An Embedded Word Training Procedure for Connected Digit Recognition

L. R. Rabiner
A. Bergh
J. G. Wilpon

Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT The "conventional" way of obtaining word reference patterns for connected word recognition systems is to use isolated word patterns, and to rely on the dynamics of the matching algorithm to account for the differences in connected speech. Connected word recognition, based on such an approach, tends to become unreliable (high error rates) when the talking rate becomes grossly incommensurate with the rate at which the isolated word training patterns were spoken. To alleviate this problem, an improved training procedure for connected word (digit) recognition is proposed in which word reference patterns from isolated occurrences of the vocabulary words are combined with word reference patterns extracted from within connected word strings to give a robust, reliable word recognizer over all normal speaking rates. In a test of the system (as a speaker trained, connected digit recognizer) with 18 talkers each speaking 40 different strings (of variable length from 2 to 5 digits), median string error rates of 0% and 2.5% were obtained for deliberately spoken strings and naturally spoken strings, respectively, when the string length was known. Using just isolated word training tokens, the comparable error rates were 10% and 11.3% respectively.

I. Introduction

Recently, several algorithms for recognizing a connected string of words based on a set of discrete word reference patterns have been proposed [1-5]. Although the details and the implementations of each of these algorithms differ substantially, basically they all try to find the optimum (smallest distance) concatenation of isolated word reference patterns that best matches the spoken word string. Thus the success of all these algorithms hinges on how well a connected string of words can be matched by concatenating members of a set of isolated word reference patterns. Clearly for talking rates that are comparable to the rate of articulation of isolated words (e.g. on the order of 100 words per minute (wpm)), these algorithms have the potential of performing quite accurately. However, when the talking rates of the connected word strings become substantially larger than the articulation rate for isolated words (e.g. around 150 wpm), then all the pattern matching algorithms tend to become unreliable (yield high error rates).

The purpose of this paper is to show how improved word reference patterns for a connected word recognizer can be obtained in a reliable and robust manner from connected word strings. We also show how such embedded training patterns can be combined with standard isolated word reference patterns to provide a training set which is capable of recognizing connected word strings spoken at normal talking rates.

The "standard" reference set for connected digit recognition for speaker trained use is the set of isolated word patterns obtained via the robust training procedure of Rabiner and Wilpon [6]. The "improved" set of word reference patterns is obtained by combining the isolated word reference patterns, with word reference patterns extracted from actual connected word strings.

A block diagram of the embedded reference word training procedure is given in Figure 1. The philosophy of the training procedure is similar to that of the robust training procedure for isolated words [6]. For each word in the vocabulary a sequence generator (i.e. a talker) produces a string of words in which the

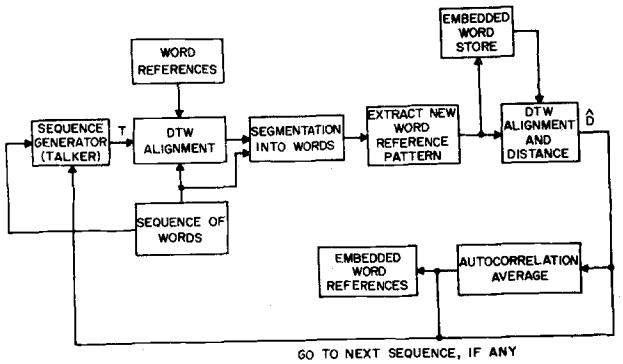

Figure 1. Block diagram of embedded digit training procedure.

desired word appears. A DTW alignment procedure matches a set of concatenated word references (corresponding to the words in the string) to the test string T, thereby providing a segmentation of T. The pattern for the appropriate reference word is extracted from T (based on the segmentation) and stored in a temporary word store. It is then compared to all previous occurrences of that word in the store using another DTW alignment procedure. For each such comparison, a distance score is obtained. If any distance score falls below a specified threshold, then the pair of tokens giving the minimum distance (among all versions in the store) are averaged (after time alignment) and the resulting reference pattern is saved in an embedded word store. This procedure is iterated until an embedded pattern is obtained for each word in the vocabulary.

