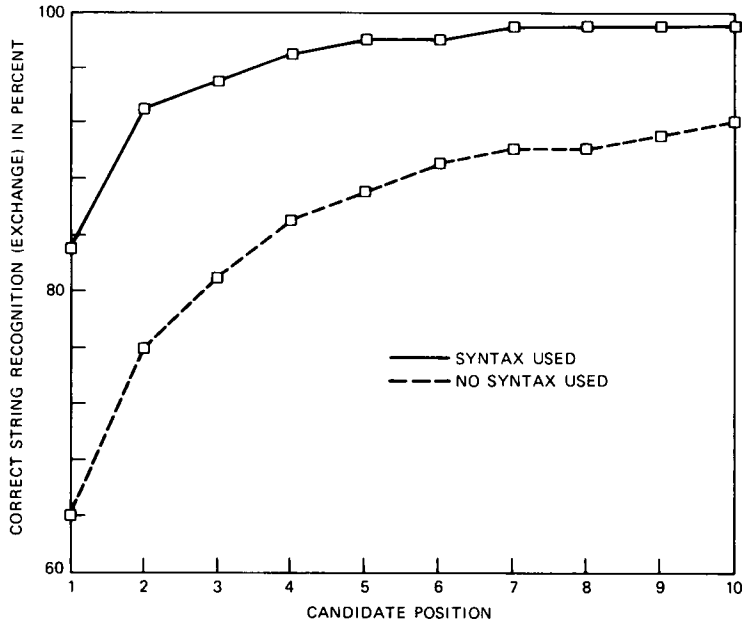

Fig. 4—Recognition accuracy on the exchange string (three digits) from the BR database as a function of candidate position (solid line is with syntax and dashed line is without syntax).

strings, was 38.2 percent. This increased to 47.8 percent when syntax was added.

Figure 4 shows a plot of exchange recognition accuracy as a function of whether the correct exchange was within the top ten candidates. The dashed line shows the results when no syntax is used, and the solid line shows the results when syntax is used. We see that the correct exchange (all three digits) has been recognized correctly in the system with no syntax 64 percent of the time and 87 percent within the top five candidates, whereas in the syntax-directed system the results are 83 percent and 98 percent, respectively. Figure 5 shows a plot of the number of times the correct exchange is within a distance Δ from the minimum possible exchange score (over the 1000 possible exchanges). For the Itakura log-likelihood ratio distance the mean distance for a correct recognition is about 0.30 and for an incorrect recognition about 0.45.[2,8] We see that within a distance of 0.25 from the minimum the correct exchange (three digits) is always present. Figure 6 shows the average number of possible exchanges within a Δ region over all strings. It shows that using syntax greatly reduces the number of recognition candidates (e.g., from 80 to 10 for a $\Delta = 0.20$).

Figure 7 shows a plot of the string accuracy of the last four digits as

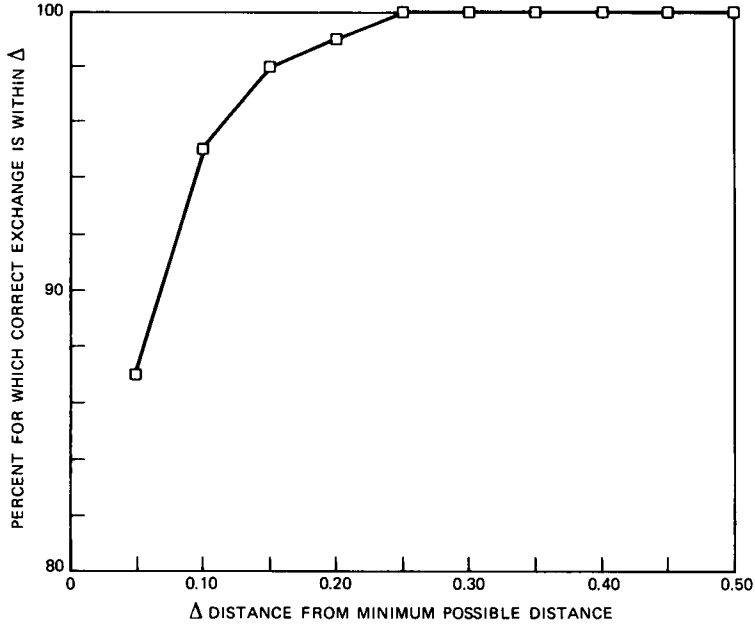

Fig. 5—Recognition accuracy on the exchange string (three digits) as a function of whether the correct exchange is within a Δ distance from the minimum possible exchange score.

