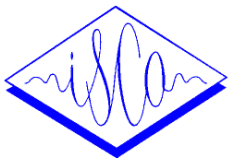

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

CONNECTED DIGIT RECOGNITION USING CONNECTIONIST PROBABILITY ESTIMATORS AND MIXTURE-GAUSSIAN DENSITIES

David M. Lubensky, Ayman O. Asadi, and Jayant M. Naik

Speech Recognition and Language Understanding Laboratory
NYNEX Science & Technology, Inc.
500 Westchester Ave., White Plains, NY 10604, U.S.A.

ABSTRACT

We report on some recent improvements to a continuous density hidden Markov model (CDHMM) based speech recognition system which is being developed for a variety of telecommunication applications. In particular, we are concerned with computing the emission probabilities of an HMM state using a combination of multi-layer perceptrons (MLPs) as probability estimators and mixture-Gaussian densities. Using MLPs as state observation estimators has been shown to improve accuracy on a speaker verification task [11]. In this paper, we describe a connected digit recognition system which incorporates both MLPs and mixture-Gaussian densities. The results are reported on the standard Texas Instruments (TI) connected digit database [9] which was digitally filtered to the telephone bandwidth (300 Hz - 3.2 kHz) and downsampled to 8 kHz. A hybrid MLP/HMM system led to 15% improvement in performance. The final string error rate is 1.7% for unknown length strings.

1. INTRODUCTION

Connected digit recognition is an extremely important task for several telecommunication applications. In particular, we focus on dialing-by-voice services and speaker verification applications in both land-line and cellular environments. Over the last few years, several high performance connected digit recognition systems have evolved [3], [4], [6], [8] and [14]. A great deal of research has gone into two major issues: (1) detailed acoustic modelling, and (2) discriminative training. The first issue is typically addressed by incorporating gender-specific models, a large number of mixtures for estimating emission probabilities of an HMM state, and higher-order time derivatives of spectral or cepstral and energy parameters.

Second, several discriminative training techniques have been explored: maximum mutual information estimation [4], corrective training [7], and either state-specific [6] or class-independent linear discriminative analysis [8].

In this paper, we show how we achieve high accuracy on a connected digit recognition task using a fairly straightforward hybrid MLP/HMM framework. In all our experiments, we trained a single hidden Markov model per word. Each model is a 10-state left-to-right Markov chain with no skip states. Training of MLPs and Gaussian mixture parameters is performed by first obtaining time-aligned state segmentation using a Viterbi decoder. The segmental k-means training procedure is then used to estimate mixture-Gaussian densities with diagonal covariance matrices. We chose the MLP structure such that the output probability density functions (pdfs) correspond to the HMM states. The MLP is trained using the backpropagation algorithm with targets specified by the segmentation.

During recognition we use time-synchronous Viterbi decoding. For each frame of speech we compute state-specific mixture likelihoods and MLP-based posterior probabilities which are then converted to scaled likelihoods by dividing them by the priors (relative state frequencies from the training data). Weighted MLP and mixture-Gaussian likelihoods are combined for the overall state observation probability.

In section 2, we describe the acoustic processing and feature analyses. Section 3 discusses the baseline HMM recognizer, followed by the description of the MLP classifier in section 4. The results are presented in section 5, and conclusions are drawn in section 6.

2. ACOUSTIC PROCESSING

The TI-Digits database was digitally filtered to the telephone bandwidth (300 Hz - 3.2kHz), downsampled to 8kHz, and pre-emphasized for spectral flattening. At intervals of 15 msec (with a 45 msec Hanning window) a 10th-order LPC analysis is performed using the autocorrelation

method. A set of 10 cepstral coefficients is derived from the LPC parameters. The resulting cepstral coefficients are then weighted by a raised sinusoidal window [12]. We also incorporate first- and second-order cepstral and the log energy derivatives, resulting in a 32-component feature vector (*10 cepstral, 10 delta-cepstral, 10 delta-delta cepstral, delta-energy and delta-delta-energy*). The first- and second-order features are approximated by a first-order polynomial over a finite length window of 5 and 3 frames respectively.

