NATURAL NUMBER RECOGNITION USING MCE TRAINED INTER-WORD CONTEXT DEPENDENT ACOUSTIC MODELS

Malan B. Gandhi and John Jacob

Lucent Technologies Bell Laboratories
2000 N. Naperville Rd., Naperville, IL 60566, USA
{mgandhi, jjacob}@lucent.com

ABSTRACT

Among applications that require number recognition, the focus has largely been on connected digit recognizers. In this paper, we introduce an acoustic model topology for natural number recognition by using minimum classification error (MCE) training of inter-word context dependent models of the head-body-tail (HBT) type. Experimental results on natural number applications involving dollar amounts and U.S. telephone numbers show that using HBT models for natural number data reduces string error rates by as much as 25% over context independent whole word models. In addition, for speech input which is strictly of connected digit type, the increase in string error rates is negligible when a natural number telephone grammar is used instead of a connected digit telephone grammar. This will enable natural number speech recognition systems to be more widely accepted because recognition accuracy is maintained while permitting a more natural and flexible user interface.

1. INTRODUCTION

Much of the work which has been reported for connected digit recognition has centered around recognition of telephone, credit card, or personal identification numbers. In the United States, most users are comfortable with the practice of speaking these digit strings as a contiguous sequence of isolated digits. For instance, when speaking the number "51" within the context of a telephone number, users are more likely to say "five one," as opposed to the way they might naturally say the same digit pair in a different context, i.e. "fifty one." In this paper we refer to the latter case as a natural number.

Previous work on natural number recognition has been reported by several authors [4-8]. In [4], Jacobsen and Wilpon compared the use of whole-word, context independent subword models with triphonic context dependent subword models for recognizing Danish telephone numbers. C. de la Torre *et al.* [7, 8] reported results for connected number recognition in Castilian Spanish using semi-continuous Hidden Markov Models (HMMs). In [4] and [7], the authors reported results for languages other than English, and noted the prevalent use of natural numbers in many countries. The use of neural networks for both context independent and context dependent natural number recognition was studied by Ma and Nelson in [5]. In [6], Ramesh and Wilpon proposed the use of state duration modeling and included a natural number speech corpus in their testing data.

As a result of assumptions and constraints on speaker behavior, digit string recognition, in its more tractable form of connected digit recognition, has been successfully addressed, with high performance recognizers reporting string accuracies in the mid 90% range. As speech recognition gains visibility and finds a wider range of applications, it is often either infeasible or undesirable to impose such restrictions on speaker behavior. This paper seeks to address two potential applications of natural number recognition. The first is in applications such as those involving financial transactions. Here the ubiquitous use of natural numbers makes it infeasible to use connected digit recognition. To make such a service usable, the recognizer must have a very high string accuracy for "naturally" spoken dollar amounts. The second involves the recognition of U.S. telephone numbers. In traditional connected digit applications, speech input containing natural numbers is either misrecognized or, at best, rejected by the system. It would be desirable to give users more flexibility in how they speak a digit string, without significantly impairing the accuracy of the system for connected digit recognition.

Section 2 describes the model topology and training procedure while Section 3 describes the databases used for training and evaluation. Experimental results are given in Section 4, followed by conclusions in Section 5.

2. MODEL TOPOLOGY AND TRAINING

The acoustic model proposed in this paper uses an inter-word context dependency modeling paradigm similar to the one which was used by Chou *et al.* in [2] for connected digit recognition. For comparison, a context independent (CI) whole-word model set is also described and tested. While the two model sets have different structures, they share many of the same processing methods.

The input speech signal is sampled at 8 kHz and passed through a first-order pre-emphasis filter with a coefficient of 0.95. A 30 msec Hamming window with a 10 msec shift is applied, followed by a 10th order LPC-derived cepstral analysis. The feature vector consists of 39 features comprised of 12 cepstral coefficients, 12 delta cepstral coefficients, 12 delta-delta cepstral coefficients, normalized log energy, and the delta and delta-delta of the energy.

