INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

Applying Large Vocabulary Hybrid HMM-MLP Methods to Telephone Recognition of Digits and Natural Numbers

Kristine W. Ma

TR-95-024

May 1995

Abstract

The hybrid Hidden Markov Model (HMM) / Neural Network (NN) speech recognition system at the International Computer Science Institute (ICSI) uses a single hidden layer MLP (Multi Layer Perceptron) to compute the emission probabilities of the states of the HMM. This recognition approach was developed and has traditionally been used for large vocabulary size continuous speech recognition. In this report, however, such a recognition scheme is applied directly to three much smaller vocabulary size corpora, the Bellcore isolated digits, the TI connected digits, and the Center for Spoken Language Understanding Numbers'93 database. The work reported here is not only on developing small baseline systems to facilitate all future research experiments, but also on using these systems to evaluate front-end research issues, and the feasibility of using context-dependency for speech recognition under the hybrid approach developed at ICSI. In addition, using the TI connected digits, the performance of ICSI's baseline system on small vocabulary size speaker-independent task is compared with those of other speech research institutes.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my research advisor, Nelson Morgan, for his guidance, the entire ICSI speech group for their help and ideas, my family for their support, and Patrick Yip for always being there.

I would also like to thank the Electrical Engineering and Computer Science department for the U.S. Department of Education Fellowship that supported my first-year graduate education.

This research is supported by a National Science Foundation grant MIP-9311980 to the International Computer Science Institute.

1 Introduction

At the International Computer Science Institute (and many other speech laboratories)[1, 6, 11, 13], the Multi Layer Perceptron (MLP) has been used to estimate emission probabilities for Hidden Markov Models (HMMs). The HMMs are used as the underlying statistical model of speech for large vocabulary speech recognition systems. Over the years, these MLP networks have expanded up to examples with 4000 hidden units and over a million free parameters, trained on millions of feature vectors. Despite the fact that fast specialized hardware is used at ICSI for training the network, we found that such large systems (requiring a week or more of training time) are not suitable for extensive experimentation with issues such as feature extraction and other aspects of acoustic modeling. Therefore, we decided that we needed to work with tasks that were much smaller, but were difficult and representative enough of larger tasks.

This report describes the research experiments that are conducted using three such small corpora and the lessons that I have learned from them. The first corpus consists of isolated digits collected over the telephone network by Bellcore. This database was used to conduct extensive testings of the robustness of various signal processing front ends (PLP, Log-RASTA, J-RASTA) in the presence of both convolution and additive noise. Such experiments were conducted earlier in [15] using the mixture Gaussian based hidden Markov toolkit (HTK). The work reported here is to integrate those front-end results with ICSI's hybrid HMM-MLP recognition system and compare the performance to that of the HTK system.

The digits corpus is very useful for conducting preliminary research experiments. However, it is an isolated word database, and therefore lessons that we learned from working with this task may not apply to our research on continuous speech recognition. Thus, after some search and discussion with others, we chose the "Numbers'93" task as our next small database. It was being developed at the Oregon Graduate Institute by researchers at the Center for Spoken Language Understanding (CSLU). It is a database of natural numbers spoken continuously and naturally over the telephone, is moderate-sized and is reasonably difficult. Using this database, a range of experiments with context-dependent networks were conducted, resulting in good improvements for this task over our standard context-independent system. Some experiments of this type were performed previously at ICSI, but few variants were explored due to computational costs for the larger tasks.

Finally, to compare our small vocabulary size recognition system performance with that of other speech research sites [2, 3, 5, 11], a recognizer was developed on the TI connected digits using the hybrid HMM-MLP context-independent approach. Since we wished to use the techniques that were developed for large vocabulary recognition on the small tasks, no specialized structures designed with knowledge of the problem (e.g. word models) were used in developing this system. We are trying to avoid solutions that might not generalize well to very large vocabulary recognition, such as whole-word models. This was likely to hurt our performance, but the role of this set of experiments is to provide a faster testbed for ideas to be used on a range of tasks.

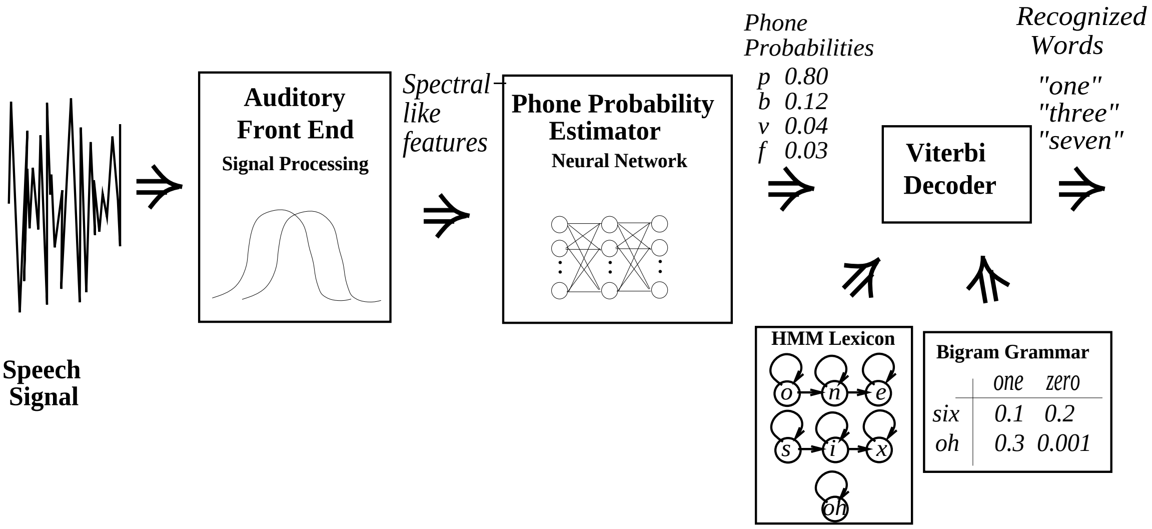

Figure 1: The hybrid HMM-MLP speech recognition system at ICSI.