The incorporation of first- and second-order time derivatives of cepstral and the log energy features typically requires some sort of normalization in order to bring all the features within the same dynamic range. In particular, feature normalization becomes an important issue when Euclidean distance is used in k-means clustering procedure. From our research in neural networks, we have found that scaling all the features between -1.0 and 1.0 leads to good performance and can be easily achieved using training data statistics. Specifically, we normalize all the features according to the following scheme:

$$\hat{f}_l(k) = \frac{f_l(k) - \mu(k)}{r(k)}, \qquad 1 \le k \le N, \qquad (1)$$

where $N = 32$ in our system, $f_l(k)$ is the feature vector at time frame l, $\mu(k)$ is the sample mean vector computed over all the training data and $r(k)$ is simply the sample range, defined as the difference between the largest observation and the smallest observation over all the training data, or

$$r(k) = max(f(k)) - min(f(k)), \qquad (2)$$

It is important to mention that using sample range $r(k)$ in Eq.(1) resulted in consistently better performance than using sample variance. In fact, on the TI-Digits database, the error rate increased by 20% when the sample range was replaced by the sample variance. We believe that with some additional scaling of log energy derivatives using the sample variance is more robust than simply using the sample range. We did not attempt to optimize the results by adjusting the normalization parameters. In all the experiments, we used Eq. (1) for feature scaling.

3. BASELINE RECOGNITION SYSTEM

To achieve high recognition accuracy on a fixed vocabulary task such as connected digits, a great deal of research has gone into designing highly specialized models: gender-specific models, context-dependent models for *oh-oh* and *three-eight* sequences and word-dependent variable duration models. From a practical point of view, connected digit recognition is an extremely important task for telecommunication applications, therefore applying vocabulary-specific knowledge sources to achieve high accuracy is a reasonable approach. The only drawback to highly specialized models is that it's not easy to extend the same modeling framework to different languages.

Our connected word recognition system is based on whole-word hidden Markov models, where the emission probabilities are estimated using mixture-Gaussian densities with diagonal covariance matrices. In all the experiments, we use 12 models, a single model per word, and one silence model. Each word model is a 10-state left-to-right Markov chain with no skips. We have found that using no skip-state topology reduces substitution and insertion errors while deletion errors remain the same. Training of Gaussian mixture parameters is performed by first obtaining time-aligned state segmentation using Viterbi decoder. The segmental k-means training procedure is then used to estimate mixture-Gaussian densities with diagonal covariance matrices.

Training of Gaussian mixture parameters consists of 2 stages. The first training stage is performed on bootstrap data, and the second phase is on all the training data.

In the first training phase, we manually-endpointed 200-isolated digit strings from 20 speakers (10 male and 10 female). These utterances were linearly segmented into 10 states for bootstrapping word models. We then ran three iterations of segmental k-means training procedure over all single-digit strings in the training database.

In the second training stage, three iterations of segmental k-means training procedure are performed on the entire training set. For each digit string, the grammar corresponding to the string is constructed with an optional silence model at the beginning, between digits and at the end. The state-transition probabilities are re-estimated during training. However, simply assigning fixed a-priori values for loop and forward transition had little effect on performance.

In recognition we use a standard null-grammar with optional silences between the digits. Time-synchronous Viterbi decoding is used for time alignment and scoring.

4. MLP CLASSIFIER

Over the last few years, a large body of research [1], [2], [5], [10], [11] and [13] has shown that feed-forward artificial neural networks (also called MLPs) can be effectively

used as probability estimators in speech recognition systems. In particular, Renals [13] has shown that on the Resource Management database, interpolating MLP context-independent phone probabilities with tied mixture context-dependent phone probabilities resulted in a significant increase in word accuracy. One of the major drawbacks of using MLPs is their requirement for computational resources during training, especially, when it comes to training a large number of context-dependent phone models [10]. We expect that within a few years, faster processing resources will reduce training times significantly, allowing wider use of MLP-based speech recognition systems.

In this paper, we interpret the outputs of a MLP as estimates of *a posteriori* probabilities (discriminant by nature) of output classes which are then used as state observation probabilities in HMMs.