For the two applications that are considered in this paper, the recognizer vocabulary consists of the following common elements: "zero," "oh," the cardinal numbers "one" through "nineteen," multiples of 10 from "twenty" through "ninety," "hundred" and "thousand." The words "a" and "and" are also allowed in both cases. In addition, the dollar amount application includes the task-specific

words "point" and "dollar," and "cent," along with their plural forms. Two additional words, "a.m." and "p.m.," are included in the vocabulary to allow the use of training data from a time-of-day database (see Section 3). Both tasks use model sets in which the entire lexicon is represented. Task specific language constraints are imposed through the grammar when evaluating the performance of these models for different tasks.

Both the context independent and context dependent models use speaker independent, continuous density left-to-right Hidden Markov Models (HMMs), with varying number of states and mixtures. The models were trained in two steps. Several iterations of maximum likelihood (ML) model building were followed by iterative discriminative minimum classification error (MCE) training as in [3].

The topology of the two model sets is described in greater detail in the following sections.

2.1. Context Independent Models

This model set consists of 40 whole-word units, one for each word in the lexicon. Each word is represented by either a 3 or 10 state Gaussian mixture model containing 8 mixtures per state. In addition, the model set contains a single state, 32 mixture silence model, and two filler models to absorb noise, namely, a 3 state, 8 mixture model representing breath and mouth noises, and a more general 3 state, 32 mixture filler model.

To train the context independent models an initial whole-word segmentation of the training data was used to build a ML bootstrap model. Each ML model set was used to obtain segmentation information for the next ML iteration. After additional iterations of ML model building, the model set was further trained using several iterations of MCE training [3] to arrive at the context independent whole-word models used in the experiments.

2.2. Context Dependent Models

Context dependent subword models are often used to model inter-word dependencies. One such model set, referred to as head-body-tail (HBT) models, has been used effectively in connected digit recognition as in [2]. The same context dependency modeling paradigm is used here to capitalize on the high performance achieved by the HBT connected digit models. The choice of HBT modeling also allows us to use a common model set to combine connected digit and natural number recognition. Head-body-tail models are a special case of subword modeling where the subword units are not phonetic units that make up a word, but rather, represent the beginning, middle, and end of a word. The center of each word, represented by the body model, is a context independent unit. Context dependency information is incorporated in the head and tail models.

The connected digit HBT model set has a lexicon of 11 words, namely the digits "one" through "nine," "zero" and "oh." In terms of HBT models, this translates into 1 body, 12 head, and 12 tail contexts (11 digits and silence) for each digit, yielding a total of 275 subword units. The natural number lexicon described above, contains 31 numbers (including "hundred" and "thousand") and 9 non-number words. To model all the contexts represented by the number models would require 31 bodies and 1984 head/tail contexts, yielding a total of 2015 subword units. In addition to the prohibitive storage and computational cost associated with such a large model set, the training data rarely furnishes sufficient instances of each context to support such exhaustive modeling. In order to reduce the model size, the non-number words in the lexicon and the numbers "hundred" and "thousand" are represented by whole-word models as in the context independent case. To further simplify the model set, the notion of shared contexts is used to represent words containing a common stem. For instance, the words "seven," "seventeen" and "seventy" all share a similar initial stem, and could reasonably expect to share the same head contexts. The same would be true of words ending with "teen" which could all share the same tail contexts. Similarly, words ending with "ty" share tail contexts. The model set is further reduced by using generalized contexts which lump head and tail contexts that are under-represented in the data set.

The final inter-word context dependent model set consists of 29 bodies, 169 heads and 170 tails. The model set also includes 11 whole-word models, two filler models, and a silence model. The whole-word and filler models have a similar structure to the context independent models. The context independent body models are represented with 4 states and 16 mixtures per state, while head and tail models are composed of 3 states with 4 mixtures per state. The whole-word context independent model set described in Section 2.1 was used to bootstrap the context dependent models. Initial bootstrap segmentation for the HBT models was obtained by uniformly dividing the whole-word segments into 10 states to obtain head, body and tail segmentation. As with the context independent model set, the HBT models were trained with several iterations of ML training followed by additional iterations of MCE training.

3. DATABASES

The four databases used for training the models are described below.