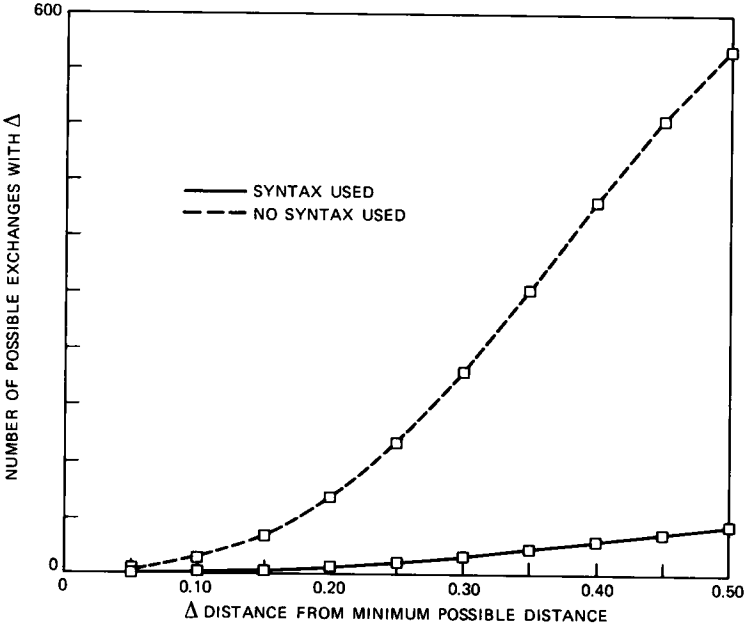

Fig. 6—Plot showing the average number of exchange candidates within a Δ region.

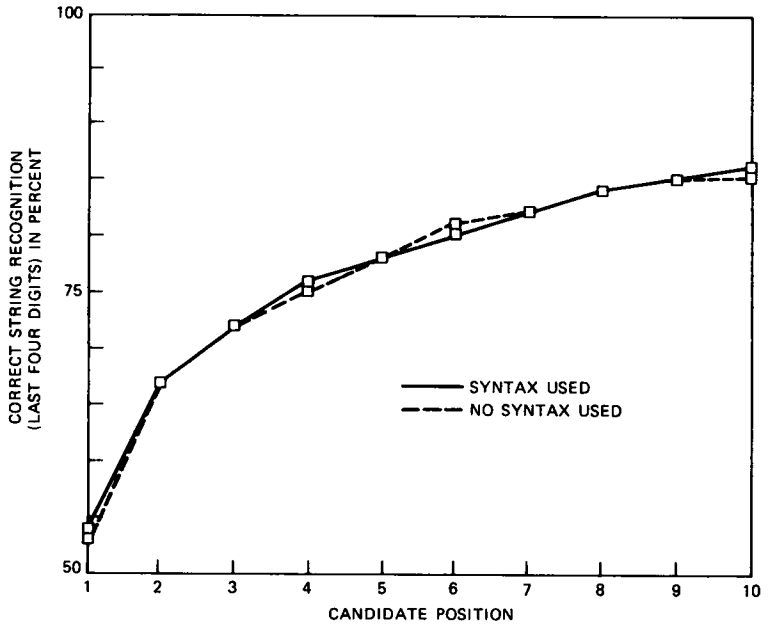

Fig. 7—Recognition accuracy on the last four digits from the BR database as a function of candidate position. The dashed line indicates no syntax was used and the solid line indicates syntax was used.

a function of whether the correct string was within the top ten candidates. The dashed line shows the results without using the additional templates generated by applying the syntactic rules to the first three digits. The solid line shows the recognition results when the original speaker-independent template set was augmented with the speaker-dependent templates determined through the syntactic rules on the first three digits. With syntax the string accuracy only improved from 53.3 to 54.5 percent and was within the top five candidates 78 percent of the time.

These results show that adding task information improved the overall system recognition accuracy. Augmenting the template set with the extra exchange utterance templates slightly improved performance on the last four digits. However, the main contribution of the syntax was to recognize the exchange more accurately.

VII. ROBUSTNESS OF SPEAKER-INDEPENDENT TEMPLATES

One of the goals of this study was to examine the robustness of speaker-independent templates created using one population of talkers under one set of transmission conditions for different populations and transmission conditions. An experiment was carried out in which

template sets created from each of the three regional databases (Portland, Baton Rouge, and Murray Hill) were tested on speech data from all three databases.

