# INTER-DIGIT HMM CONNECTED DIGIT RECOGNITION USING THE MACROPHONE CORPUS

*Yu-Hung Kao and Lorin Netsch*

PO Box 655303, MS 8374, Texas Instruments Incorporated, Dallas, TX 75265
{yhkao, netsch}@csc.ti.com

## ABSTRACT

Continuous digit recognition over the telephone channel is a key technology for many telecommuncations applications such as voice dialing, automatic banking, and credit card number entry. Speech recognizers usually acheive high performance by modeling the acoustics in Hidden Markov Models (HMMs) using large numbers of multivariate Gaussian mixtures with assumed diagonal covariance in order to model the variability of different speakers and channel conditions. In this paper, we present a system that uses single mixture 16 feature Gaussian distributions with an assumed identity covariance to achieve 1.0% word error and 5.7% sentence error rate on the Macrophone corpus. We found that inter-digit modeling, discriminant training, and per-utterance adaptation can each contribute about 30% reduction in error rate. Using this approach, we can realize a system with relatively low memory requirements.

## 1. INTRODUCTION

Many telecommunication speech recognition applications require rapid, accurate continuous-digit recognition technology. Typical applications include voice-dialing, automatic banking, and credit card number entry. Many speech recognizers acheive high performance by modeling digits as HMMs with large numbers of multivariate Gaussian mixtures to accommodate the variability of different speakers and channel conditions. In this paper, we implement a recognition system that models each digit as multiple HMM segments which depend on context in order to capture acoustic phenomena associated with coarticulation between digits. The HMM acoustic models are single mixture 16 feature Gaussian distributions with an assumed identity covariance matrix. We constrain the use of the multiple HMM segments for each digit by a grammar in order to enforce consistency of model usage based on the intended modeling of coarticulation. We find that a combination of inter-digit modeling, discriminant training, and per-utterance adaptation can each contribute about 30% reduction in error rate. Using this approach, we can realize a system with relatively low memory and metric computation requirements.

The second section of this paper introduces the motivation for and construction of inter-digit HMM models. The third section describes corpora used for training and testing. The fourth section presents the results of experiments performed using the inter-digit models. In the conclusion we address related areas for future research.

## 2. INTER-DIGIT CONTEXT MODELING

Coarticulation between words is a common phenomenon in continuous speech. One expects that the pronunciation characteristics associated with the end of a digit influences the pronunciation of beginning of the next digit, and vice-versa. This leads to a large number of possible acoustic signals between digits. When we examined digit likelihood scores and feature variances, we observed significantly larger scores and variances at the beginning and ending of word HMM models than during the middle portion of the digit. Figure 1 shows a typical example of the nominal expected scores for the digit “three”. The horizontal axis is the distribution number that would be encountered along the nominal path through the HMM.

These observations suggest that inter-digit context is important for modeling of continuously spoken digits. To model the inter-digit context we implemented the simple strategy of dividing each digit model into three equal-length HMM contexts: left, center, and right. The left and right models are affected by inter-word context, and thus require separate acoustics (and hence HMMs) for different contexts. We modeled the center HMM as consistent over all left and and right contexts. Other papers [3] have reported improvements by using similar partitions. This is by no means the best partition. We believe the partition should be based on variance analysis. Nevertheless, the partition should be

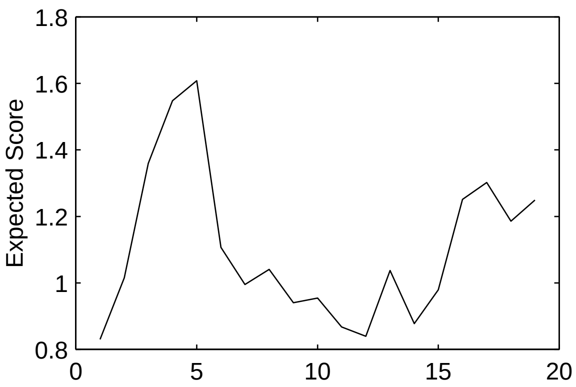

Figure 1: Expected Score for Three

indicative of improvements that may be achieved. We also assume the context effects will depend only on the immediate neighboring phone. For example, *zero* and *oh* both end with phone /ow/, therefore, their effect on the following digit should be the same. Constructing models with these assumptions results in 418 models with a total of 2612 acoustic distributions, about six times that of using whole word models. An example of the labeling of the three HMM model segments used to recognize the digit "two" in the various contexts is outlined as follows (bbb indicates a word boundary):