- DB1: This database consists of connected digit strings, ranging in length from 1 to 16 digits, with an average string length of 11.8. These strings were spoken strictly as connected digits. This database was collected over the U.S. telephone network under different conditions such as data collection efforts, a live service, and field trials covering many dialectical regions in the U.S. 13,714 strings were used for training.
- DB2: The Macrophone Corpus of American English Telephone Speech was data collected by SRI and is distributed by Linguistic Data Consortium (LDC). The data was collected in 8-bit mu-law digital format over T1 telephone lines. We used two subsets of this database. The first consists of strings of dollar amounts (e.g., fourteen thousand three hundred and ninety seven dollars and twenty one cents). 10,294 strings were used for training. The second consists of people saying the time of day (e.g., eleven fifty three a.m.). 552 strings were used for training.
- DB3: The NYNEX PhoneBook database, also distributed by LDC, consists of data collected from a T1 telephone line in 8-bit mu-law digital format. We used a subset of spontaneously spoken strings of natural numbers and dollar amounts. 2341 strings were used for training.
- DB4: This is a local database consisting of phone numbers spoken over the telephone network as either connected digits or natural numbers. 475 strings were used for training.

For evaluating the performance of our natural number models, we used a testing corpus consisting of strings from some of these databases. For the dollar amount application, we used 1291 strings from DB2 and 276 strings from DB3. For the application of U.S. telephone number recognition, we used 465 strings from DB3 and 2986 strings from DB1. None of the strings in the testing corpus were used for training the models.

4. EXPERIMENTAL RESULTS

We chose various natural number tasks for evaluating the performance of our model sets. These tasks involve recognition of:

- dollar amounts,
- U.S. telephone numbers spoken as connected digits,
- U.S. telephone numbers spoken as either connected digits or natural numbers.

Since these natural number tasks differ considerably in the vocabulary used, task dependent constraints were imposed during evaluation of performance. These constraints were defined using a Backus Naur Form (BNF) grammar compiler [1]. In all tasks, the grammar was defined to allow unlimited sequences of filler models only at the beginning of the sentence.

In the following sections, we compare performance between the baseline context independent whole-word ML trained model set and the context independent whole-word model set obtained from MCE training. In addition, for each of the tasks mentioned above, we also compare performance, in terms of string error rate, between the MCE trained context independent whole-word model set and the MCE trained context dependent HBT model set.

4.1. Dollar Amount Recognition

For the dollar amount task, the grammar allows only the following:

- a natural number dollar amount followed by a natural number cent amount (e.g., "two hundred and fifty nine dollars and seventy three cents," "a thousand and nineteen dollars," "thirty eight cents").
- a connected digits dollar amount followed by a connected digit cent amount (e.g., "two five nine point seven three dollars," "two five nine dollars and seven three cents," "two five nine point seven three").

For the dollar amount task, insertions or deletions of the words "and" and "a," and substitutions between "dollar" and "dollars" and between "cent" and "cents" are not counted as errors since they do not change the meaning of the sentence.

Table 1 compares performance of the baseline ML trained context independent whole-word models with the MCE trained context independent whole-word models for the DB2 testing data. It is observed that the MCE trained models reduce the word error rate by 18.73% and the string error rate is reduced by 17.39%.

Table 1: Performance Comparison of Training Methods for Context Independent Models

Training Method (1291 strings)	Baseline ML Model	MCE Trained	Error Reduction
Word Error Rate	3.15%	2.56%	18.73%
String Error Rate	10.22%	8.44%	17.39%

Table 2 illustrates the performance comparison of string error rates between the MCE trained context independent models and the MCE trained HBT models for the DB2 and DB3 dollar amount tasks.

Table 2: String Error Rate for DB2 and DB3 Dollar Amount Tasks

Database (no. strings)	CI Models String Err.	HBT Models String Err.	Error Reduction
DB2 (1291)	8.44%	7.44%	11.85%
DB3 (276)	8.70%	7.61%	12.53%

Table 2 shows that modeling inter-word context dependencies reduces the string error rate by 11.85% for the DB2 testing data and by 12.54% for the DB3 testing data.

