DURATIONAL CONSTRAINTS FOR NETWORK-BASED CONNECTED DIGIT RECOGNITION

Marcia A. Bush

Schlumberger Palo Alto Research, 3340 Hillview Ave., Palo Alto, CA 94304, USA. [1]

This paper examines the influence of durational constraints on recognition accuracy in an acoustic-phonetically based, speaker-independent connected digit recognizer. The constraints are expressed using a set of finite-state pronunciation networks, together with specifications of minimum and maximum allowable durations for network primitives. The recognizer was tested on a corpus of 1232 5-digit and 7-digit strings, with and without a priori knowledge of string length. Recognition accuracies ranged from 33.9% to 94.6% and from 91.6% to 96.8%, for unknown and known string lengths, respectively, depending on the particular durational constraints incorporated in the network models.

INTRODUCTION

The word models used in the connected digit recognizer described here consist of a set of finite-state pronunciation networks, in which primitive branches correspond to meaningful acoustic-phonetic units (Table 1). Unlike networks based on the hidden Markov model formalism, these word models allow for the convenient expression of acoustic-phonetic constraints which are manifest over portions of an utterance longer than a single time frame. One example of such a constraint is segment duration. [2]

This paper examines recognizer performance as a function of the minimum and maximum allowable durations for primitives in two types of network: 1) a baseline network formed by simply connecting in parallel the isolated digit models shown in Table 1; and 2) a set of networks which incorporate additional paths representing prepausal lengthening for the digits *oh* and *eight*. Constraints on minimum duration were found to have the greatest influence on recognition accuracy, particularly when recognition was performed without a priori knowledge of digit string length. Prepausal durational constraints proved useful in reducing a common class of digit insertion errors.

The digit recognizer incorporates a set of generalized acoustic pattern matchers and a dynamic programming search in addition to the pronunciation network models. Details of the recognition framework, and of signal preprocessing, are provided in [1] and [2].

CORPUS

The corpus used in the recognition experiments consisted of the adult-talker, 5-digit and 7-digit subset of the training portion of Texas Instruments' multi-dialect connected digits database [3]. The utterances of half of the talkers (27M, 29F, 1232 tokens) in this subset were used for training the recognizer and the utterances of the remaining half (28M, 28F, 1232 tokens) were used for testing. These two corpora will be referred to as TRNA-57 and TRNB-57 respectively.

An initial version of the recognition system was trained on 616 handlabelled 5-digit strings from TRNA-57, and run over the entire TRNA-57 corpus [2]. The segmentations generated for correctly identified tokens in this experiment defined a set of bootstrapped training data, which were used in all of the experiments reported here. Statistics on minimum and maximum segment duration were collected for both the handmarked and bootstrapped data, and used in specifying the durational constraints in the network models.

RESULTS

Table 2 shows recognition data for corpus TRNB-57 using the baseline network (unknown string length) and various constraints on segment duration. As indicated in the first three columns, recognition accuracy ranges from 33.9% when the minimum allowable duration is a single frame (10 msec), as in first-order hidden Markov models, to 93.2% using the minimum durations for the bootstrapped training data. During the bootstrapping experiment, very short durations (i.e., those falling in the bottom 5% of the distributions for each segment type) were penalized, with the result that minimum durations for the bootstrapped training data were typically 1 to 2 frames longer than for the handmarked utterances. The main effect of prohibiting very brief segments is to reduce the number of digit insertion errors from 1407 to 33.

Not surprisingly, constraints on mimimum segment duration have a less dramatic effect on recognizer performance when string length is known a priori. As shown in the first two columns of Table 3, recognition accuracy increases from 91.6% with minimum allowable durations of a single frame to 96.8% using the bootstrapped minima.

In the experiments just described, the maximum allowable segment duration was 1.5 times that observed for the bootstrapped data. Comparison of columns 3 and 4 in Table 2, and of columns 2 and 3 in Table 3 indicate that imposing tighter constraints on maximum segment duration (i.e., the bootstrapped maxima) has virtually no effect on recognition accuracy with the baseline network.

Table 4 shows recognition data for corpus TRNB-57 using networks which require prepausal lengthening for the digit *oh* (column 1) or for both *oh* and *eight* (columns 2-4). These networks were motivated by the observation that the most consistent errors using the baseline network were *oh* and *eight* insertions following the third digit of a 7-digit string. (Presumably, talkers used a "telephone number" grouping in producing these tokens.) *Oh*'s were most often inserted after the digits *oh*, *two* and *zero*, and *eight*'s after *two*, *three* and *eight*. Prepausal lengthening was required for each of the vocalic segments in the two digits, with the degree of lengthening estimated from the two sets of training data.