| Left Context | Center Context | Right Context |
|---|---|---|
| bbb_twoL_twoC | twoC | twoC_twoR_bbb |
| ow_twoL_twoC | | twoC_twoR_zeroL |
| n_twoL_twoC | | twoC_twoR_ohL |
| twoR_twoL_twoC | | twoC_twoR_oneL |
| iy_twoL_twoC | | twoC_twoR_twoL |
| fourR_twoL_twoC | | twoC_twoR_threeL |
| fiveR_twoL_twoC | | twoC_twoR_f |
| sixR_twoL_twoC | | twoC_twoR_s |
| | | twoC_twoR_eightL |
| | | twoC_twoR_nineL |

All HMM models used in this paper are finite duration (no state self-loops), but states can share the same acoustic distribution. We have observed consistently better performance for finite duration than infinite duration HMM models for the telephone digit recognition task. The number of states in each model is proportional to the average duration of each digit. We allow up to three states to share the same acoustic distribution and hence a maximum of three input frames can be explained by a single distribution. We also allow a skip of states to bypass a given distribution to accommodate different speaking rates. The multivariate Gaussian acoustic distribution model for each state consists of only a 16 element mean feature vector; all state covariances are assumed to be identity. We also construct separate male and female HMM models.

## 3. CORPORA

In this paper, we use the publicly available Macrophone corpus [1] for all training of HMM models. This corpus contains a total of 204,160 utterances from 5005 telephone calls. Each call provides between 1 and 44 utterances. The calls are partitioned into sets for training, development testing, and evaluation testing. The training set contains 4005 calls, the development test set contains 500 calls, and the evaluation test data are split into five sets of 100 calls per set. We use the training set to train the HMM models and the development testing set as one of the three corpora used for testing. We did not use the evaluation test set. Of the 44 utterances in each call, three utterances (utterances 10, 13, and 31) are digit strings (with possible *dash* and *area code* embedded). We used male and female utterances, but excluded utterances from children. We also excluded utterances containing out of vocabulary words, but retained utterances containing noise such as [mouth_noise] and [uh].

The syntax for Macrophone digits utterances is:

```
dddd dddd dddd
dddd dash dddd dash dddd
ddd ddd dddd
area code ddd ddd dddd
ddddd dd
ddddd dash dd

where 'd' stands for one digit.
```

The selection process mentioned above yields 4053 male training utterances and 5543 female training utterances, totaling 9596 training utterances. The same selection criteria results in 1207 test utterances from the development testing set with the same syntax as the training set.

In order to validate performance over multiple collection environments, we use two other corpora for additional testing. One testing corpus consists of 1390 ten-digit utterances from the Voice Across America long-distance telephone corpus collected by Texas Instruments [2]. The second corpus consists of 1676 ten-digit utterances recorded from a challenging noisy field recognition test conducted by Texas Instruments.

In this paper, we constrain the number of digits recognized per utterance by using a length-constrained grammar. Additional HMMs are included to model silence, breath sounds, and other noise phenomena.

## 4. EXPERIMENAL RESULTS

### 4.1. Baseline Performance

To establish a baseline comparison we use HMM word models. The resulting 22 digit HMMs utilize a total of 438 acoustic distributions. Performance is given in Table 1 below, which shows word correct, word substitution, word deletion, word insertion, word error and sentence error in percent.

