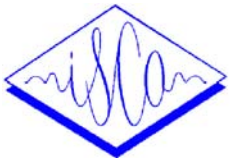

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

IMPROVEMENT ON CONNECTED DIGITS RECOGNITION USING DURATION CONSTRAINTS IN THE ASYNCHRONOUS DECODING SCHEME

Miroslav Novak
IBM Watson Research Center - Human Language Technologies Group
P.O. Box 218, Yorktown Heights, NY 10598, USA
email: novak@watson.ibm.com

1 ABSTRACT

This paper describes the use of an explicit word duration model in the environment of a HMM based time asynchronous stack search decoder. The benefit of the method is demonstrated on the task of connected digit recognition. Analysis of typical errors observed on this task suggests that appropriate word duration modeling can improve recognition accuracy. Duration model based on the Gamma Distribution, applied as a post-processing step during iterations of the search algorithm, reduces the error rate of the baseline system by 14%.

2 INTRODUCTION

Current connected digit string recognition systems can achieve very high decoding accuracy. Typically the word error rates are bellow 1% (for clean speech, such as the TI digits data set). It has been observed the the distribution of errors is not uniform among all digits [3]. Certain digits (in specific context) are more likely to be decoded incorrectly. For example, the digit *OH* is frequently misrecognized, especially in a context of one or more repeated *OH*s. Detailed analysis of the cause of this phenomenon and appropriate corrective actions can lead to significant reduction of the error rate of the connected digit recognition system. The fact that the involved digits are associated with short acoustic segments suggests that proper duration modeling can reduce the error rate.

Approaches to explicit duration modeling can be divided into two major groups. The first approach incorporates the duration model into the acoustic model [2]. While good performance can be achieved, the overhead represented by increased demand on training data and cpu is usually prohibitive. The second approach relies on the independently trained duration model used as a post-processor, to perform re-scoring of N-best sequences. The drawback of this technique is the inherent limitation of a post processing scheme - the correct path can be pruned out during the earlier steps of decoding. However the post-processing approach is very popular for it's simple implementation and minimal overhead in both model training and decoding. This technique is usually used in the time synchronous environment (Viterbi decoder). Implementation in a non-Viterbi decoder (e.g. maximum likelihood based decoder) poses certain complications, since the exact duration of a each model node is not known. This work shows that that post-processing type can be still efficiently used in such situations.

3 ANALYSIS OF ERRORS

The TI digits corpus [5] was chosen as a test corpus. It has been observed that certain digits (*OH*,*EIGHT*) are more likely to be involved in decoding errors. Moreover, the context in which these digits appear, plays an important role. The experiments confirmed the previous observations, practically all deletion errors involve either *OH* or *EIGHT*. The bulk of these deletion appear in the context of repetition of the same digit. Since the TI corpus is relatively rich on string of repeated *OH*s and *EIGHT*s, it was possible to analyze these errors and determine probable cause of this behavior. Table 1 shows the counts of deletions and insertions of these two digits. All error count are also shown as percentages of the total number of errors. As can be seen, deletions and insertions of *OH* and *EIGHT* represent 71% of all errors.

It is straightforward to implement explicit duration modeling in the time asynchronous stack search decoding scheme [1]. It allows to incorporate the post-processing explicit duration model into the search during the forward pass. The N-best re-scoring is elimi-

nated, while the inexpensive post-processing approach is used. The core of the decoder is a stack, which keeps several best hypothesis during the search, each of which is characterized by a sequence of words, the most probable end time, total score of the path and an end-time distribution. This distribution, properly normalized, represents the probability density function of the exact end time of a particular path. It is usually limited to a certain time interval around the most probable ending time. It is obtained by the forward pass computation of the total likelihood of the HMM representing the evaluated path. This distribution is used as initial distribution for the following word models.

The shape of this distribution is typically characterized by a distinct global maximum (the most probable end time) and exponential decay in both directions away from this maximum. If the distribution contains two or more local maxima, it could mean that two distinct utterances of the same digit are being matched by a single model. An explicit word duration model can be used to eliminate the false local maxima and emphasize the correct maximum.

