# Improvements in Connected Digit Recognition Using Higher Order Spectral and Energy Features

S5.25

*J. G. Wilpon*
*C.-H. Lee*
*L. R. Rabiner*

AT&T Bell Laboratories
Murray Hill, New Jersey 07974

## ABSTRACT

The problem of recognizing strings of connected digits is crucial to a number of applications such as voice dialing of telephone numbers, automatic data entry, credit card entry, PIN (personal identification number) entry, entry of access codes for transactions, etc. Algorithms for connected digit recognition, based on whole word reference patterns, have become increasingly sophisticated and have been shown capable of achieving high recognition performance. Much of this complexity is derived from the design of specialized word models suitable solely for connected digit recognition.

In this paper we show how we can apply the improved acoustic modeling techniques (using a continuous density hidden Markov model framework), developed for large vocabulary speech recognition applications, to the problem of connected digit recognition with no changes made to the basic modeling techniques and with no vocabulary specific information used. The improved modeling techniques adopted in this study include an improved feature analysis procedure, that incorporates higher order cepstral and log energy time derivatives, and an improved acoustic resolution procedure, that uses more Gaussian mixture components per state to characterize the acoustic variability in each state of the model. Using these techniques, string accuracies of 98.6% for unknown length strings and 99.2% for known length strings were achieved on the standard Texas Instruments connected digits database. These string accuracies are a factor of 2 better than those previously reported using the same modeling procedures [4], and are even somewhat better than those reported by Doddington using specialized modeling techniques for the digits [1].

## 1. Introduction

Connected digit recognition is an extremely important speech recognition task because of its application to such problems as credit card validation, catalog ordering, and digit dialing by voice. In the last several years, several highly successful algorithms for recognizing spoken connected word strings from word prototypes have evolved [1-4]. These algorithms, all based on statistical pattern recognition methods, have achieved great success when applied to the problem of connected digit recognition. The reasons for this success are twofold. First, the recognition algorithms are optimal in the sense that they find the string of digit reference patterns that best (in some objective sense) matches the spoken digit string. Second, there have been several highly successful training procedures developed which derive the digit reference patterns from a training set of fluent connected digit strings [1-5].

To achieve high accuracy on connected digit recognition, a great deal of research has gone into devising specialized word models suitable for connected digit recognition, which rely on the use of separate male/female models, context dependent models for the digits *two* and *four*, confusion class models for likely digit confusions, etc. (see, for example, Doddington [1]). With the use of these specialized modeling techniques, very high performance connected digit recognition accuracy was achieved on the standard Texas Instruments (TI) connected digit database (1.5% string error rate, 0.5% word error rate for unknown length strings). Although excellent performance was obtained on this one task, the techniques and modeling ideas were highly specialized to the digits vocabulary, and hence might not be straightforwardly generalizable to other applications (e.g. large vocabulary speech recognition).

In an effort to improve the performance of continuous density, hidden Markov model (CDHMM) based, large vocabulary recognition systems, the feature analysis was extended to incorporate higher-order time derivatives of cepstral and log energy parameters. The motivation behind this change was the observation that, by including higher order information about the time derivative of the cepstral vector and log energy parameter, a more complete 2-dimensional (time and frequency) representation of the time-varying speech signal is obtained. This had the effect of reducing the overall word recognition error rate by 30% when evaluated on the DARPA 1000-word Resource Management connected speech database [6].

In this paper, we show that the improved acoustic modeling techniques, developed for large vocabulary speech recognition applications, work extremely well on the problem of connected digit recognition, with no changes made to the basic modeling techniques. Using the improved feature set with a *single* hidden Markov model (HMM) per digit, we achieved a string error rate of 1.4% for unknown length strings (with word error rate, which includes substitutions, insertions and deletions of 0.48%) and a string error rate of 0.77% for known length strings (with word error rate of 0.23%) using the standard TI connected digits database. The error rate using the improved model is less than half that of our earlier study, which had fewer spectral parameters and less acoustic resolution in each model state [4]. Hence, without using any vocabulary specific or speaker specific knowledge, the results based solely on improved modeling techniques are comparable to (or actually slightly better than) the best published results obtained using highly specialized models and processing [1].