Initially, a template set was created from a Murray Hill speech database that consisted of 100 talkers, 50 male and 50 female. The data were collected under laboratory conditions over local Private Branch Exchanges (PBXs). A clustering analysis was performed and a set of 12 speaker-independent templates per word was created. This template set has been tested extensively in other experiments (see Refs. 2, 8, 11, 14, 16, and 17). For testing purposes, another group of 100 talkers (disjoint from the training population) each provided one replication of the digits vocabulary.

The template set used to represent Portland data was a 30-template-per-word set created from the "cleanest" speech obtained in the PO study.[8] For testing the entire 11,035-digit database was used. For comparison purposes a 30-template-per-word reference set was used to model the Baton Rouge database. For testing purposes the entire 7973-digit testing set was used.

Table VI shows the results for all cross recognition tests. The symbol <AVG> stands for the averaging of recognition results over all three databases given a particular training or testing dataset. In order not to distort the averages (since each database had a different number of tokens), a simple nonweighted averaging was performed. It is felt that there was sufficient data in each regional database to make this result meaningful.

Figure 8 (a graphical form of Table VI) shows the results when

Table VI—Cross template and testing set recognition accuracy

Training Set	Testing Set	Recognition Accuracy in Percent
BR	BR	84.4
BR	PO	85.8
BR	MH	92.3
PO	PO	93.1
PO	BR	76.8
PO	MH	91.6
MH	MH	98.4
MH	PO	77.4
MH	BR	62.3
BR	<AVG>	87.4
PO	<AVG>	87.1
MH	<AVG>	79.3
<AVG>*	BR	74.3
<AVG>	PO	85.4
<AVG>	MH	94.1

* <AVG> = averaged over all three databases

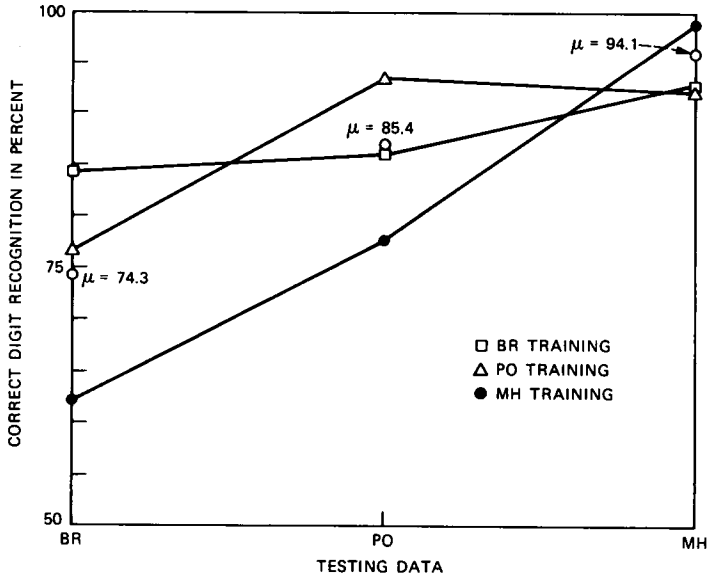

Fig. 8—Recognition results for each regional template set as a function of the testing set.

testing each of the template sets against each of the testing datasets, where μ is the average recognition accuracy over all three testing sets given a training set. Shown is recognition accuracy as a function of which testing set was used. This figure shows that the best recognition results for each of the test sets occurred, not surprisingly, using the template set also created from data in the same region. The recognition performance was best with MH data, then with PO data, and last with BR data. Also, notice the greater variation in recognition accuracies as the self-recognition scores decline. For example, the PO and BR templates performed about the same against MH testing data and about 7 percent worse than MH templates, whereas the PO and MH template sets performed, respectively, 8 and 21 percent worse than the BR template set when tested with BR data.

In Fig. 9, the results are shown as a function of individual digits. As was the case in the earlier PO study, we see that the MH templates do not adequately represent the speaking style or noise conditions present in the PO or BR testing data. For most of the digits in the BR and PO testing population the templates from PO and BR yielded much better results.

Figure 10 shows the results in a different context. Shown is recognition accuracy as a function of the template set used. Interestingly, the BR template set performed the best over all three testing conditions. (Even though the same average accuracy was obtained with the