An example of such situation is shown in figure 1. The uttered sequence was *TWO OH OH*. The solid line represents the end-time distributions of each word. When the first *OH* is uttered, the the curve peaks at the correct end time and then decreases during the pause between the utterances. But at the beginning of the next utterance of the same word, the curve starts to rise again and peaks at end of this second utterance. On the same picture, the result of a forced extension of the second utterance of *OH* using the end-time distribution of the previous *OH* is shown. Due to the penalty imposed by the language model, the achieved final probability is actually lower and the final path containing only a single utterance of *OH* is selected as the best path.

Decreasing the language model probability would certainly solve this particular problem. But the weight with which the language model probability is applied is a global parameter which is usually tuned for the optimal ratio between insertion and deletion errors. A solution with only local effect would be more appropriate.

4 EXPLICIT MODELING

Conventional Hidden Markov Models (HMM) implicitly model the duration probability distribution of each state $p_i(\tau)$ by a Geometric distribution, i.e. $p_i(\tau) = a_{ii}^{\tau-1}(1-a_{ii})$ where i is the state, τ is the duration and a_{ii} is the state self transition probability.

This geometric distribution is usually inappropriate for the duration distribution modeling of real words. Other types of distributions had been proposed for the explicit state and word duration modeling. For example, the Gamma distribution was found to produce a high quality fit to the duration distribution observed on the training data [4].

$$p(x) = \frac{\alpha^k}{\Gamma(p)} e^{-\alpha x} x^{k-1} \qquad \alpha > 1, p > 0 \qquad (1)$$

In figure 1, the duration distribution of each word (obtained from the Gamma distribution model (1) is shown (dashed line) positioned relatively to the end of the previous word. Obviously, the second and main peak of the distribution lies completely out of the range of acceptable durations of the word *OH* . Figure 2 shows how the situation changes after the end time distribution of the previous word is weighted by the duration distribution - the second peak practically disappears and the sentence is recognized correctly.

The advantage of the asynchronous approach is apparent here. Since the whole end-time distribution is avaliable for inspection, certain techniques can be applied to make a decision if the use of the duration model is appropriate. For example, if only one distinct peak in the distribution is identified, there is usually little need for use of the duration model, moreover, unnecessary penalties imposed by the model on unusually long or short pronunciations can be avoided. On the other hand, if the two peaks with a relatively deep gap between them are found the duration model clearly helps to emphasize the right peak.

Since the duration model is applied at each iteration during the forward recognition pass, there is no need of a post-processing step, while the advantages of the post-processing approach are fully utilized.

5 RECOGNITION EXPERIMENTS

For recognition experiments, a system was built using the TI digits database. The system uses different acoustic models for sub-phonetic units in different contexts. These instances of context dependent classes are identified by growing a decision tree from the training data. The acoustic feature vectors that characterize the training data at the leaves are modeled by a mixture of Gaussians pdf's, with diagonal covariance matrices (a total 4300 Gaussians were used). As far as the output distributions on the state transitions of the model are concerned, rather than expressing the output distribution directly in terms of the feature vector, the IBM system expresses it in terms of the rank of the leaf [6]. The acoustic front end uses a FFT based filter

	OH		EIGHT	
	ins	del	ins	del
without model	36 (19%)	59 (32%)	10 (5%)	28 (15%)
with model	39 (24%)	49 (30%)	11 (7%)	12 (7%)

Table 1: Comparison of errors counts

	without model	with model
Word error rate	0.65 %	0.56 %
String error rate	1.94 %	1.67 %

Table 2: Comparison of error rates

bank followed by cepstral rotation. Frame energy and dynamic parameters ($\Delta + \Delta\Delta$) were added to each feature vector. Sentence based cepstra mean removal was used.

Viterbi segmentation of the training data was performed to obtain data for the Gamma distribution model estimation. A weighting factor was applied to each word duration model to match the dynamic range of the end point distributions. The new end-time distribution is computed as::

$$p'_e(t) = w_m * p_\Gamma(t - t_{start}) * p_e(t) \qquad (2)$$

where $p_e(t)$ is the original end-time distribution, t_{start} is the beginning of the word (the most probable end time of the previous word) and $p_\Gamma(x)$ is the trained Gamma distribution model (1). The proper weighting factor w_m was found experimentally.