In Section 2, we discuss the improved feature analysis techniques. In Section 3 we present recognition results from a



CH2977-7/91/0000-0349 $1.00 © 1991 IEEE

series of tests on the TI connected digits database.

## 2. Improved (Expanded) Feature Analysis

A comprehensive description of the complete hidden Markov model based connected word speech recognizer is given in References 2 and 5. In this section, we focus our discussion on the improved front-end feature analysis developed for a large vocabulary recognition system. All other signal processing in the recognizer is essentially identical to that described in Reference 2.

Since we are using a continuous density HMM approach for characterizing each of the models, it is fairly straightforward to incorporate new features into the feature vectors. Specifically, we study the incorporation of higher order time derivatives of short-time cepstral features and log energy features into our continuous speech recognition system. These include: the second cepstral derivatives (called the *delta-delta cepstrum*), the log energy derivative (*delta energy*), and the second log energy derivative (*delta-delta energy*), into our continuous speech recognition system.

### 2.1 Second Order Cepstral Time Derivatives

The incorporation of first order time derivatives of cepstral coefficients has been shown useful for both speech recognition and speaker verification. Thus we were interested in investigating the effects of incorporating higher order cepstral time derivatives. There are several ways to incorporate the second order time-derivative of the cepstral coefficients. Most of the existing approaches evaluate the second derivatives (called the *delta-delta cepstrum*) as the least squares fit to the second difference of each of the cepstral parameters defined over a finite time window. The degree of success in using such a strategy for the delta-delta cepstrum computation has been mixed.

One of the earliest continuous speech recognition systems which used the delta-delta cepstral features, (which were computed as the time derivatives of the first order time derivatives) was reported on by Ney [7]. Ney tested the system for speaker independent recognition of the DARPA 1000-word Resource Management task, and showed a very significant improvement in word accuracy (over the system without higher order cepstral information) when testing the recognizer without using any grammar (i.e. perplexity of 991). The same type of evaluation was studied in the Bell Labs, PLU-based, large vocabulary recognition system [6]. We found that direct incorporation of delta-delta cepstral features gave a 10% word error rate reduction (over that achieved without using delta-delta cepstral features).

From our research in large vocabulary recognition, it was shown that the window size for the second order cepstrum should be of duration 3 frames of first order analysis data (i.e. an overall window duration of 7 frames of speech contributed to the delta-delta cepstrum at a given time). We used the same window length (namely 105 msec with a 15 msec frame rate) in this study. The $m^{th}$ delta-delta cepstral coefficient at frame $l$ was approximated as

$$\Delta_2 \hat{c}_l(m) = G_1 \left[\Delta \hat{c}_{l+1}(m) - \Delta \hat{c}_{l-1}(m)\right] \qquad (1)$$

where $\Delta \hat{c}_l(m)$ is the estimated weighted $m^{th}$ delta cepstral coefficient evaluated at frame $l$ and $G_1$ is a scaling constant which was fixed to be 0.375.

We augment the original 24-dimensional feature vector (12 cepstral 12 delta cepstral coefficients) with 12 additional features derived from Eq. (1) giving a 36-dimensional feature vector. We observed that the second order cepstral analysis produces very noisy observations based on a 45 msec window and a 15 msec frame shift. Of equal concern is the effectiveness of each of the additional features. In Ney [7], a pre-selected set of delta and delta-delta cepstral features was used. An automatic feature selection algorithm, e.g. a principal component analysis, should be used to determine the relative importance of all spectral analysis features. (However we did not use such a feature selection procedure here. This study will be undertaken in the future.)