2 Baseline System

The hybrid HMM-MLP speech recognition system at the International Computer Science Institute (ICSI) has the same basic structure as described in [1]. Its overall architecture is shown in Figure 1. A single hidden layer MLP is employed to estimate the posterior phonetic class probabilities, which are then converted, using Bayes' rule, to likelihood probabilities for Viterbi alignment in the HMM framework.

The three-layer MLP (Figure 2) is trained by using stochastic gradient descent, and relative entropy as the error criterion. A sigmoid function is applied to the hidden layer units, and softmax (exponential of the unit's weighted sum normalized by the sum of exponentials for the entire layer) is used as the output nonlinearity. The network has an input layer of 153 units spanning a window of 9 frames, where each frame consists of 8 cepstra, 8 delta cepstra, and 1 delta energy features. All input features for the network are normalized to have zero mean and unit variance. For the Bellcore digits and the Numbers'93 tasks, the MLP was scaled down from our usual range of 500-4000 hidden units to 200 hidden units. For context-independent recognition, the output layer has 61 units, corresponding to one unit per phonetic class.

ICSI currently uses a decoder called Y0 [13] that applies the standard synchronous Viterbi algorithm. The lexicon consists of single pronunciation word models with repeated states to enforce minimum phonetic durations. The pronunciations are based on the most likely TIMIT pronunciations. A null grammar is used for both the Bellcore and the TI digits recognizers. The language model for the Numbers'93 context-dependent experiments will be described later.

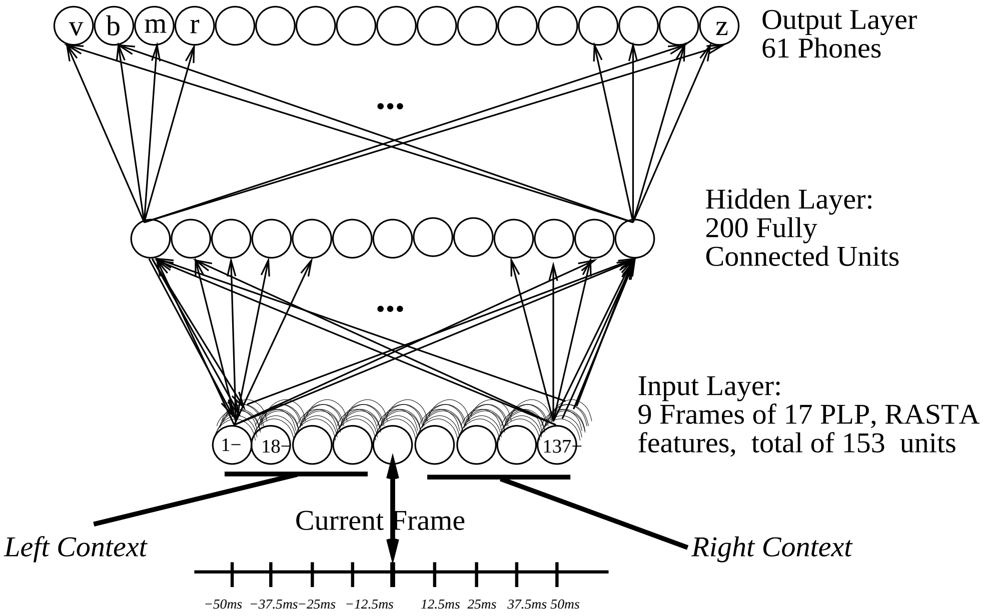

Figure 2: Use a single hidden layer MLP to estimate the posterior phonetic class probabilities $p(q_i|x)$.

3 Training Procedure

Log-RASTA-PLP [8] is used as the acoustic pre-processor for both the TI digits and the Numbers'93 experiments. Each frame of the feature vector represents 25 msec of speech, with 12.5ms overlap of consecutive frames. Log-RASTA-PLP was chosen for its robustness to linear spectral distortions in speech signals that are often introduced by communication channels. This is important for the Numbers'93 database because it is a very realistic set of data that have gone through the public-switched telephone network.

The training of the recognition system is “bootstrapped” from NTIMIT. Such “bootstrapping” is necessary whenever the targeting task is not phonetically labeled. Sometimes, even though a corpus such as the Numbers'93 database is phonetically hand transcribed, we found from past experiences that it is useful to pre-train a neural network from a much larger data set and use it to initiate our task dependent training. The Numbers'93 corpus has about 100,000 frames of training data, while ten times this many frames are available from NTIMIT. The Bellcore digits database has about 30,000 frames of training data for each of the four experiments. The NTIMIT (Network TIMIT) database was collected by transmitting the TIMIT database speech signals over the telephone network. We use NTIMIT (rather than TIMIT) because we are developing recognition systems for telephone bandwidth data.

The first step in the training procedure is to perform a feedforward pass of the digits or Numbers'93 data through a pre-trained NTIMIT net, followed by a phonetic time-alignment of the new corpus using the Y0 Viterbi decoder. This process estimates a set of preliminary target label for the training data. A new MLP is then trained on this set of preliminary alignments. Using this MLP, we reestimate a new set of target labels and so forth. For the training to converge, three or four iterations of forced Viterbi alignment are found to be sufficient. Within each iteration, an independent cross-validation set is used to control the learning rate and to decide when to stop the training. The initial learning rate is kept constant at 0.008 until frame level cross-validation rate improves by less than 0.5%, at which point the learning rate is reduced by a factor of 2. This continues until performance stops improving. This procedure is essentially unchanged from the parameter settings that we used for training up systems with vocabularies of 1000 to 20,000 words, with networks that had over a million parameters.

4 The Bellcore Isolated Digits Front-End Experiment

4.1 Database

The isolated digits task has 13 words in the vocabulary: *1, 2, 3, 4, 5, 6, 7, 8, 9, zero, oh, yes, no.* They are spoken by 200 speakers over the telephone network, producing a total of 2600 utterances. To efficiently use the database, the corpus is divided into four sections to conduct four separate experiments. For each experiment, 150 speakers were used for training (a non-overlapping subset of these were used for cross-validation), while the remaining 50 speakers were used for testing. I report the average of the four experiments as the recognition error rate. The digits database has about 30,000 frames of training data for each of the four experiments.