The test set of the TI digits database (8700 sentences) was decoded first by a conventional decoder, without any duration model. In the second step, the above described method was used to apply the duration model. No constraints were imposed on the length of the decoded digit string. Table 1 compares the number of insertions and deletions of digits *OH* and *EIGHT*. Use of the duration model cause a slight increase in the number of insertions, but at same time significantly reduces the number of deletions, especially in case of the digit *EIGHT*. Total error rate (computed both word and string-wise) is shown in Table 2. As can be seen, the application of the duration model reduces the error rate by 14%.

6 CONCLUSION

The presented method helps to eliminate the most fre-

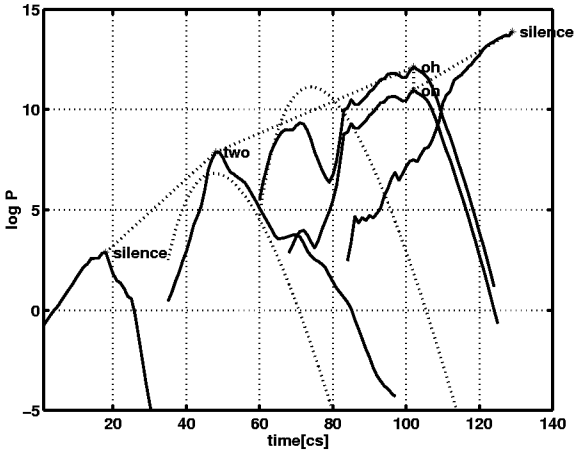

Figure 1: Decoding of a sequence *TWO OH OH* without a duration model

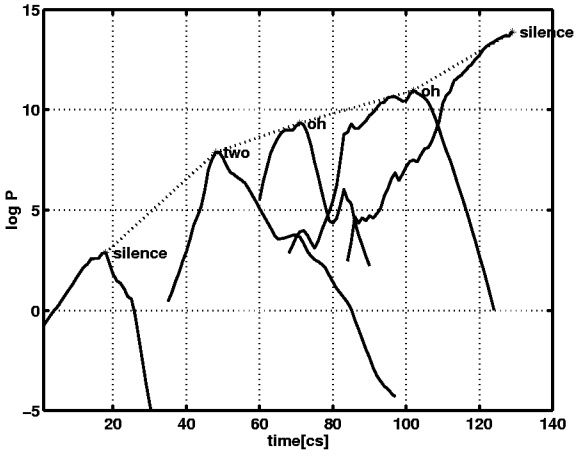

Figure 2: Decoding of a sequence *TWO OH OH* with a duration model

quent type of errors in the connected digits task - deletion of short words. Explanation of causes of these errors was presented and it was shown that the duration modeling can be effectively used to reduce the error rate. The achieved accuracy of the system is competitive with the performance reported on the TI digits task [2] [3] [4].

REFERENCES

[1] Gopalakrishnan, P.S. , Bahl, L.R., Mercer, R.L., "A tree search strategy for large vocabulary continuous speech recognition", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 572-575, May 1995

[2] Levinson, S.E., "Continuously variable duration hidden markov models for automatic speech recognition", Computer, Speech and Language, vol. 1, no. 1, pp. 29-45, 1986

[3] Rabiner, L.R. , Wilpon, J.G., Soong, F. K., "High performance connected digit recognition using hidden markov models", vol.37, ASSP-37, pp. 1214-1225, August 1989.

[4] Burshtein,D., "Robust parametric modeling of durations in hidden markov models", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 548-551, May 1995

[5] Leonard, R.G., "A database for speaker independent digit recognition" Pro. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 42.11.1-4, Mar 1984

[6] Bahl,L.R., de Souza, P.V., Gopalakrishnan, P.S., Nahamoo,D., Picheny M.A., " Robust Methods for using Context-Dependent features and models in a continuous speech recognizer", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 1994