### 2.2 Log Energy Time Derivatives

Extending the feature vector to include first order time derivatives of the log energy values, known as delta energy, has been shown to be useful [6]. Several systems use both log energy and delta energy parameters as features. In order to use the energy parameter effectively, careful normalization is required. In our large vocabulary system, the log energy parameter was normalized syllabically. We did not include the log energy parameter directly in the feature vector; instead we used the log energy parameter to assign a penalty term to the likelihood of the observed feature vector. However, we have found, from our work in large vocabulary recognition [6], that the delta and delta-delta energy parameters are more robust and more effective recognition features, and therefore we augmented the 36-dimensional feature vector with the delta and delta-delta energy features. Similar to the evaluation of the delta cepstrum the delta energy at frame $l$ is approximated as a linear combination of the log energy parameters in a 5 frame window centered at frame $l$. Since the log energy parameter has a wider dynamic range in value, we replace $G$ in Eq. (1) with a smaller constant (0.0375) for the evaluation of the delta energy. Again, we did not attempt to optimize the results of the $k$-means clustering algorithm by adjusting the normalization constant.

The second order time derivatives of the energy parameters, called the delta-delta energy, are computed similar to the way the delta-delta cepstral features are evaluated in Eq. (1). Again we use a window of 3 frames, and the constant $G_1$ is fixed to be 0.375. Starting with the 24-element feature vector, by adding delta-delta cepstrum, delta energy and delta-delta energy to the feature set, for every frame $l$, we have a 38-element feature vector.

## 3. Experimental Evaluation and Results

### 3.1 Speaker Independent Connected Digits Database

To evaluate the performance of the connected digit recognizer, in a speaker independent mode, we used the standard TI connected digits database [8], as distributed by the National Institute of Standards and Technology (NIST). This database contains connected digit strings from 225 adult talkers (equally distributed among male and female talkers), and was conveniently divided into training and testing sets, for consistency of comparison of results among the different

researchers using this database. This database was dialectically balanced with an equal mix of talkers from 22 dialectical regions. At least 10 talkers (5 male, 5 female) from each dialectical region were included in the database.

As provided by the NIST, the digitized strings were sampled at a 20 kHz rate. For consistency with the telephone bandwidth, all strings were digitally filtered to a 3.2 kHz bandwidth, and downsampled to a 6.67 kHz rate. This downsampled version represents a much lower data rate than what was used in other connected digit recognition systems [1,3]. A total of 8568 training strings and 8578 testing strings were used (a small number of the strings on the digital tapes were unreadable).

### 3.2 Recognition Results for Different Feature Analysis

Table 1 presents the results from a series of recognition experiments to determine the effect of sequentially adding components to the feature analysis vector. In all the experiments, we trained a single hidden Markov model per word, based on a continuous density model, with state observation densities approximated by mixture-Gaussian densities with diagonal covariance matrices. Each model was a standard left-to-right design with 10 states and 64 Gaussian mixture components per state. For this task, the segmental *k*-means training procedure was used [5]. For comparison purposes, a baseline experiment was run using the experimental design in Reference 4, namely using only 9 Gaussian mixtures per state. The recognition error rate (using a single HMM per word) in this test was 3.9% and 2.4% for unknown and known length strings, respectively (this reduces to 2.8% and 1.6% using 2 HMMs per word). We can see from the table that, as new features are added, the recognition string error rate is reduced from 2.9% (with a word error rate of 0.5%), for unknown length (UL) strings (1.8% for known length (KL) strings, with a word error rate of 0.2%) using only the cepstrum and delta cepstrum features, to 1.4% for unknown length strings (0.8% for known length) using both the 1st and 2nd order cepstrum and energy features. The table shows that a feature vector consisting of the cepstrum, delta cepstrum, delta energy and delta-delta energy (a 26 component vector) yielded the same results (2.1% string error rate) as using the cepstrum, delta cepstrum and delta-delta cepstrum (a 36 component vector). This alone represents a 30% reduction in the string error rate. By combining the cepstral and energy derivatives into a single feature vector (a 38 component vector), the string error rate was further reduced to 1.4% (an addition 30% string error rate reduction). Additionally, we evaluated the improved recognition system on the training database itself and achieved a 0.3% and 0.05% string error rate on unknown and known length strings respectively. The error rates of this improved model (which is based solely on improved acoustic parameterization) are almost a factor of *three* less than the baseline result which used fewer spectral features and thus less acoustic resolution in each model state, and are comparable to the best results obtained using highly specialized models (i.e. models tailored to the specific vocabulary) and processing [1].