4.2 Front-End Experiment

Developing speech recognizers that are robust under realistic acoustic environment has been and still is a major problem for the speech community. Speech distortion can be categorized into two types. There is linear spectral distortion (i.e., convolutional noise), which is typically introduced by microphones and the telephone channel. There is also additive noise such as stationary or slowly varying background noise.

Three types of features (PLP, Log-RASTA, J-RASTA) were used to test the robustness of our hybrid recognizer in the presence of convolution and additive noise. Perceptual linear predictive analysis (PLP) is an extension of linear predictive analysis that takes into account some aspects of human sound perception [7]. Log-RASTA is based on PLP but also aims at reducing the effect of linear spectral distortion. J-RASTA tries to handle both linear spectral distortion and additive noise simultaneously. [8]

The purpose of this set of experiments is two-fold. First, we want to verify that the front-end research carried out earlier in [15, 8] on a mixture Gaussian based system will perform equally well using ICSI's hybrid approach. Secondly, we want to compare the performance of the hybrid approach to that of the HTK based system.

4.3 Result

The experimental results are reported in Table 1 and Table 2. All systems are trained with clean (original data collected by Bellcore) data and tested with data that are artificially corrupted with convolutional and/or additive noise (see [15] for detailed description of the noise sources). Systems that are trained without delta features use a 400 hidden unit MLP, while systems with delta features use a 200 hidden unit MLP. Thus the number of parameters in both cases is kept at about 42.8k. All hybrid trainings are bootstrapped from the best NTIMIT nets that are trained on the same front-end feature (although there is evidence from a small test set that bootstrapping from a partially trained NTIMIT net improves overall performance on the new task). Within each jackknived section, the results on all four noise scenarios are from the same Viterbi iteration that yields the best score on clean data.

Table 1: Bellcore isolated digits result using the hybrid approach with delta features.

	clean	filtered	10 dB SNR	10dB+filt
PLP+Deltas	1.4	12.5	42.3	53.7
Log-RASTA+Deltas	1.2	1.8	31.5	27.9
J-RASTA+Deltas	1.2	4.6	9.2	12.4

Table 2: Bellcore isolated digits result using the hybrid vs. HTK approach with no delta features.

	clean	filtered	10 dB SNR	10dB+filt
PLP (HMM+NN)	2.2	15.8	43.8	54.9
PLP (HTK)	5.0	24.9	38.0	50.4
J-RASTA (HMM+NN)	2.0	6.0	10.9	14.8
J-RASTA (HTK)	3.2	5.6	14.8	17.8

The results indicate that:

1. Both types of RASTA feature perform better than PLP on the Bellcore Digits database. J-RASTA is the best front-end feature to use, except when the speech distortion is mainly due to convolutional noise. In this case, Log-RASTA should be applied.

2. From Table 2, we observe that the hybrid HMM-MLP approach performs slightly worse than that of the HTK recognizer under three situations: 1) apply PLP on 10dB SNR speech, 2) apply PLP on 10dB SNR and linear distorted speech, and 3) J-RASTA on linear distorted speech (this last case is not statistically significant at the 5% level). However, one should note that these are precisely the situations where we do not have to pay too much attention to because of item 1. In other words, the hybrid approach performs better when the RASTA features are applied appropriately according to the noise scenario. In comparing the two systems, one should note that the hybrid approach uses 42.8k parameters, while the HTK system

uses only 7.1k parameters [15]. However, in previous work in the group, the number of parameters was increased by un-tying variances, etc. This did not appear to help performance for this task [12]. Another major difference between the two systems is that the HTK recognizer is whole-word model based, whereas the hybrid approach is phonemic based. I should also note that the local expertise at ICSI is focused on the hybrid approach. It is certainly possible that HTK experts could extract better performance from this system.

A confusion matrix of the recognition result (using J-RASTA) on clean speech is shown in Table 3. The leftmost column indicates the correct answer and the topmost row is the answer returned by the recognizer. From the mapping, we see that the error is spread across words except for the word *no*. There are 6 incidences where the word *no* is being confused with the word *zero*, 2 incidences with the word *two*, and another 2 incidences with the word *oh*. If one were to evaluate the performance only on the digits by taking out the two control words *yes* and *no* from the test set, the word error rate would be reduced from 1.2% to 0.8%.

Table 3: Confusion matrix on Bellcore isolated digits tested on clean speech.

	z	1	2	3	4	5	6	7	8	9	o	y	n
z	200	0	0	0	0	0	0	0	0	0	0	0	0
1	0	200	0	0	0	0	0	0	0	0	0	0	0
2	2	0	197	0	0	0	0	0	0	0	0	0	1
3	0	0	0	199	0	0	0	1	0	0	0	0	0
4	0	0	0	0	199	1	0	0	0	0	0	0	0
5	0	0	0	0	1	197	0	1	0	1	0	0	0
6	0	0	0	0	0	0	199	1	0	0	0	0	0
7	0	0	0	0	0	1	0	199	0	0	0	0	0
8	0	0	0	0	0	0	0	0	200	0	0	0	0
9	0	1	0	1	0	2	0	1	0	195	0	0	0
o	0	0	0	0	0	2	0	0	0	1	197	0	0
y	0	0	0	0	0	0	0	0	0	0	0	199	1
n	6	1	2	0	0	0	0	0	0	1	2	1	187

5 The Numbers'93 Context-Dependent Experiment

5.1 Database

As noted earlier, the Numbers'93 corpus is a continuous-speech database collected by the CSLU at the Oregon Graduate Institute. It consists of numbers spoken spontaneously over telephone lines on the public-switched network. These numbers are extracted from the addresses spoken by the callers of CSLU's Spelled and Spoken Names Corpus [4]. The Numbers'93 database consists of 2167 speech files of spoken numbers produced by 1132 callers. I used 1534 of these utterances for training (1117749 frames) and development (86,630 frames), saving the remaining utterances for final testing purposes. There are 36