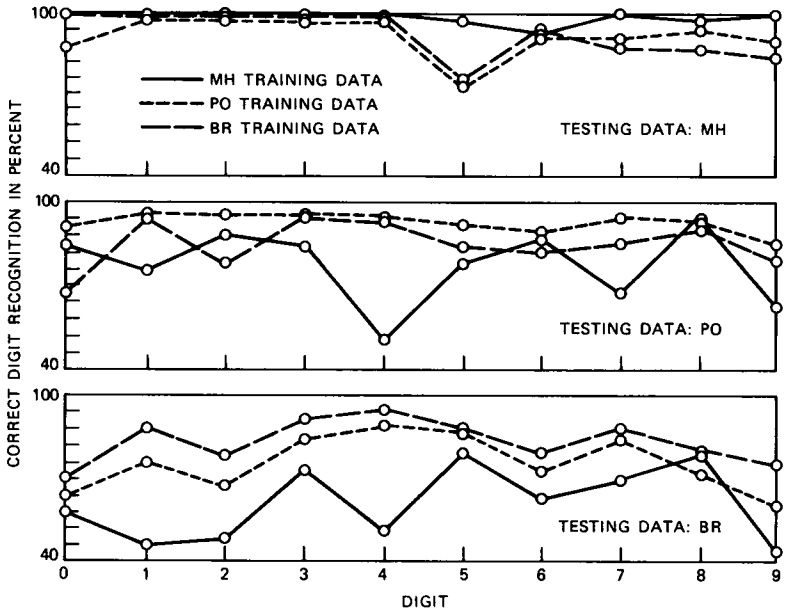

Fig. 9—Recognition results for each regional template set as a function of testing set, on a per-digit basis.

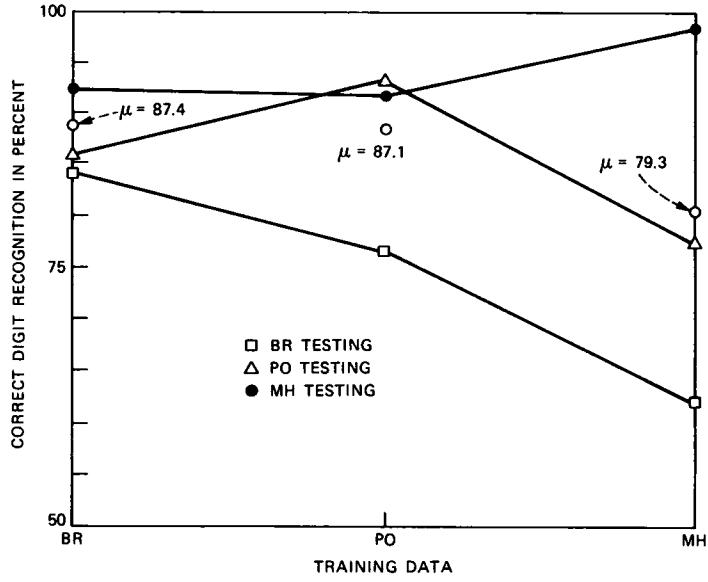

Fig. 10—Recognition accuracy for each regional testing set as a function of the template set used.

BR and PO templates, the BR templates yielded a standard deviation of 4.3 percent compared with 9.1 percent for the PO template set.) The MH template set performed significantly worse than either the BR or PO template sets, with an average recognition accuracy of 79.3 percent and a standard deviation of 18.1 percent.

Figure 11 shows the per-digit recognition results for each testing data set given a particular template set. For most digits in the MH testing set, the template sets created from BR and PO data yielded as good a recognition accuracy as did the templates created from MH data. However, again we see the converse not to be true, that is, the MH templates yielded significantly poorer recognition results when tested against PO and BR testing data than did template sets created from those regions.

Tables VII through IX show confusion matrices generated from each of the above recognition experiments. Shown are only those confusions that occurred more than 3 percent of the time. In Table VII, results are shown for each testing set when recognition was performed using the BR template set. In each of the BR and PO testing sets the biggest error was the spoken word oh being confused for four. Possible explanations for the confusions have been given in Section 6.1. This problem did not occur in the MH data as all talkers