Figure 1 shows cumulative plots, for the testing set based on both UL (part a) and KL (part b) strings, of the percentage of talkers with string error rate above a threshold. The median UL string error rate is 0.4% (with 58% of the talkers having no errors) and the median KL string error rate is 0.2% (with 74% of the talkers having no errors). The UL string error rates are lower, by a factor of about 4, than the average error rates reported in Table 1, showing that a large percentage of the string errors were generated by a small fraction of the talkers. This result was noted by Bush and Kopec [3] and Rabiner, *et al* [2].

### 3.3 Error Analysis

An analysis of the string errors using the 38 component feature vector and 64 mixture components per state (i.e. our best results) shows the following:

- The most confusable pair of words was *two* being mis-recognized as *oh* (12 out of 68 substitutions). The next most confusable pair was *five* mis-recognized as *four* (5 times). All other confusions occurred less than 5 times. In 27 (out of the 68) of the word substitutions, the word *oh* was the substituted word.
- 25 of the 42 deletions were for the digit *oh* and 15 were for the digit *8*
- The most inserted digit was the digit *oh* (19 out of 27 insertions).
- The digit *two* was the most mis-recognized digit (14 out of 68 mis-recognitions) and *zero* the second most mis-recognized digit (11 times).

The above analysis shows that in about half the digit errors, the word *oh* was involved. This result is to be expected since *oh* can be spoken rather rapidly and therefore is a prime candidate for digit insertion, deletion, or substitution.

An analysis was also made of errors in the recognition of the training set and trends very similar to those discussed above were found.

### 3.4 Effects of Number of Mixtures Per State

To demonstrate the effects of using greater or fewer than 64 mixtures per state, an experiment was run, in which a set of HMM's was designed using the improved feature analysis described above with 9, 16, 32, 64 and 128 mixtures per state. Nominally these figures represent the maximum number of mixtures per state. When insufficient data existed to design the appropriate set of mixtures, the actual number of mixtures per state was reduced appropriately. Again, only a single HMM was designed for each digit. The results of this experiment, in the form of string error rates for UL strings and KL strings for different number of mixtures per state are given in Table 2. It can be seen that as the number of mixtures per state goes from 9 to 32, a 33% reduction in string error rate occurs (from 2.4% to 1.6% for unknown length strings). For runs with 32 to 64 mixture components per state, a small but significant reduction in string error rate was observed (down to 1.4%). However, increasing the number of mixture components per state from 64 to 128 yielded no real performance difference for UL strings. It should be noted that the computation of the local likelihood scores doubles for each factor of 2 increase in the number of mixture components per state. Hence, for real-time implementations one must consider the added value of increased number of mixture components per state.

## 4. Summary

In this paper we have presented results that demonstrate major improvements in our ability to recognize relatively unconstrained strings of connected digits (i.e. strings up to 7 digits in length). We have shown that by incorporating information about the time derivatives of the cepstral coefficients and log energy features, along with instantaneous cepstral coefficients, we can significantly improve recognizer performance. We have also demonstrated the flexibility of using continuous density HMM approaches. The techniques we developed in CDHMM-based large vocabulary recognition can be applied directly to smaller recognition tasks with little or no change in basic modeling strategy.

The incorporation of first and second order time derivatives of the cepstral and energy parameters improved recognition performance significantly, with a top performance of 1.4% and 0.8% string error rate for UL and KL strings, respectively. This is almost a 3-to-1 reduction in the number of errors made using an earlier system, which did not have such a fine acoustic representation, and is comparable to the best published results on this database, which used highly stylized models for each digit. In addition, we have used this improved feature analysis on several other speaker independent connected word databases and have also seen equally impressive reductions in error rates.