Table 1. Baseline HMM Word Models

| Corpus | Corr | Sub | Del | Ins | Err | S. Err |
|---|---|---|---|---|---|---|
| Mac | 97.5 | 2.0 | 0.5 | 0.4 | 2.9 | 17.0 |
| VAA | 95.0 | 3.6 | 1.4 | 0.3 | 5.3 | 26.5 |
| Fld | 91.8 | 4.7 | 3.5 | 0.5 | 8.8 | 35.4 |

### 4.2. Inter-digit models

We then used the 418 inter-digit models trained over the Macrophone corpus to recognize the same utterances as the baseline test. Recognition using the 418 inter-digit models provides notable improvement over the baseline word HMMs as indicated in Table 2.

Table 2. Inter-Digit Models

| Corpus | Corr | Sub | Del | Ins | Err | S. Err |
|---|---|---|---|---|---|---|
| Mac | 98.4 | 1.4 | 0.3 | 0.5 | 2.1 | 11.2 |
| VAA | 97.3 | 2.4 | 0.3 | 0.2 | 2.9 | 18.1 |
| Fld | 94.4 | 3.4 | 2.2 | 0.3 | 5.9 | 26.1 |

## 5. PER-UTTERANCE ADAPTATION

Speaker and environment adaptation can improve recognition performance significantly. Common improvements utilize methods of adapation such as types of cepstral normalization or HMM parameter adaptation [6] [7].

Using the inter-digit models, we tried two simple types of adaptation based only on the utterance to be recognized. One method used simple normalization by subtraction of the utterance log spectral parameter mean from each input utterance frame parameter vector.

The second method adapted the mean vectors of the HMM models using a single affine transformation, $\hat{r} = Ar + b$. We also investigated simplifying assumptions of $A = I$ or $A$ diagonal. To calculate the affine transformation, we constructed two alignments of the test utterance. The first alignment was based on the original models. The affine transformation was computed to minimize the distance between the Viterbi aligned test utterance and corresponding model acoustic distributions. The second alignment was based on the affine transformed models. The transformation should adapt all models to represent the input utterance better. This is basically an unsupervised adaptation based on the same test utterance. Because silence usually accounts for a large portion in the utterance, in computing the transformation, we did not give each input frame equal weight. Instead, we gave each model frame (those that occur in the alignment) equal weight. Because silence is only one frame of model, the silence portion will not inordinately influence the transformation. That is, the transformation is based mostly on speech portion of the test utterance.

Results using both types of adaptation gave improvements in performance. Recognition results using log spectral mean subtraction are shown in Table 3. Tables 4 and 5 show results using affine adapatation with just a bias term and with a diagonal assumption for the matrix $A$. The results in Tables 4 and 5 show that word error was significantly higher for the affine adaptation method. This is due to a large deletion error caused by non-alignment resulting from the beam-limited Viterbi search. When the recognizer does not align an utterance, then all words of the utterance are reported as deletions.

Table 3. Mean Subtraction

| Corpus | Corr | Sub | Del | Ins | Err | S. Err |
|---|---|---|---|---|---|---|
| Mac | 99.0 | 0.8 | 0.2 | 0.4 | 1.4 | 7.5 |
| VAA | 98.4 | 1.5 | 0.1 | 0.1 | 1.6 | 12.0 |
| Fld | 97.6 | 2.2 | 0.2 | 0.2 | 2.6 | 17.6 |

Table 4. Affine with $A$=I

| Corpus | Corr | Sub | Del | Ins | Err | S. Err |
|---|---|---|---|---|---|---|
| Mac | 98.9 | 0.9 | 0.2 | 0.3 | 1.4 | 8.1 |
| VAA | 98.2 | 1.5 | 0.3 | 0.1 | 1.9 | 12.9 |
| Fld | 95.7 | 2.2 | 2.1 | 0.2 | 4.6 | 19.9 |

Table 5. Affine with $A$ Diagonal

| Corpus | Corr | Sub | Del | Ins | Err | S. Err |
|---|---|---|---|---|---|---|
| Mac | 99.0 | 0.9 | 0.1 | 0.3 | 1.3 | 7.9 |
| VAA | 98.5 | 1.3 | 0.3 | 0.1 | 1.7 | 12.9 |
| Fld | 96.0 | 1.9 | 2.1 | 0.3 | 4.3 | 18.7 |

We have also tried using two affine transformations: one for higher energy speech input frames, and another for low energy and silence input frames. We note that there is a natural bi-modal distribution of frame energies. The classification was done by a simple energy threshold. However, the two transformations did not provide better performance than a single affine transformation.