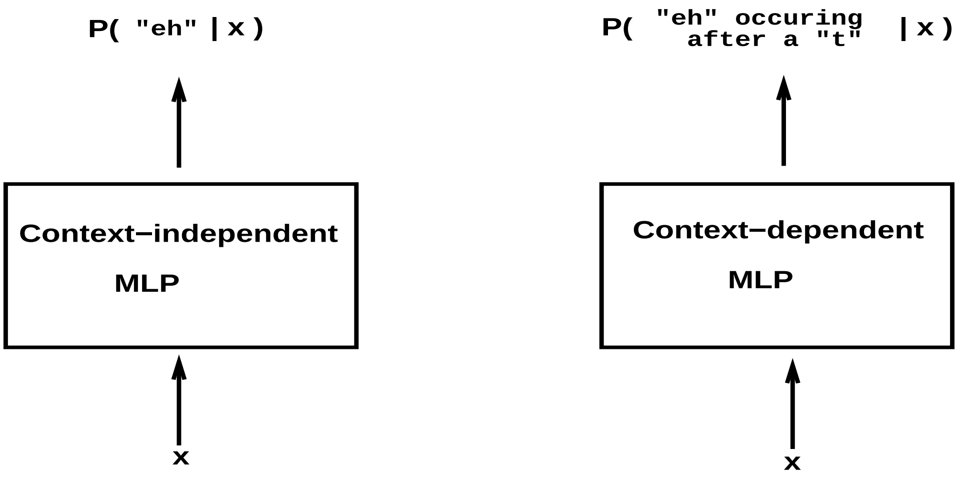

Figure 3: Context-independent vs. context-dependent system.

words in the vocabulary, namely *zero, oh, 1, 2, 3,...,20, 30, 40, 50,...,100, 1000, a, and, dash, hyphen,* and *double.* This task, including a much larger collection of spontaneously spoken numbers, is being made publically available from CSLU.

5.2 Context-Dependent Experiments

One way to model speech coarticulatory effects in speech recognition is to use context-dependent phonetic models. For each of the experiments described in this section, the phonetic models are reformulated to capture various degrees of contextual information to do recognition. This is achieved by training different types of context-dependent MLPs to compute the emission probabilities required by the context-dependent state model for the Viterbi algorithm.

The basic difference between a context-independent MLP and a context-dependent MLP is illustrated in Figure 3. Using the hybrid approach for context-independent recognition, the MLP is trained to estimate the probability of a phonetic class given the current feature vector. For context-dependent recognition, the MLP is trained to discriminate not only among phonetic classes, but also the phonetic contexts under which each phonetic class occurs in. The example shown in Figure 3 is a left-biphone context-dependent MLP.

I experimented with three types of context dependence. They were

I. single state generalized triphone models,

II. single-state triphone models, and

III. multiple state phonetic models with generalized biphones.

Figure 4 provides an example for each one of them. The topmost Markov chain illustrates

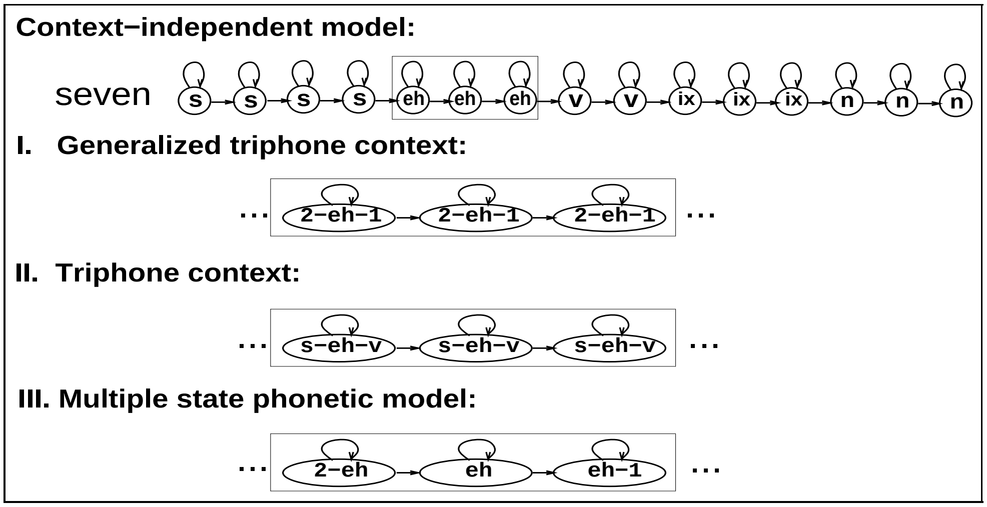
Context−independent model:
seven
s
s
s
s
eh
eh
eh
v
v
ix
ix
ix
n
n
n
I. Generalized triphone context:
2−eh−1
2−eh−1
2−eh−1
II. Triphone context:
s−eh−v
s−eh−v
s−eh−v
III. Multiple state phonetic model:
2−eh
eh
eh−1

Figure 4: Examples for generalized triphone, triphone, and multiple state phonetic model.

a context-independent model for the word *seven*. Under triphone context-dependent modeling, each phoneme is further classified according to its left and right phonemic context. In generalized triphone context, it is only classified according to the generalized category to which its left and right neighbor belongs. There are eight generalized broad categories, and they are clustered according to the place of articulation (see Appendix A). In an alternate approach to modeling of contextual effects, the multiple state phonetic model method expands the single phonetic state model into a three state model—a context-independent middle state, a generalized left-biphone dependent first state, and a generalized right-biphone dependent last state.

5.3 Context-Dependent Training Procedure

All three approaches are bootstrapped from a similar baseline system as described in Section 3, except that the generalized triphone and the triphone methods require an MLP with a larger output layer of 90 and 111 units respectively, while the multiple state phonetic model approach utilizes a different connectionist architecture similar to the one describe in [6]. To support the multiple state phonetic model formulation a connectionist probability estimator consisting of 17 MLPs is used, with 8 nets corresponding to each of the 8 generalized left-biphones, 8 nets for the generalized right-biphones, and 1 net for the context-independent states (see Figure 5).