We believe that a better feature analysis is one of the key techniques we can benefit most from in the area of acoustic modeling for both connected digit recognition as well as large vocabulary speech recognition.

## REFERENCES


1. G. R. Doddington, "Phonetically Sensitive Discriminants for Improved Speech Recognition," *Proc. ICASSP 89*, Glasgow, Scotland, pp. 556-559, May 1989.
2. L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "High Performance Connected Digit Recognition Using Hidden Markov Models," *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, Vol. 37, No. 8, pp. 1197-1213, Aug. 1989.
3. M. A. Bush and G. E. Kopec, "Network-Based Connected Digit Recognition," *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, Vol. ASSP-35, No. 10, pp. 1401-1413, Oct. 1987.
4. Rabiner, L. R., Lee, C. H., Juang, B. H., and Wilpon, J. G., "HMM Clustering for Connected Word Recognition System," *Proc. of ICASSP '89*, Glasglow, Scotland, pp. 405-408, May 1989.
5. L. R. Rabiner, J. G. Wilpon, and B. H. Juang, "A Model-Based Connected-Digit Recognition System Using Either Hidden Markov Models or Templates," *Computer Speech and Language*, Vol. 1, No. 2, pp. 167-197, Dec. 1986.
6. C.-H. Lee, E. P. Giachin, L. R. Rabiner, R. Pieraccini and A. E. Rosenberg, "Improved Acoustic Modeling for Continuous Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop*, Somerset, PA., June 1990.
7. H. Ney, "Acoustic-Phonetic Modeling Using Continuous Mixture Densities for the 991-Word DARPA Speech Recognition Task," *Proc. ICASSP 90*, pp. 713-716, Albuquerque, NM, April 1990.
8. R. G. Leonard, "A Database for Speaker-Independent Digit Recognition," *Proc. 1984 ICASSP*, pp. 42.11.1-4, March 1984.


| Analysis Type | Size of Feature Vector | String Error Rate (%) | |
|---|---|---|---|
| | | UL | KL |
| cep + dcep | 24 | 2.9 | 1.8 |
| cep + dcep + de | 25 | 2.2 | 1.3 |
| cep + dcep + de + dde | 26 | 2.1 | 1.2 |
| cep + dcep + ddcep | 36 | 2.1 | 1.2 |
| cep + dcep + ddcep + de | 37 | 1.7 | 1.0 |
| cep + dcep + ddcep + de + dde | 38 | 1.4 | 0.8 |

TABLE 1

cep = cepstrum
dcep = delta cepstrum
ddcep = delta-delta cepstrum
de = delta energy
dde = delta-delta engery

| # Mixtures Per HMM State | String Error Rate (%) | |
|---|---|---|
| | UL | KL |
| 9 | 2.4 | 1.2 |
| 16 | 2.1 | 1.1 |
| 32 | 1.6 | 0.7 |
| 64 | 1.4 | 0.8 |
| 128 | 1.4 | 0.9 |

TABLE 2

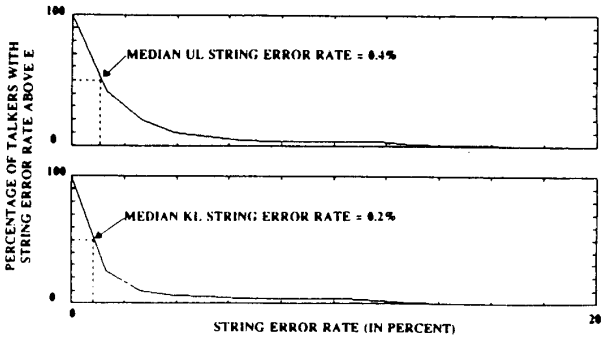


FIGURE 1