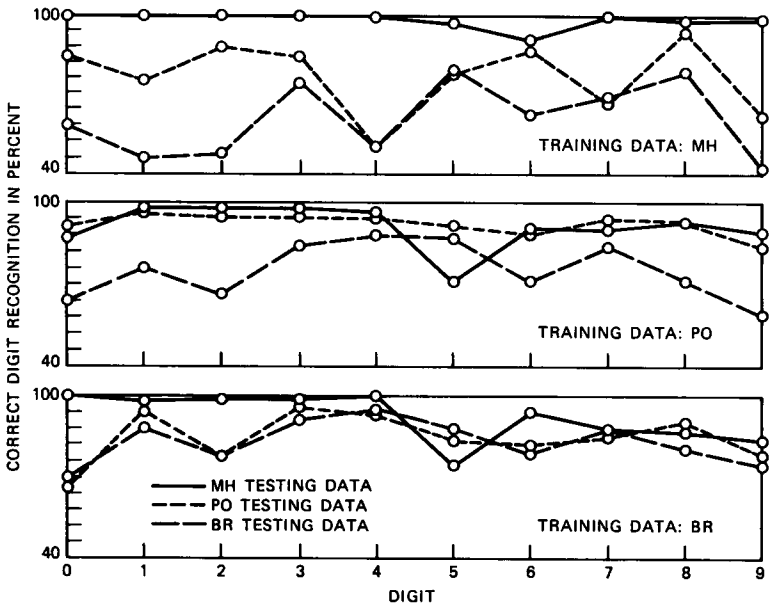

Fig. 11—Recognition accuracy for each regional testing set as a function of the template set used, on a per-digit basis.

Table VII—Confusion matrix for each testing set when recognition was performed using the BR template set

Spoken Digit	Recognized Digit									
	0	1	2	3	4	5	6	7	8	9
					(a) BR Testing Data					
0	69.6				15.6			5.1		
1		88.2			5.3					
2	4.6		78.2				4.7	5.4		
3				91.3			3.0			
4					95.4					
5						87.8				6.9
6				4.2			79.3		11.6	
7								88.4		
8				7.5			5.6		81.2	
9		3.1				12.8		4.6		74.9
					(b) PO Testing Data					
0	66.3				27.7					
1		94.2								
2	8.1		77.8		8.1			4.5		
3				95.7						
4		3.3			93.6					
5		6.8				84.0				5.5
6							82.4	6.8	5.1	
7								85.5		5.0
8							5.8		90.3	
9		4.4		4.1		8.9				79.0
					(c) MH Testing Data					
0	100.0									
1		98.0								
2			99.0							
3				99.0						
4					100.0					
5		4.0			5.0	75.0		4.0		11.0
6			3.0				94.0			
7			10.0					87.0		
8	3.0		3.0	3.0			4.0		87.0	
9		9.0		5.0						84.0

used the word zero. Notice that the reverse confusion (i.e., four misrecognized as zero or oh) did not occur.

Table VIII shows the confusion matrices when using the template set generated from MH data. The only confusion when tested against MH testing data was six versus seven. Notice when testing against BR data that all digits are misrecognized as seven a large percent of the time. For this testing data there are many major confusions.

Table IX shows the confusions generated from training with PO data. As with the other template sets the five-nine confusion is prominent. Also, the MH testing set produced a large confusion between the digits nine and one. In examining Tables VII through IX the template set generated from BR speech data yielded fewer major

Table VIII—Confusion matrix using template set generated from MH data

Spoken Digit	Recognized Digit									
	0	1	2	3	4	5	6	7	8	9
					(a) BR Testing Data					
0	58.4		11.8				4.7	11.8	6.5	
1	7.8	46.3			6.4	20.1		4.0		8.3
2			47.9	3.3			7.0	19.4	19.9	
3				74.2			7.0	5.4	7.5	
4	28.6	3.1			50.6	8.5		3.8		
5	3.3					79.6	7.4	4.2		3.9
6				4.0			62.6	5.0	24.9	
7			4.6			3.0	12.3	69.4	5.1	3.0
8				7.4				4.7	79.2	
9						26.7	14.7	10.2		43.2
					(b) PO Testing Data					
0	84.3		5.6					3.3		
1	6.3	72.9				7.8			4.3	6.3
2			86.4					6.0		
3				87.7					5.4	
4	34.0				48.7	8.8				
5	3.2					80.4	5.6			6.5
6							85.8		7.5	
7						3.2	9.2	74.8	3.0	3.2
8									95.3	
9				4.4		12.6	12.3	3.9		64.3
					(c) MH Testing Data					
0	100.0									
1		100.0								
2			100.0							
3				100.0						
4					99.0					
5						97.0				
6							91.0	6.0		
7								100.0		
8									98.0	
9										99.0

confusions (over all three test sets) than either the PO or MH template sets.

Summarizing this experiment, the template set created from a subset of BR speech data was quite robust over different populations and noise conditions, yielding an average recognition accuracy of 87.4 percent. Additionally, the MH template set, which was created under laboratory conditions, provided poor recognition results when tested under "real-word" recording conditions.

In a final experiment the template sets created from the PO, BR, and MH data were combined together to form one large template set with 72 templates per word (i.e., 30 templates per word from each of the PO and BR sets, and 12 templates per word from the MH template set). The testing set for this experiment was the combined testing sets from PO, BR, and MH.