Incorporating prepausal lengthening for the digit *oh* serves to reduce the number of *oh* insertions from 19 to 10 relative to the baseline situation (Table 5, columns 1 and 2), and to increase overall recognition accuracy from 93.0% to 93.8% (Table 2, column 4, and Table 4, column 1.) Adding

[1] After 1 Aug 86: Division of Engineering, Box D, Brown University, Providence, RI 02912, USA.

[2] As used in this paper, the term *segment* refers to the acoustic-phonetic primitives listed in Table 1.

prepausal lengthening for *eight* reduces the number of *eight* insertions from 17 to 11 (columns 2 and 3, Table 5) and increases overall accuracy to 94.2% (Table 4, column 2).

Virtually all of the prepausal *oh* and *eight* insertions which remain after these two network modifications occur following the digits *two* and *three*. Several of these errors can be eliminated by increasing the maximum allowable durations for the vocalic portions of *two* and *three* from 1.0 to 1.5 times their bootstrapped values (Table 5, column 4), increasing overall recognition accuracy to 94.6% (Table 4, column 3). (Additional *eight* insertions can be eliminated by allowing a noisy or breathy "release" segment after these same two digits.) Allowing looser maximum durational constraints for all segments results in a small decrease in recognizer performance (Table 4, column 4), in contrast to experiments with the baseline network.

SUMMARY

The experiments described above illustrate the importance of appropriate durational constraints for high-accuracy network-based connected digit recognition. Modeling duration in the current system is facilitated by the use of network primitives corresponding to meaningful acoustic-phonetic units.

REFERENCES

[1] M. Bush and G.Kopec, "Network-based connected digit recognition using explicit acoustic-phonetic modeling", in *Proceedings, 1986 IEEE International Conference on Acoustics, Speech and Signal Processing*, Tokyo, Japan (Apr 1986).

[2] M. Bush and G.Kopec, "Network-based connected digit recognition", submitted for publication in *IEEE Transactions on Acoustics, Speech and Signal Processing* (Mar 1986).

[3] G. Leonard, "A database for speaker-independent digit recognition", in *Proceedings, 1984 IEEE International Conference on Acoustics, Speech and Signal Processing*, San Diego, CA (Mar 1984).

Digit	Network Primitives
oh	OW1 OW2 OW3
1	WAH1 WAH2 N
2	(TS) TR UW1 UW2
3	TH RIY1 RIY2
4	F AOR1 AOR2
5	F AY1 AY2 V
6	S IH KS KRS
7	S EH V AX N
8	EY1 EY2 (TS) (TR)
9	NI AY1 AY2 NF
zero	Z IYR1 IYR2 ROW1 ROW2

Table 1: Network primitives for the baseline pronunciation network. Parentheses indicate optional segments.

Segment Duration: Minimum / Maximum	1 frame / 1.5xBS	HM / 1.5xBS	BS / 1.5xBS	BS / BS
% correct	33.9	86.2	93.2	93.0
string length errors	800	118	47	49
matches	7270	7306	7328	7338
substitutions	121	79	49	47
insertions	1407	124	33	43
deletions	1	7	15	7

Table 2: Recognition data for corpus TRNB-57 using the baseline network and various constraints on segment duration. Unknown string length. HM = handmarked TRNA-5, BS = bootstrapped TRNA-57.

Segment Duration: Minimum / Maximum	1 frame / 1.5xBS	BS / 1.5xBS	BS / BS
% correct	91.6	96.8	96.8
string length errors	–	–	–
matches	7257	7350	7349
substitutions	122	42	43
insertions	13	0	0
deletions	13	0	0

Table 3: Recognition data for corpus TRNB-57 using the baseline network and various constraints on segment duration. Known string length. BS = bootstrapped TRNA-57.

Segment Duration	BS, except as noted			
Prepausal Minimum Lengthening	OW 1.75xBS	OW - 1.75xBS EY - 1.5xBS		
Maximum Lengthening	None	None	UW,RIY 1.5xBS	All 1.5xBS
% correct	93.8	94.2	94.6	93.8
string length errors	40	34	30	39
matches	7338	7338	7338	7328
substitutions	47	47	47	49
insertions	34	28	24	25
deletions	7	7	7	15

Table 4: Recognition data for corpus TRNB-57 using networks with prepausal lengthening and various constraints on segment duration. Unknown string length. BS = bootstrapped TRNA-57.

Segment Duration	BS, except as noted			
Prepausal Minimum Lengthening	None	OW 1.75xBS	OW - 1.75xBS EY - 1.5xBS	
Maximum Lengthening	None	None	None	UW,RIY 1.5xBS
oh	19	10	10	7
8	17	17	11	10

Table 5: *Oh* and *eight* insertion errors for corpus TRNB-57 for various networks and constraints on segment duration. Unknown string length. BS = bootstrapped TRNA-57.