The only unspecified aspect of the embedded training procedure is the word sequence generator. We have considered two sequence generators — a non-coarticulated (NC) sequence generator, and a coarticulated (CO) sequence generator. For both cases the desired digit was the middle digit of a 3-digit sequence. The NC sequences had the property that the preceding and following digits had different manners of production (at the boundary) than the middle digit. Similarly the CO sequences had the property that either the preceding or the following digit (or both) had a similar manner of production as the middle digit.

II. Evaluation of the Embedded Word Training Procedure

To study the effectiveness of the embedded word training procedure, 18 talkers (9 male, 9 female) each trained a connected digit recognizer in the following manner:

1. A set of isolated digit templates was created for each talker using the robust training procedure of Rabiner and Wilpon [6]. A single template was created for each of the 10 digits. An 11^{th} template was created for the digit 8 in which the talker was requested to speak the digit 8 without releasing the final /t/. For the normal 8 template, each talker was told to release the final /t/. We designate the isolated digits training set as IS.
2. Four sets of embedded templates were created for each talker using the robust training procedure of Figure 1. The four sets of templates differed in the speed at which the talker spoke

CH 1746-7/82/0000 · 1621 $ 00.75 © 1982 IEEE

the 3-digit training sequences (normal — NR, and deliberate — DL), and the degree of digit coarticulation — CO, and non-coarticulation — NC). Thus the four embedded training sets were denoted as CO.DL, CO.NR, NC.DL, and NC.NR.

The composite reference sets used for testing were created by considering various combinations of the isolated digits test set with one or more of the embedded digits test sets.

Two sets of testing data were obtained. The first set, denoted as TS.DL, consisted of 40 randomly chosen (i.e. a different set of 40 strings for each talker) digit strings of varying length from 2 to 5 digits, each string spoken deliberately (i.e. carefully articulated). The second set, denoted as TS.NR, consisted of the same 40 digit strings (for each talker) as in TS.DL, but instead spoken at a normal rate.

3.1 Evaluation Tests

For each of the two test sets (TS.DL and TS.NR) and for each talker, results are presented on the following reference sets:

1. IS — Isolated digits only, 1 template per digit, 11 digits.
2. IS ⊕ NC.DL — IS combined with NC.DL — 2 templates per digit.
3. IS ⊕ CO.NR ⊕ NC.DL — 3 templates per digit.

The recognition test consisted of using the LB based DTW algorithm [5] to match the spoken test strings by the best sequence of reference patterns, regardless of length. In performing the matches, the parameters of the LB algorithm were set as follows:

1. For all references and test sets the parameters δ_{END}, M_T, and ϵ were given values of $\delta_{END} = 4$, $M_T = 1.6$ and $\epsilon = 20$.*
2. The parameters δ_{R_1} and δ_{R_2} were made to vary with each subset of reference patterns in the following manner. For isolated digit patterns (in all reference sets) the values $\delta_{R_1} = 4$ and $\delta_{R_2} = 6$ were used. For NC.DL reference patterns the values $\delta_{R_1} = 2$, $\delta_{R_2} = 3$ were used. For CO.NR reference patterns the values $\delta_{R_1} = \delta_{R_2} = 0$ were used.

The logic for this choice was that the NC.DL patterns could be shortened somewhat, but not as much as the isolated patterns. However the CO.NR patterns could not be shortened at all since they came from highly reduced normally spoken digit sequences.

Before presenting results of the recognition tests, we first give some statistics on the rate of talking of each of the talkers.

3.2 Talking Rate Statistics

The statistics on the overall average talking rate (as a function of number of digits in the string) is given in Figure 2. The talking rates are given in terms of words-per-minute (WPM). For deliberately spoken strings, the average talkers rate varies from 99 to 156 WPM (across the 18 talkers). Thus a high degree of rate variability exists across talkers for deliberately spoken strings of digits. For naturally spoken digit strings the average talking rates varies from 118 to 193 WPM (across the 18 talkers), again pointing out the high degree of variability in normal talking rates.

The plots of Figure 2, however, show that when averaged across talkers, the talking rate for different length strings does not vary as markedly as for different talkers. For deliberately spoken digit strings, the slowest average rate is 126 WPM for 2-digit strings, and the average rate increases to 134 WPM for 4-digit strings. Almost no increase in average talking rate is found for 5-digit strings over that for 4-digit strings. For normally spoken digit strings the same trends of the average talking rate are seen across different length strings. Thus the average rate for 2-digit strings is 150 WPM and it increases to 170 WPM for 4-digit strings. However the average rate for 5-digit normally spoken strings (171 WPM) is essentially the same as for 4-digit normally spoken strings.