### 5.1. Discriminant Training

We used Viterbi training to obtain the above results. For a small vocabulary application, such as digit recognition, discriminant training has shown significant improvement [5] [4]. To include this type of discriminant processing with the inter-digit models, we used a simple gradient algorithm to adjust the feature mean vectors. We performed two recognition passes on the digit training corpus: one supervised, the other unsupervised. Confusion pairs were identified whenever the two alignments disagreed, and a gradient method was applied to adjust the mean vectors. Including this gradient type of adaptation yielded further improvements. Table 6 shows the results of combining inter-digit models, mean subtraction and gradient training. Table 7 gives the results for affine adaptation with diagonal $A$ and gradient training.

Table 6. Mean Subtraction and Gradient

| Corpus | Corr | Sub | Del | Ins | Err | S. Err |
|---|---|---|---|---|---|---|
| Mac | 99.3 | 0.6 | 0.1 | 0.3 | 1.0 | 6.5 |
| VAA | 98.9 | 1.0 | 0.1 | 0.1 | 1.2 | 9.4 |
| Fld | 98.0 | 1.7 | 0.3 | 0.3 | 2.3 | 15.2 |

Table 7. Affine with Diagonal $A$ and Gradient

| Corpus | Corr | Sub | Del | Ins | Err | S. Err |
|---|---|---|---|---|---|---|
| Mac | 99.2 | 0.6 | 0.2 | 0.2 | 1.0 | 5.7 |
| VAA | 98.5 | 1.0 | 0.5 | 0.1 | 1.6 | 10.0 |
| Fld | 96.1 | 1.5 | 2.4 | 0.2 | 4.1 | 15.8 |

## 6. CONCLUSIONS

The results presented in this paper demonstrate that inter-digit modeling can provide improved digit recognition performance by modeling the coarticulation that occurs between digits. Results presented indicate that this improvement is in addition to improvements obtained by utterance adaptation and discriminative training. On the Macrophone corpus each of the methods provides about 30% improvement in performance.

We noticed during training that word boundary left and right context models were preferred over other left and right context models, leading to an imbalance in training between word boundary context and other context models. We are investigating this phenomenon and designing means of balancing the training.

Obviously, a single mixture acoustic distribution for each state consisting of only a mean vector can not model all of the variability of a state. For example, single mixtures cannot model dialect variations within the digits. We are presently conducting experiments with multiple-mixture inter-digit models. We also plan to determine if we can add mixtures on an *as needed* basis, to keep the number of acoustic distributions at a minimum.

## 7. REFERENCES

[1] J. Bernstein et al. "Macrophone: An American Telephone Speech Corpus for the Polyphone Project," ICASSP 94, Vol 1, pp. 81-84.

[2] B. Wheatley, J. Picone, "Voice Across America : Toward Robust Speaker-Independent Speech Recognition for Telecommunications Applications," Digital Signal Processing, Apr. 1991, pp. 45-63.

[3] T. Matsuoka et al. "Elaborate Acoustic Modeling for Japanese Connected Digit Recognition," IEEE ASR Workshop, Snowbird, Utah, Dec. 1995.

[4] L. Netsch, "Discriminant Methods for Word Acoustic Modeling in Speaker-Independent Digit Recognition," TI Tech Report, 1995.

[5] W. Chou et al. "Minimum Error Rate Training Based on the N-Best String Models," ICASSP 93, Vol 2, pp. 652-655.

[6] F. Liu et al. "Environment Normalization for Robust Speech Recognition Using Direct Cepstral Comparison," ICASSP 94, Vol. 2, pp. 61-64.

[7] A. Sankar , C.H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition," IEEE Trans. Speech and Audio Processing, vol4, no. 3, pp. 190-202, May 1996.