Table IX—Confusion matrix generated from training with PO data

Spoken Digit	Recognized Digit									
	0	1	2	3	4	5	6	7	8	9
					(a) BR Testing Data					
0	63.8		12.5		5.7		3.6	6.8		
1		76.2			13.4	5.6				
2			66.8	5.4			12.7	4.2	8.0	
3				84.6			3.8			
4	5.0		3.4		88.5					
5						87.7		4.7		3.1
6				5.8			72.1	3.5	15.5	
7							5.8	84.9		
8				10.0			7.9	5.6	72.5	
9						19.4		12.5		60.1
					(b) PO Testing Data					
0	90.9									
1		95.2								
2			95.0							
3				94.2						
4	3.8				93.4					
5						93.1				3.0
6							87.0	3.2	4.2	
7								93.3		
8									95.3	
9						9.2				85.0
					(c) MH Testing Data					
0	87.0		9.0		4.0					
1		98.0								
2			98.0							
3				98.0						
4			3.0		97.0					
5					8.0	72.0		3.0		14.0
6			4.0				91.0	5.0		
7			7.0					91.0		
8				3.0					94.0	
9		7.0								90.0

A recognition accuracy of 90.9 percent was achieved under these conditions. In examining a histogram of template usage, several templates were used more often for incorrect recognitions than for correct recognitions. A test was carried out in which template sets were created as subsets of the full 72-template-per-word set. These subsets were chosen such that the Net Percent Correct Recognition (NPCR) per template, as defined over all three testing sets (i.e., the percentage of the time that the template was used for a correct score minus that when used incorrectly), was greater than a threshold. Figure 12 shows a plot of the total number of templates used (dashed line) and the recognition accuracy (solid line) as a function of the threshold. As indicated by the results, 20 templates of the original 720-template set yielded only incorrect recognitions. Also, most templates yielded a NPCR of 50 percent. As we look for a NPCR of greater than 50

percent, the number of templates that qualify goes down exponentially. Only 144 of the original 720 templates yielded an NPCR of 100 percent.

The recognition curve (solid line) shows a similar shape. We see that recognition accuracy stays constant for NPCRs of less than 80 percent, then falls rapidly to 68 percent for an NPCR of 100 percent. These two curves show that the total number of templates can be reduced by 22 percent from 720 to 560 (or an average of 56 templates per word) without reducing the overall recognition accuracy.

Table X shows a confusion matrix for the combined template set (only entries greater than 3 percent are shown). The results indicate that this template set yielded fewer confusions than did either of the individual template sets, with the major confusions being zero (oh)-four, nine-five, and six-eight.

VIII. DISCUSSION

The results described in earlier sections show that:

1. Based on the collection of speech data in Portland and Baton Rouge, it is clear that significant problems exist prompting casual telephone customers to speak digit strings in an isolated format.

2. One problem that existed in the Portland study was the inability to detect words automatically in nonideal environments. With the use of the top-down endpoint detection algorithm,[12] this problem was greatly reduced in the Baton Rouge tests.

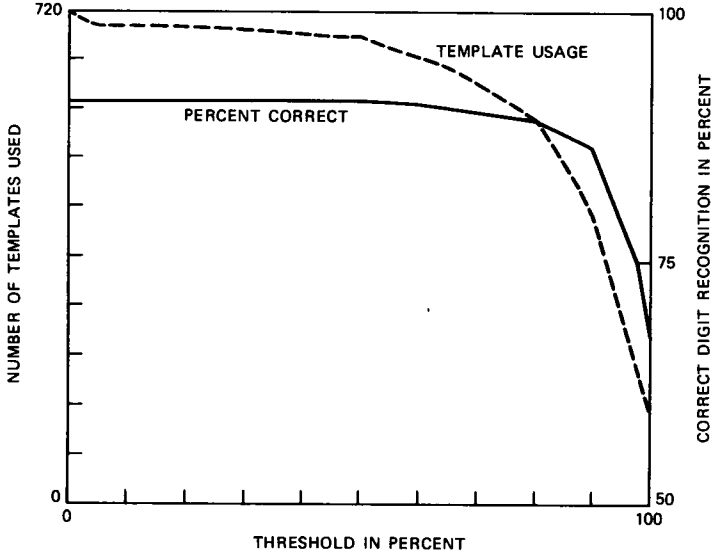

Fig. 12—Plot showing recognition accuracy and number of templates used as a function of template use threshold.