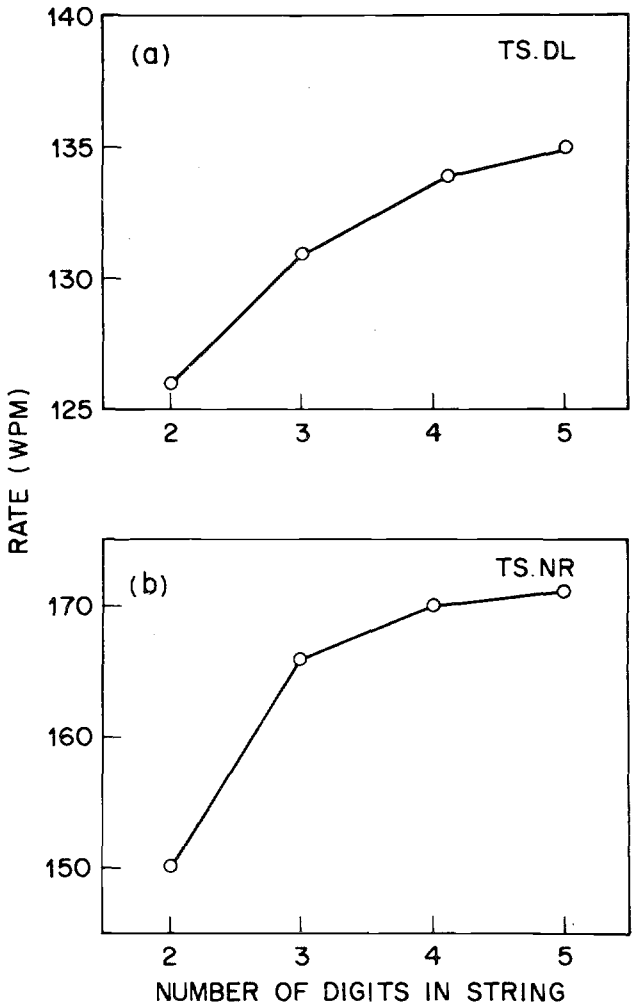

Figure 2. Plot of average talking rate as a function of the number of digits in the string for TS.DL (part a) and TS.NR data (part b).

3.3 Recognition Test Results on Large Data Base

The results of the recognition tests on the 18 talker data base are given in Tables I and II, and are plotted in Figure 3. Tables I and II give average and median string error rates (averaged over talkers and various length strings) for the 3 reference sets for TS.DL data (Table I) and TS.NR data (Table II). Included in the table are string error rates based on the top 1, 2, 3, 4, and 5 candidates (regardless of string length) and for the case in which the string length was known apriori (KL) — i.e. we only considered strings of the proper length. Results are given for the median error rate, the average error rate for all 18 talkers, the average error rate for 17 of the 18 talkers (omitting the one with the highest error rate) and the average error rate for 16 of the 18 talkers (omitting the 2 talkers with the highest error rates).

Figure 3 shows plots of the average and median error rates as a function of the top n candidates (n=1,2,3,4,5) for both TS.DL and TS.NR data using the reference set IS ⊕ NC.DL ⊕ CO.NR, which provided the best overall results.

The results of Tables I and II and Figure 3 show the following:

1. For deliberately spoken strings (TS.DL) the median string error rate is 5% on the top candidate, and falls to 2.5% on the top 2 candidates for reference set IS ⊕ NC.DL ⊕ CO.NR. Using reference set IS alone the median error rate is 16.3% on