Thus in approach I and II, the probability estimator is trained to estimate the posterior

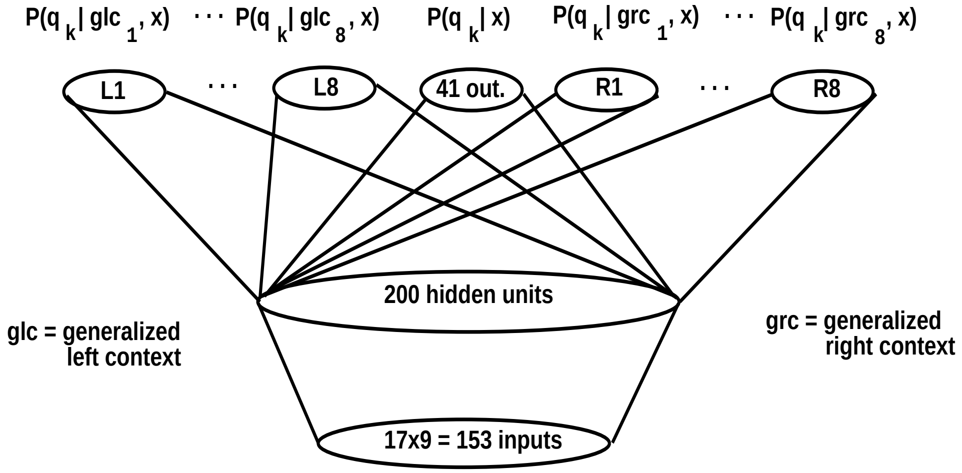

Figure 5: The posterior probability estimator for the context-dependent multiple state phonetic model.

probability $p(q_i, lc_j, rc_k|x)$, i.e. the probability of the current frame being a phoneme q_i occurring between left context class lc_j and right context class rc_k. This probability is related to the likelihood $p(x|q_i, lc_j, rc_k)$ by Bayes' rule as follows:

$$p(x|q_i, lc_j, rc_k) = \frac{p(q_i, lc_j, rc_k|x)p(x)}{p(q_i, lc_j, rc_k)}. \tag{1}$$

In approach I, lc_k and rc_k refer to the eight generalized left biphone, and the eight generalized right biphone, respectively. In approach II, they refer to the 41 left biphone, and the 41 right biphone classes, respectively. The prior $p(q_i, lc_j, rc_k)$ can be estimated by counting over the training examples.

In approach III, the set of 17 MLPs is trained to estimate $p(q_i|x, c_j, s_k)$, where q is one of the 41 phonetic classes, x is the feature vector, c is the generalized biphone context class, and s indicates whether it is the left, right, or context-independent states. Again, this posterior probability can be converted to likelihood via Bayes' rule as follows:

$$p(x|q_i, c_j, s_k) = \frac{p(q_i|x, c_j, s_k)p(c_j, s_k|x)p(x)}{p(q_i, c_j, s_k)}, \tag{2}$$

where $p(c_j, s_k|x)$ is computed by an additional MLP trained to discriminate contexts and states, and the prior $p(q_i, c_j, s_k)$ can be estimated by counting over the training examples.

The major problem encountered with training context-dependent systems is the lack of data for training highly specific phonetic context classes. One solution, adapted from [6], is to initialize the context specific MLP training with weights from a more general context net. Thus, the generalized triphone context-dependent network is trained by bootstrapping from the previously trained context-independent MLP. Similarly, the context-dependent triphone net and the set of multiple state model MLPs are trained by bootstrapping from the generalized triphone context-dependent MLP. To limit the number of free parameters and training time for the context-dependent nets, only the hidden-to-output weights were trained. No degradation in recognition performance was found in comparison with training all the weights, as determined by a pilot experiment that was done on the generalized triphone context-dependent training.

5.4 Language Model

For the Numbers'93 task, the language model is a class bigram derived from the statistics of the training set. This language model was chosen because the Numbers'93 training set is not large enough to provide examples for all possible word transitions between numbers. A simple bigram was therefore not sufficient. I also experimented with a bigram with add-one smoothing, but a simple class bigram was found to be more appropriate. The ten classes are clustered as follow:

- *oh*
- *0*
- *1, 2, 3, 4, 5, 6, 7, 8, 9*
- *11, 12, 13, 14, 15, 16, 17, 18, 19*
- *10, 20, 30, 40, 50, 60, 70, 80, 90*
- *100, 1000*
- *and*
- *dash, hyphen*
- *double*
- *a*

Using this class bigram improves the context-independent recognition performance from 10.7% (null grammar) to 8.2% word error rate.

5.5 Result from Pilot Experiments

The results on the Numbers'93 development set for the context-dependent experiments are shown in Table 4. From the results under the column labeled "Context-dependent Prior", one sees that incorporating context dependency does improve performance, but the recognition performance on the Numbers'93 development set did not continue to improve when finer contextual units are incorporated. This is probably due to the lack of training data for those context specific examples (even though we smoothed the training with the context-dependent net). One solution is simply to use a more general context classification, or methods such as dynamic multi-level context clustering as proposed in [10]. The first three context-dependent experiments do not take into account of cross-word context. However, the fourth experiment does model cross-word context, which resulted in an improvement from 6.9% to 6.2% on word recognition error rate using "Context-dependent Prior".

Table 4: Word recognition error in % on the Numbers'93 development set for the context-dependent experiments.

	CD Prior	Mix Prior	Trained Params	Active Params
Context-Independent (200HU)	8.2	-	42.8K	38.8K
Gen-triphone Context	6.9	6.1	48.8K	48.6K
Triphone Context	6.7	6.1	52.8K	52.8K
Multi-phonetic-state Context	6.8	6.7	170.0K	58.2K
Gen-triphone w/ Xword-Context	6.2	6.2	79.8K	72.0K
Context-Independent (400HU)	7.2	-	85.6K	77.6K

The results under "Mix Prior" is obtained using a slightly different procedure from that of Equation 1 or Equation 2 in converting context-dependent posterior to context-dependent likelihood. The following heuristic is used [6]:

$$p(x|q_i, c_j) = p(q_i, c_j|x)(\alpha_i^j \frac{1}{p(q_i)} + (1 - \alpha_i^j)\frac{1}{p(q_i, c_j)})p(x), \tag{3}$$

where,

$$\alpha_i^j = \frac{N_{ci}(i)}{N_{ci}(i) + b(N_{cd}(i,j))}.$$

The number $N_{ci}(i)$ is the amount of training examples for phonetic class i for the context-independent net, whereas $N_{cd}(i,j)$ is the number of training examples for the context-dependent net for the phonetic class i and for the context class j. The parameter b is adjusted to optimize performance. Note that, as usual, the factor $p(x)$ is ignored during recognition as it is independent of the state.