Table X—Confusion matrix—combined template set versus all testing data

Spoken Digit	Recognized Digit									
	0	1	2	3	4	5	6	7	8	9
0	86.5				5.6					
1		94.5								
2			90.7							
3				93.9						
4					94.4					
5						92.5				3.4
6							85.7		7.1	
7								92.0		
8							3.3		90.5	
9						7.5				84.2

3. The recognition results obtained from the BR tests were worse than those obtained in the Portland study (84.2 and 93.1 percent, respectively, on comparable sized reference sets). Since in the Portland experiment 50 percent of all data was eliminated from testing and 50 percent of the remaining data needed human interaction to correct endpoint failures, we feel the results from the Baton Rouge study, which eliminated no data and did not allow for endpoint corrections, more accurately demonstrate our current capabilities.

4. The addition of syntactic constraints on the isolated word recognizers output increased the overall recognition system accuracy—from 84.4- to 87.2-percent digit accuracy and from 38.2- to 47.8-percent string accuracy (i.e., seven-digit telephone number).

5. The template set created from a subset of BR data is quite robust over different populations and noise conditions, averaging 87.4 percent over the three regional data sets. By creating a combined template set based on the templates generated from speech data from Portland, Baton Rouge, and Murray Hill, a recognition accuracy of 91 percent was obtained when tested on 20,000 tokens of PO, BR, and MH data.

These results are very encouraging, as they indicate that regional "speaker-independent" template sets may not be required to obtain the highest recognition accuracy possible over all regions. However, since for each regional database the best recognition scores occurred using training and testing data from that region, having regional templates will improve accuracies in the individual regions.

To compute the end-to-end recognition system performance number, several intermediate results must be combined together. Figure 13 shows the combination of all steps in the recognition system. Starting with all calls that were handled by our recognition system, initially 20 percent abandoned the transaction. Of the 80 percent of calls remaining, 52 percent required some form of operator assistance to complete the call and 17 percent contained some connected input.

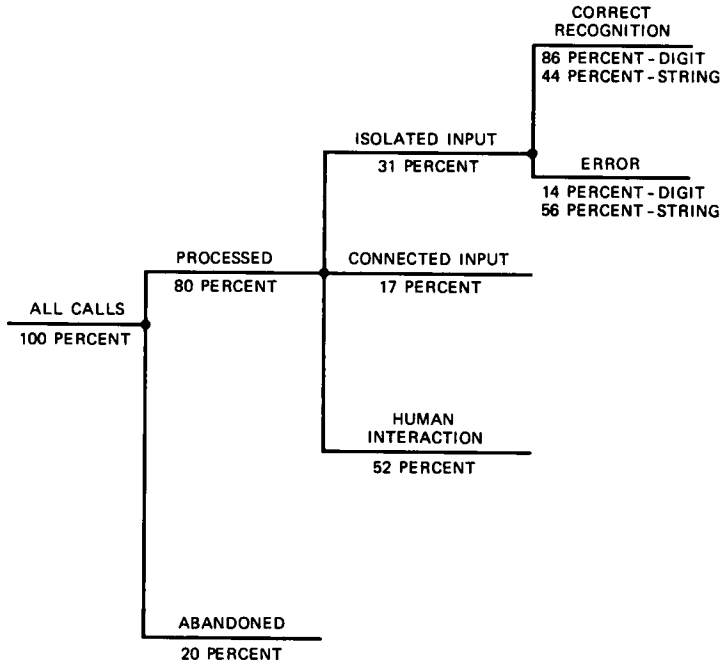

Fig. 13—End-to-end recognition system performance from isolated digit input.

This left only 24 percent of the calls consisting solely of isolated digit input. Therefore, we can see that the end-to-end system performance was 21.3 percent on digits and 11 percent on strings (where a string consisted nominally of seven isolated digits). The full recognition system was able to handle automatically 11 percent of *all* calls received. If such a system were to be implemented, hopefully over a period of time customers would learn the required task. This would greatly reduce the number of transactions needing manual assistance and the number of calls containing connected input. Once connected digit recognition has achieved the same performance as isolated speech, the restriction of isolated input can be relaxed. These improvements should greatly increase end-to-end system performance.

IX. SUMMARY

Results have been presented from a series of speech recognition experiments on a speech database obtained from 7373 telephone customers speaking in an actual telephone environment in Baton Rouge, Louisiana. The best performance of 86.3-percent correct digit recognition was obtained when a set of speaker-independent templates

was created from a subset of the data and tested on the remaining data.

We described a syntax-directed recognition system that incorporates information about a seven-digit telephone number task. System accuracy was shown to improve by 9.4 percent.

Finally, a series of recognition tests was performed to quantify the robustness of speaker-independent templates created under one set of recording conditions and tested under another. Additionally, a template set was created from a subset of each of the regional templates sets. A recognition accuracy of 91 percent was obtained when tested against 20,000 isolated tokens from PO, BR, and MH data.