We use a straightforward MLP structure with three layers: an input layer with 32 units corresponding to features described in Section 2, a hidden layer with 80 units which can be interpreted as representing tied mixtures and an output layer with 111 units (11 words * 10 states + 1 silence) such that the output pdfs correspond to the HMM states.

Training of MLPs parameters is performed by first obtaining time-aligned state segmentation using a Viterbi decoder. The MLP is trained using the backpropagation algorithm with targets specified by the segmentation. Our initial training targets were obtained by using mixture-Gaussian segmentation. The MLP training may be viewed as performing state-specific discrimination.

The MLP trained on the TI-Digits database is relatively small, a total of 11,631 weights. Training this network required about 60 passes through the training database of about 1 million feature vectors. On a SPARCstation-10 the system converged in 2 weeks.

5. EXPERIMENTAL RESULTS

Experiments to test the performance of the connected digit recognizer were performed on the TI connected digits database[] as distributed by NIST. The CD version of the database was originally sampled at a 20kHz rate. For consistency with the telephone bandwidth, the database was digitally bandpass filtered from 300Hz to 3.2kHz and downsampled to 8 kHz. A total of 8623 training digit strings and 8700 testing digit strings were used.

In Table 1, we present results from a series of recognition experiments designed to demonstrate that increasing the number mixtures significantly reduces the error rate at the expense of a significant increase in computational requirements; doubling the number of mixtures results in roughly a factor of 2 increase in computational and memory requirements.

# Mixtures	Sub	Del	Ins	Wer	Ser	# Params
1	1.12	0.56	0.75	2.43	7.09	7040
4	0.87	0.34	0.20	1.41	3.78	28160
8	0.61	0.29	0.16	1.06	2.94	56320
16	0.43	0.26	0.14	0.83	2.38	112640
32	0.34	0.21	0.12	0.67	1.98	225280

Table 1: Error rates in [%] for unknown-length digit strings. Sub - substitution; Del - deletion; Ins - insertion; Werr - word error rate; Serr - string error rate. # Params is based on 64 parameters (32 means + 32 variances) * number of mixtures * number of states (excluding silence model).

In Table 2 we show results obtained with the MLP estimate of a probability and a number of experiments demonstrating performance improvements when the MLP likelihood estimations are combined with the mixture-Gaussian likelihoods:

$$\log\left(P(x|b_j)\right) = w_1 \log\left(\frac{P_{mlp}(b_j|x)}{P(b_j)}\right) + w_2 \log P_{Gm}(x|b_j) \quad (3)$$

where P_{mlp} corresponds to the MLP state observation probability, P_{Gm} is the mixture-Gaussian likelihood, and $P(b_j)$ is the prior probability of a state based on relative state frequencies from the training data. In all the experiments, we used a single set of weights $w_1 = 1.5$ and $w_2 = 1.0$; these were empirically determined by performing a small set of recognition experiments.

The results show the following:

- Error rates obtained with a relatively small MLP structure (only 11631 weights, Table 2) are similar to the results derived with the 16 mixtures per state system (112640 parameters, Table1). Clearly, using the MLP for estimating state observation probabilities is both effective in terms of accuracy and efficient in terms of the CPU and memory run-time requirements.
- From Table 2, we observe that combining the MLP and mixture-Gaussian likelihood estimators consistently improves performance with a relatively small increase in computational requirements.

- The most confusable pair of words was *four* being confused with *oh* (14 out of 87 substitutions) closely followed by 11 occurrences of digit *five* confused with *nine*. Out of 52 deletions, digits *oh* and *eight* were deleted 39 and 9 times respectively. As expected, the most frequently inserted digit was *oh* (14 out 29 insertions) followed by *eight* (8 insertions).