The idea behind this method is that normally when we convert a context-dependent posterior to a likelihood, we simply divide it by the context-dependent prior (Equation 1, 2), but since the context-dependent training was initialized with context-independent weights, the conversion factor should take into account of both the context-independent and context-dependent priors. The weighting factor α takes into account of the number of training examples used for both the context dependent net and the context independent net.

The last two columns of the table indicate the number of parameters used for each of the experiments. "Trained Params" refers to the actual size of the MLP used for recognition, and "Active Params" includes only the number of parameters that supported active output units. For example, the size of the context-independent MLP is (153 input + 61 output) x 200 hidden = 42.8k parameters. But since the Numbers'93 corpus only contains 41 phonetic classes, therefore the number of "active parameters" is (153 input + 41 output) x 200 hidden = 38.8k. That is, 20 of the outputs were always trained with targets of "0" because these phones never occurred in the training set. These 20 outputs were only used for consistency with our bootstrapping NTIMIT network.

Looking at the results obtained from this set of context-dependent experiments, we felt that it is necessary to verify whether the improvements obtained using generalized-triphone with cross-word context-dependency was mainly due to the increase in the number of parameters or from using a more refine context-dependent acoustic model. Hence, a new context-independent system was developed using a 400 hidden unit MLP. Comparing the results from the last two rows of Table 4, generalized triphone with cross-word context-dependency model improves performance from 7.2% word error to 6.2% word error (on the Numbers'93 development set) with similar number of parameters. Thus the previous improvements from 8.2% (context-independent) to the 6.2% (cross-word context-dependent) word error rate was partly due to the number of parameters increase and partly due to incorporating context-dependency.

New language model

A class bigram might be over smoothing the word pair transition probabilities because the corpus was obtained from street addresses and zip codes. To test this, a new language model was obtained by merging the class bigram with a bigram by averaging their corresponding word pair transition probabilities. This improves the context-independent (200HU) word level recognition error rate from 8.2% to 7.5%, and the generalized triphone with cross-word context-dependency model from 6.2% to 5.8% (Table 5). All subsequent experiments on the Numbers'93 task are performed using this merged language model.

Table 5: Word recognition error in % on the Numbers'93 development set with an improved language model.

	Class bigram	Merged grammar
Context-Independent (200HU)	8.2	7.5
Gen-triphone w/ Xword-Context	6.2	5.8

5.6 Result on Final Test Set

Table 6: Word recognition error in % on the Numbers'93 final test set for the context-dependent experiments. This set of results were obtained using the same parameter settings as the development set.

	Word Error	Trained Params	Active Params
Context-Independent (200HU)	11.5	42.8K	38.8K
Context-Independent (400HU)	11.6	85.6K	77.6K
Gen-triphone Context (Mix Prior)	9.2	48.8K	48.6K
Gen-triphone w/ Xword-Context (Mix Prior)	8.1	79.8K	72.0K

For the Numbers'93 test set, increasing the size of the context-independent MLP estimator to include context-dependent units yields an error rate reduction of 30%. This is at the cost of an 86% increase on parameter size. On this test set, all the improvement appears to be due to incorporating contextual information. The final test set provided by CSLU contains 591 sentences, but 207 of these were eliminated from the final testing because they contain ordinal numbers, which did not appear in the training nor the development sets. I feel that a much larger test set may be required in order to have a more precise evaluation of the various systems. Nevertheless, the results from both the development and the test sets indicates that incorporating context-dependence improves recognition performance significantly.

The most confusable pair of digits in the Numbers'93 database is *four* and *oh*. Moreover, a significant percentage of the errors in the Numbers'93 database is due to the confusibility between utterance strings such as *four ten* vs. *fourteen*, *six eighty* vs. *sixty*, *eleven* vs. *seven*, *fifteen* vs. *fifty* vs. *sixty* vs. *sixteen*, etc. Also, the word *oh* accounts for more than half of the deletion error rate.

6 Evaluation on the TI Connected Digits

Both the Bellcore Isolated Digits and the Numbers'93 database have been used by other speech research laboratories, but the results have not been widely published. Despite its small size, the Numbers'93 corpus is difficult due to its inherently high confusibility factor in the vocabulary, spontaneity of the utterances, and the acoustic channel effects introduced by the telephone network. CSLU plans to distribute the Numbers corpus, and so we hope to see how other sites compare on this task in the coming year.

In the meantime, we tested our recognizer on the standard TI/NIST Connected-Digits Recognition Task ("TI-Digits") [9]. While this corpus is inherently less realistic since it was recorded in an artificial studio situation with wide bandwidth, it has been used by many well developed systems [2, 3, 5, 11], so that it will be a calibration point for our methods on small vocabulary size tasks. Ever since this corpus was made available in 1984, it has became a quasi standard for benchmarking small vocabulary speaker-independent recognition systems. The error rate has reduced by more than a factor of 5 since the first published results.

Most of the current state of the art TI-Digits recognizer systems use whole word modeling, cross word context-dependency, gender dependent models, and more elaborate training procedures. The following was obtained from a literature search:

- The best recognition performance is currently from Bell Laboratories [3]. Their system yields word and string error rates of 0.24% and 0.72%, respectively. They use LPC-12, delta, and delta-delta features for the front end; whole word model with inter-word triphone context dependency, and continuous density HMM for acoustic modeling. The system uses a new training technique that minimizes string error rate, and N-best for decoding.