REFERENCES

1. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-23*, No. 1 (February 1975), pp. 67–72.
2. L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker Independent Recognition of Isolated Words Using Clustering Techniques," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-27*, No. 4 (August 1979), pp. 336–49.
3. L. R. Rabiner and J. G. Wilpon, "Speaker Independent, Isolated Word Recognition for a Moderate Size 54 Word Vocabulary," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-27*, No. 6 (December 1979), pp. 583–7.
4. J. G. Wilpon, L. R. Rabiner, and A. F. Bergh, "Speaker Independent Isolated Word Recognition Using a 129-Word Airline Vocabulary," J. Acoust. Soc. Amer., *72*, No. 2 (August 1982), pp. 390–6.
5. L. R. Rabiner and J. G. Wilpon, "A Simplified Robust Training Procedure for Speaker Trained, Isolated Word Recognition Systems," J. Acoust. Soc. Amer., *68*, No. 5 (October 1980), pp. 1069–70.
6. L. R. Rabiner and J. G. Wilpon, "Considerations in Applying Clustering Techniques to Speaker Independent Word Recognition," J. Acoust. Soc. Amer., *66*, No. 3 (September 1979), pp. 663–73.
7. L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and W. J. Keilin, "Isolated Word Recognition for Large Vocabularies," B.S.T.J., *61*, No. 10, Part 1 (December 1982), pp. 2989–3005.
8. J. G. Wilpon and L. R. Rabiner, "On the Recognition of Isolated Digits From a Large Telephone Customer Population," B.S.T.J., *62*, No. 7 (September 1983), pp. 1977–2000.
9. F. Pirz and K. Bauer, unpublished work.
10. L. R. Rabiner, J. G. Wilpon, and A. E. Rosenberg, "A Voice Controlled Repertory Dialer System," B.S.T.J., *59*, No. 7 (April 1980), pp. 571–92.
11. B. Aldefeld, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "Automated Directory Listing Retrieval System Based on Isolated Word Recognition," J. Acoust. Soc. Am., *68*, No. 5 (November 1980), pp. 1271–6.
12. J. G. Wilpon, L. R. Rabiner, and T. Martin, "An Improved Word-Detector Algorithm for Telephone-Quality Speech Incorporating Both Syntactic and Semantic Constraints," AT&T Bell Lab. Tech. J., *63*, No. 3 (March 1984), pp 479–98.
13. J. G. Wilpon and L. R. Rabiner, "A Modified *K*-Means Clustering Algorithm for Use in Speaker Independent Isolated Work Recognition," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-33*, No. 3 (June 1985).
14. L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-29* (August 1981), pp. 777–85.
15. S. E. Levinson, L. R. Rabiner, A. E. Rosenberg and J. G. Wilpon, "Interactive Clustering Techniques for Selecting Speaker Independent Reference Templates for Isolated Word Recognition," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-27*, No. 2 (April 1979), pp. 134–41.
16. A. E. Rosenberg and K. L. Shipley, "Evaluation of An Isolated Word Recognizer in Talker-Dependent and Talker-Independent Modes Using a Large Telephone-

Band Database," Conf. Rec., 1984 IEEE Int. Conf. Acoust., Speech, Signal Processing, March 1984.
17. C. S. Myers and L. R. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-29*, No. 2 (April 1981), pp. 284–97.
18. L. R. Rabiner, "On the Applications of Energy Contours to the Recognition of Connected Word Sequences," AT&T Bell Lab. Tech. J., *63* (December 1984), pp. 1981–95.
19. L. R. Rabiner, A. F. Bergh, and J. G. Wilpon, "An Improved Training Procedure for Connected-Digit Recognition," B.S.T.J., *61* (July–August 1982), pp. 981–1001.

AUTHOR

Jay G. Wilpon, B.S., A.B. (cum laude in Mathematics and Economics, respectively), 1977, Lafayette College, Easton, Pa.; M.S. (Electrical Engineering/Computer Science), 1982, Stevens Institute of Technology, Hoboken, N.J.; AT&T Bell Laboratories, 1977—. Since June 1977 Mr. Wilpon has been with the Acoustics Research Department at AT&T Bell Laboratories, Murray Hill, N.J., where he is a Member of the Technical Staff. He has been engaged in speech communications research and presently is concentrating on problems of speech recognition. He has published extensively in this field and has been awarded several patents. His current interests lie in training procedures, speech detection algorithms, and determining the viability of implementing speech recognition systems for general usage.