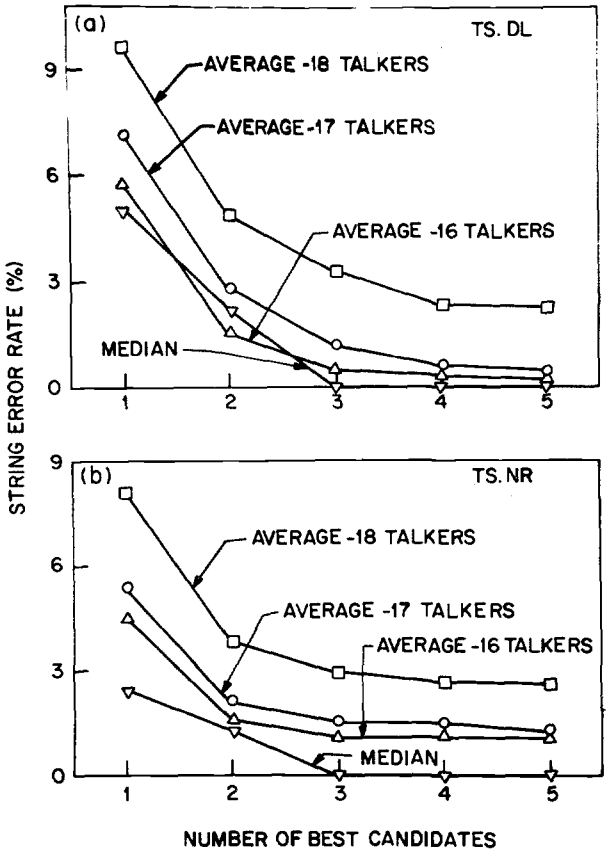

Figure 3. Plots of string error rates versus number of best candidates for TS.DL (part a) and TS.NR (part b) data.

Reference Set	Statistic	TS.DL Number of Best Candidates KL	1	2	3	4	5
Isolated Training (IS)	MEDIAN	10.0	16.3	8.8	6.3	5.0	5.0
	18 Talkers	11.4	20.0	11.8	9.9	8.5	8.3
	17 Talkers	9.1	17.6	9.3	7.8	6.6	6.5
	16 Talkers	7.5	16.3	8.3	6.4	5.5	5.3
IS ⊕ NC.DL	MEDIAN	2.5	10.0	2.5	2.5	2.5	2.5
	18 Talkers	5.7	14.1	7.4	5.3	5.0	4.9
	17 Talkers	3.1	12.2	4.9	3.1	2.9	2.8
	16 Talkers	2.0	10.5	3.4	2.0	1.9	1.9
IS ⊕ NC.DL ⊕ CO.NR	MEDIAN	0	5.0	2.5	0	0	0
	18 Talkers	3.3	9.7	4.9	3.3	2.6	2.5
	17 Talkers	1.3	7.2	2.8	1.2	0.6	0.4
	16 Talkers	1.1	5.8	1.9	0.5	0.3	0.2

Table I

Average and Median String Error Rates
(Percentage Errors) for TS.DL Data
For the 3 Reference Sets

Reference Set	Statistic	TS.NR Number of Best Candidates KL	1	2	3	4	5
Isolated Training (IS)	MEDIAN	11.3	13.8	10.0	7.5	7.5	7.5
	18 Talkers	16.3	19.0	11.7	10.3	10.1	9.7
	17 Talkers	13.1	15.9	9.3	8.1	7.9	7.5
	16 Talkers	13.0	15.5	9.1	8.0	7.8	7.3
IS ⊕ NC.DL	MEDIAN	5.0	7.5	3.8	3.8	3.8	3.8
	18 Talkers	8.3	11.5	6.4	5.8	5.4	5.3
	17 Talkers	5.3	8.4	4.0	3.7	3.2	3.1
	16 Talkers	4.8	7.8	3.3	3.1	2.7	2.7
IS ⊕ NC.DL ⊕ CO.NR	MEDIAN	2.5	2.5	1.3	0	0	0
	18 Talkers	5.8	8.1	3.8	2.9	2.6	2.5
	17 Talkers	3.7	5.4	2.1	1.5	1.5	1.3
	16 Talkers	3.4	4.5	1.6	1.1	1.1	1.1

Table II

Average and Median String Error Rates
(Percentage Errors) for TS.NR data
for the 3 Reference Sets

the top candidate and is still 8.8% on the top 2 candidates. Using reference set IS ⊕ NC.DL the median error rate is 10.0% on the top candidate and 2.5% on the top 2 candidates.