- Another speech site is Philips Laboratory [5]. Their system also uses whole-word model, and with gender dependency and background noise model. Training and recognition were done using Viterbi search. One interesting twist is that they use continuous Laplacian (rather than Gaussian) emission density for the HMM. Their report is concentrated on reducing recognition error rate by performing energy thresholding, spectrum normalization, linear discriminant analysis, and increasing the acoustic resolution. As a result, the error rate is 0.28% on word level and 0.84% on string level.

- Another effort comes from Centre de recherche infromatique de Montréal (CRIM) [2]. They use Mel-cepstral-12, delta, and delta-delta features for the front end. Again, a whole-word gender dependent model was used with triphone context dependent and discrete HMM. This group used a discriminant training algorithm based on maximum mutual information estimation (MMIE). The MMIE training is initialized with a few iterations of Baum-Welch. Their best performance is 0.28% word error rate and 0.84% string error rate.

- The last speech site was from NYNEX [11]. They use a less elaborate system that does not model gender dependency nor context dependency. Nevertheless, their system is still whole-word model based. They use a hybrid MLP/HMM approach with 10 states per word. LPC-10, delta, and delta-delta features were used for front-end processing. The resulting error rate is 0.89% on the word level and 2.51% on the string level. Their best result is obtained by combining the likelihood estimations from this MLP system with mixture-Gaussian likelihoods, and the resulting error rate is 0.59% on the word level and 1.7% on the string level.

6.1 Database

The TI connected digits database was collected in an acoustically treated sound room and digitized at 20 kHz. It has 11 words in the vocabulary: *1, 2, 3, 4, 5, 6, 7, 8, 9, zero, oh.* Only the adult speakers are used for training and testing. There are 225 speakers, each providing 77 utterances:

- 22 isolated digits
- 11 two-digit sequences
- 11 three-digit sequences
- 11 four-digit sequences
- 11 five-digit sequences
- 11 seven-digit sequences

A total of 8623 sentences (1,202,889 frames) are used for training and development, while 8700 sentences are used for testing.

6.2 Experiment

The speech research community has been using this database to develop telephone bandwidth digits recognizer, and therefore it is common to downsample the data to 8kHz prior to training and testing. However, there doesn't seem to be a standard for this procedure. Some research sites bandpass filtered the data from 100Hz to 3.8kHz, some did not mention filtering in their paper, and some bandpass filtered it from 300Hz to 3.2kHz. We decided to bandlimit the signal to telephone bandwidth so that we could at least approximately compare with other sites' results. For our experiment, the speech data is digitally filtered to telephone bandwidth (300Hz - 3.2kHz) and downsampled to 8kHz. The lowpassing process is performed as part of the decimation procedure. Highpass is then added to eliminate low frequency components of the signal. The frequency responses of the highpass (127 taps) and lowpass (183 taps) filters are given in Figure 6 and Figure 7, respectively. Both filters are designed using the Kaiser windowing method via esps. Log-RASTA-PLP [8] is then used as the acoustic pre-processor.

The size of the system and the training procedure are as described in Section 2 and Section 3, respectively.

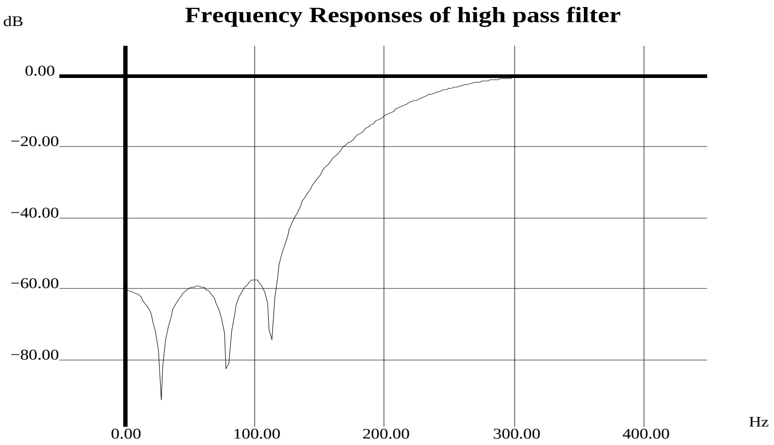

Figure 6: Frequency response of the highpass filter for processing the TI connected digits.

6.3 Result

A recognizer was developed on the TI connected digits using the hybrid HMM-MLP context-independent approach. Since we wished to use the techniques that were developed for large

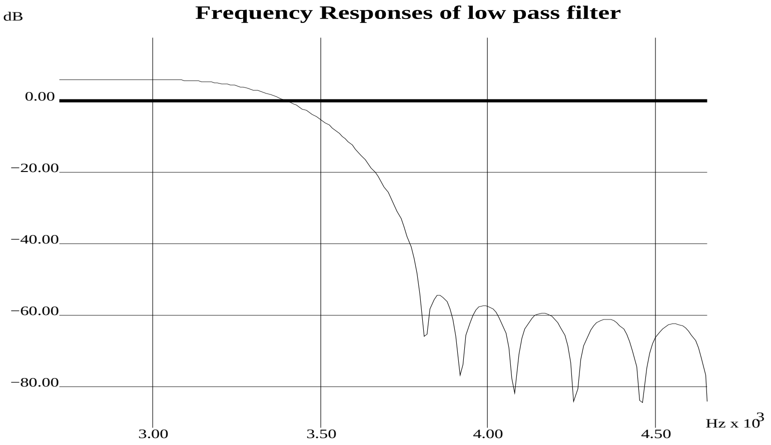

Figure 7: Frequency response of the lowpass filter for processing the TI connected digits.

vocabulary recognition on the small tasks, no specialized structures designed with knowledge of the problem were used in developing this system. However, to fully take advantage of the tremendous number of training patterns, we used a larger MLP than we had used on Bellcore digits or Numbers'93, with 1000 hidden units. The recognition result after four iterations of training was 0.9% word error rate and 2.7% string error rate.