State obs Estimate	Sub	Del	Ins	Wer	Ser	# Params
MLP	0.49	0.26	0.14	0.89	2.51	11631
MLP+ 4 Gm	0.65	0.27	0.18	1.1	3.07	11631+ 28160
MLP+ 8 Gm	0.48	0.22	0.15	0.86	2.4	11631+ 56320
MLP+ 16 Gm	0.36	0.22	0.12	0.70	2.07	11631+ 112640
MLP+ 32 Gm	0.30	0.18	0.10	0.59	1.72	11631+ 225280

Table 2: Results with the MLP and Gaussian mixtures

6. CONCLUSIONS

In this paper, we presented results which demonstrate major improvements in recognition rates by combining discriminative MLP-based state observation estimates and mixture-Gaussian likelihoods.

From the recognition results (Table 1) we conclude that a 33% reduction in string error rate is obtained by simply increasing the number of mixtures per state, i.e. from 8 to 32. Of course, these performance improvements come at a price; greater CPU and memory resources, as well as larger training database requirements.

When we combine the MLP and mixture-Gaussian likelihoods, performance of the system is consistently better than simply using mixture-Gaussian or MLPs alone. We think that a hybrid system with the MLP and 16 mixtures per state offers both good performance and reasonable load on computational resources.

In all the experiments reported in this paper, the MLP was trained with only one 32-component feature vector as input. We expect that using multiple frames of left- and right-context will improve performance of our hybrid system.

REFERENCES

[1] Austin, S., Zavaliagkos, G., Makhoul, J., and Schwartz, R., "Speech Recognition Using Segmental Neural Nets," ICASSP-92, San Francisco, CA, March 1992, pp. I-625-629.

[2] Bourlard, H., and Morgan, N., "Continuous Speech Recognition by Connectionist Statistical Methods," IEEE Transactions on Neural Networks, Vol. 4, No.6, pp. 893-909, November 1993.

[3] Buhrke, E., Cardin, R., Normandin, Y., Rahim, M., and Wilpon, J.,"Application of Vector Quantized Hidden Markov Modeling to Telephone Network Based Connected Digit Recognition,", ICASSP-94, Adelaide, South Australia, April 1994, pp. I-105-108.

[4] Cardin, R., Normandin, Y., and De Mori, R., "High Performance Connected Digit Recognition Using Codebook Exponents," ICASSP-92, San Francisco, CA, March 1992, pp. I-505-508.

[5] Chigier, B., and Leung, H., "The Effects of Signal Representations, Phonetic Classification Techniques, and the Telephone Network," ICSLP-92, Banff, Canada, October 1992, pp. 97-100.

[6] Doddington, G., "Phonetically Sensitive Discriminants for Improved Speech Recognition," ICASSP-89, Glasgow, UK, May 1989, pp.556-559.

[7] Gauvain, J., and Lee, C., "Improved Acoustic Modeling with Bayesian Learning," ICASSP-92, San Francisco, CA, March 1992, pp. I-481-484.

[8] Haeb-Umbach, R., Geller, D., and Ney, H., "Improvements in Connected Digit Recognition Using Linear Discriminant Analysis and Mixture Densities," ICASSP-93, Minneapolis, Minnesota, April 1993, pp. II-239-242.

[9] Leonhard, R., "A Database for Speaker-Independent Digit Recognition," ICASSP-84, San Diego, CA, March 1984,pp.42.11.1-42.11.4.

[10] Lubensky, D., "Generalized Context-Dependent Phone Modeling Using Artificial Neural Networks," EUROSPEECH-93, Berlin, Germany, September 1993, pp. 1477-1480.

[11] Naik, J., Lubensky, D., "A Hybrid HMM-MLP Speaker Verification Algorithm for Telephone Speech," ICASSP-94, Adelaide, South Australia, April 1994, pp. I-153-156.

[12] Rabiner, L., Wilpon, J., and Soong, F., "High Performance Connected Digit Recognition Using Hidden Markov Models," ICASSP-88, New York, NY, April 1988, pp. 119-122.

[13] Renals, S., Morgan N., Bourlard, H., Cohen, M., and Franco, H., "Connectionist Probability Estimators in HMM Speech Recognition," IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 1, pp. 161-174, January 1994.

[14] Wilpon, J., Lee, C., Rabiner, L., "Improvements in Connected Digit Recognition Using Higher Order Spectral and Energy Features," ICASSP-91, Toronto, Canada, May 1991, pp. 349-352.