2. For deliberately spoken strings the median error rates with known length strings are 0% for reference set IS ⊕ NC.DL ⊕ CO.NR, 10% for reference set IS, and 2.5% for reference set IS ⊕ NC.DL.
3. For deliberately spoken strings the average error rate scores for all 18 talkers are significantly larger than the median error rate scores. However, when the talker with the highest error rate is omitted (i.e. only 17 talkers are used) the average and median error rate scores are comparable.
4. For normally spoken strings (TS.NR) the median string error rate is 2.5% on the top candidate and falls to 1.3% on the top 2 candidates using reference set IS ⊕ NC.DL ⊕ CO.NR. Using reference set IS alone the median error rate is 13.8% on the top candidate and 10% on the top 2 candidates. Using reference set IS ⊕ NC.DL the median error rates are 7.5% on the top candidate and 3.8% on the top 2 candidates.
5. For normally spoken strings the median error rates with known length strings are 2.5%, 11.3% and 5.0%, for reference sets IS ⊕ NC.DL ⊕ CO.NR, IS, and IS ⊕ NC.DL respectively.
6. For normally spoken strings the average error rate scores for all 18 talkers are significantly larger than the median error rate scores. Again when the talker with the highest error rate is omitted (the same one as for deliberate strings) the average and median error rate scores become comparable.

III. Conclusions

From the above set of results we can draw the following conclusions:

1. The inclusion of embedded digit training led to significant improvements in digit recognition accuracy for both deliberately and normally spoken digit strings.
2. Somewhat better recognition accuracy was obtained when the string length of the test sequence was known than when no knowledge of string length was used. This result is due to the high probability of inserting extraneous short digits (the embedded 2's and 8's) in deliberately (and often naturally) spoken strings.
3. The accuracy with which connected digit strings could be recognized can be made essentially independent of the talking rate, especially if one can take advantage of knowing in advance the length of the digit string.
4. Both coarticulated and non-coarticulated embedded digits training patterns aid in recognizing connected strings of digits.
5. There are some talkers (1 of the 18 tested here) for whom all the training procedures failed. For this one talker the string error rate exceeded 50% on all reference and test sets. No obvious or clear explanation is available for this result except

   for the fact that all the training procedures indicated a high degree of variability in speaking digits for this talker (i.e. it took the maximum number of iterations to obtain each template set). This high degree of variability basically implied that no single set of reference patterns could adequately match the digits of this talker. Hence highly inaccurate connected digit recognition resulted.

The above results and the conclusions drawn from them indicated that the improved training procedure provided, in general, robust, reliable digit patterns that greatly aided the connected digit recognizer in recognizing connected digit strings for these talkers at essentially any reasonable talking rate.

IV. Summary

An improved training procedure for extracting reference word templates for connected word recognition systems has been described. The resulting reference patterns essentially model the word characteristics when embedded in connected word strings. Hence a reference set consisting of both isolated word patterns and embedded word patterns has the capability of providing reliable recognition of connected word strings spoken at natural rates. In an evaluation of this procedure it was shown (using a digit vocabulary) that high string accuracy could be obtained, if the length of the digit string was known apriori, for deliberately and naturally spoken digit strings. If the length of the digit was not known, the string error rate was somewhat higher due to the problem of inserting short digits into the matching sequence.

References

[1] T. K. Vintsyuk, "Element-Wise Recognition of Continuous Speech Composed of Words from a Specified Dictionary," *Kibernetika*, Vol. 2, pp. 133-143, April 1971.

[2] J. S. Bridle and M. D. Brown, "Connected Word Recognition Using Whole Word Templates," *Proc. Inst. Acoustics*, Autumn Conf., pp. 25-28, Nov. 1979.

[3] H. Sakoe, "Two Level DP-Matching — A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-27, No. 6, pp. 588-595, Dec. 1979.

[4] L. R. Rabiner and C. E. Schmidt, "Application of Dynamic Time Warping to Connected Digit Recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-28, No. 4, pp. 377-388, Aug. 1980.

[5] C. S. Myers and L. R. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 2, pp. 284-297, April 1981.

[6] L. R. Rabiner, and J. G. Wilpon, "A Simplified, Robust Training Procedure for Speaker Trained, Isolated Word Recognition," *J. Acoust. Soc. Am.*, Vol. 68, No. 5, pp. 1271-1276, Nov. 1980.