Comparing with other corpora that I have been working with, the TI connected digits corpus is a much more uniform speech database. In corpora such as the Bellcore Isolated Digits Task and the CSLU Numbers'93 task, there are about 3 or 4 utterances that were spoken at rates that are faster than twice the average speaking rate of the sentences from the database. Both corpora have less than 2000 sentence strings. However, the TI corpus has 17,323 sentence strings and the fast speaking rate phenomenon did not occur at all. In light of the database consistency and the vast amount of training data, using whole-word modeling should yield a significant performance improvement. However, since whole-word modeling is not the objective of ICSI's research on large vocabulary size continuous speech recognition, a more appropriate approach would be to experiment with new training algorithms, gender dependency, and cross-word context dependency in order to bring the performance of this particular system up to those of other speech research institutes.

7 Conclusion

In this paper, I have reported various experimental results on extending ICSI's speech recognition method, developed for large vocabulary size continuous speech recognition, onto three small size corpora of digits and numbers. From working with these three small databases, we have found that training and recognition could be performed in virtually the same way that we have done for our large vocabulary size tasks, and that further specialized knowledge was not required. A series of front-end experiments were made possible due to the moderate size of the Bellcore isolated digits database. Moreover, a wide range of experiments with context-dependent networks were also performed, resulting in good improvements for this task over our standard context-independent system. Comparing the performance of our baseline system on the two digits tasks, it is found that the Bellcore digits database (1.2% word error rate) appears to be a more difficult task than the TI digits corpus (0.9% word error rate) even though the former corpus consists of only isolated words. This highlights the distinction between data collected under a realistic environment verses data collected under artificial conditions. It may also reflect the greater amount of training date for the TI task. Finally, the TI connected digits database was used to evaluate our baseline recognition approach on speaker-independent small vocabulary size tasks. The error rate from our baseline system is higher than a number of other dedicated systems reported over the last few years, but we believe that a more comparable result can be obtained if we were to incorporate cross-word context-dependency, use whole-word model, and more elaborate training procedures. It is hoped that other laboratories will soon publish results on more realistic digits/numbers corpora, so that we will be able to verify whether our results are more comparable for such tasks.

A The Place of Articulation

The place of articulation is defined as the location in the vocal tract of the blockage or restriction of the airflow in the production of sounds. The TIMIT phone set is clustered as follow [14]:

0. Silence: *bcl, dcl, gcl, pcl, tcl, kcl, h#, pau, epi, q*
1. Labial (upper or lower lip): *b, p, f, th, v, dh, m, em*
2. Alveolar (tongue touching gum ridge): *d, t, dx, jh, ch, s, sh, z, zh, n, nx, en, l, el*
3. Velar (soft palate): *g, k, ng, eng, hh, hv*
4. Rhotocized: *r, er, axr*
5. Bilabial + glide: *w, aw, ow*
6. Palatal glide + front vowels: *y, iy, ih, eh, ey, ae, ay, oy, ix*
7. Mid to back vowels: *aa, ah, ao, uh, uw, ux, ax, ax-h*

References

[1] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach.* The Kluwer International Series in Engineering and Computer Science. VLSI, Computer Architecture, and Digital Signal Processing, Boston, Massachusetts, Kluwer Academic Publishers, 1994.

[2] Regis Cardin, Yves Normandin and Evelyne Millien. "Inter-Word Coarticulation Modeling and MMIE Training for Improved Connected Digit Recognition," Proceedings of the International Conference on Acoustics Speech and Signal Processing, II.243-246, Minneapolis, MN, 1993.

[3] W. Chou, C.-H. Lee, B.-H. Juang, "Minimum Error Rate Training of Inter-word Context Dependent Acoustic Model Units in Speech Recognition", Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan, September 1994.

[4] R.A. Cole, M. Fanty, and T. Lander, "Telephone Speech Corpus Development at CSLU," Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan, September 1994.

[5] R. Haeb-Umbach, D. Geller, H. Ney, "Improvements in Connected Digit Recognition Using Linear Discriminant Analysis and Mixture Densities," Proceedings of the International Conference on Acoustics Speech and Signal Processing, II.239-242, Minneapolis, MN, 1993.

[6] Horacio Franco, Michael Cohen, Nelson Morgan, David Rumelhart, and Victor Abrash, "Hybrid Neural Network/Hidden Markov Model Continuous-Speech Recognition," Proceedings of the International Conference on Spoken Language Processing, pp. 915-918, 1992.

[7] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis for Speech," J. Acoust. Soc. Am., pp. 1738-1752, 1990.

[8] Joachim Koehler, Nelson Morgan, Hynek Hermansky, H. Guenter Hirsch, and Grace Tong,"Integrating RASTA-PLP into Speech Recognition," Proceedings of the International Conference on Acoustics Speech and Signal Processing, Adelaide, Australia, 1994.

[9] R. Leonard, "A Database for Speaker-Independent Digit Recognition," Proceedings of the International Conference on Acoustics Speech and Signal Processing, pp.42.11.1-42.11.4, San Diego, CA, 1984.

[10] D. Lubensky, "Generalized Context-Dependent Phone Modeling Using Artificial Neural Networks," Proceedings of *EUROSPEECH'93*, September 21-23, Berlin, Germany.

[11] D. Lubensky, A.O. Asadi, and J.M. Naik, "Connected Digit Recognition Using Connectionist Probability Estimators and Mixture-Gaussian Densities," Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan, September 1994.

[12] Nelson Morgan, Personal Communication, 1995.

[13] T. Robinson, L. Almeida, J.M. Boite, H. Bourlard, F. Fallside, M. Hochberg, D. Kershaw, P. Kohn, Y. Konig, N. Morgan, J.P. Neto, S. Renals, M. Saerens, & C. Wooters. (1993). "A Neural Network Based, Speaker Independent, Large Vocabulary, Continuous Speech Recognition System: The WERNICKE Project," Proceedings of *EUROSPEECH'93*, September 21-23, Berlin, Germany.

[14] Gary Tajchman, Personal Communication, 1994.

[15] Grace C.H. Tong. "Combating Additive Noise and Spectral Distortion in Speech Recognition Systems with JAH-RASTA," Masters Thesis, University of California at Berkeley, 1994.