4.2. U.S. Telephone Number Recognition

U.S. telephone numbers are 7, 8, 10, or 11 digits. In the case of 8 and 11 digit numbers, the leading digit is always a 0 or 1. The first digit of a 7 and 10 digit phone number cannot be a 0 or 1. The first 3 digits of a 10 digit phone number comprise the area code while the next 3 digits comprise the exchange. When a telephone number is spoken, the recognizer must recognize the number without any knowledge of the number of digits it contains.

For the U.S. telephone number application, we impose the following constraints on the recognition grammar:

- the leading zero/oh/one is optional,
- the area code is optional and can be spoken as connected digits or, in the case of the special area codes 500, 700, 800, or 900, as "five hundred," "seven hundred," "eight hundred," or "nine hundred,"
- The 3 digit exchange can only be spoken as connected digits,
- The last four numbers can be spoken as connected digits, natural number pairs (e.g.,"twelve oh five," "seventeen fifty three"), a digit from 1 through 9 followed by the word "thousand" (e.g., "nine thousand"), as a number between eleven and ninety nine followed by the word "hundred" (e.g., "twenty five hundred"), or as two digits from 1 through 9 followed by the word "hundred."

The DB3 testing database consists of telephone numbers spoken completely as connected digits, as natural numbers, or as a combination of both. This database was tested with the natural number phone grammar.

The results for string error rate for this database are given in Table 3. It is observed that the HBT context dependent models reduced the string error rate by 20.0% over context independent whole-word models.

Table 3: String Error Rate for DB3 US Telephone Number Task (465 strings)

MCE Trained CI Models String Error	MCE Trained HBT Models String Error	Error Reduction
8.60%	6.88%	20.00%

A subset of DB1 contains telephone numbers which are spoken strictly as connected digit strings. In order to study the effect on connected digit recognition accuracy of allowing users more flexibility in speaking styles, this data was tested with two telephone number grammars. The first grammar was a strictly connected digit telephone grammar. The second grammar was a natural number telephone grammar which allowed telephone numbers to be spoken in a more natural way. This grammar was an expansion of the first grammar to include natural number speech input in addition to strictly connected digit speech input. The results of testing this data with the two grammars are given in Table 4.

Table 4: String Error Rates for DB1
US Telephone Number Task (2986 strings)

Grammar Used in Recognition	CI Models String Err.	HBT Models String Err.	Error Reduction
Connected Digit Phone Grammar	12.12%	9.04%	25.41%
Natural Number Phone Grammar	12.22%	9.21%	24.63%

Table 4 shows that the string error rate for strictly connected digit speech input increased only very slightly when a natural number telephone grammar was used instead of a connected digit telephone grammar. The increase in string accuracy was 0.83% for the CI model set and 1.88% for the HBT model set. This is due to the higher perplexity of the natural number grammar. In addition, it is also observed that modeling inter-word context dependencies improved recognition performance over context independent whole-word models. The string error rate reduced by 25.41% for the connected digit telephone grammar, while the reduction for the natural number telephone grammar was 24.63%.

5. CONCLUSIONS

In this paper, we have addressed two applications of natural number recognition by using minimum classification error trained inter-word context dependent models of the head-body-tail type. The first application allows recognition of spoken dollar amounts in financial transactions. The second is in the recognition of U.S. telephone numbers. Experimental results indicate that string error rate reductions of up to 25.41% can be achieved with these models over more traditional context independent whole-word models. In addition, we have also shown that for speech input which is strictly of connected digit type, the increase in string error rate is negligible when a natural number telephone grammar is used instead of a connected digit telephone grammar. Thus, the methods described in this paper make it possible for a speech recognition system to accept both connected digits and natural numbers with high recognition accuracy. Applications that incorporate these techniques will be more usable and more widely accepted because accurate recognition is maintained while permitting a natural and flexible user interface.

6. ACKNOWLEDGMENTS

The authors acknowledge R. Chengalvarayan and P. Ramesh for helpful discussions on natural number modeling, and C. Mitchell for support of the software used for training.

7. REFERENCES

[1] M. K. Brown, J. G. Wilpon, "A grammar compiler for connected speech recognition," *IEEE Transactions on Signal Processing, Vol. 39, No. 1,* pp. 17-28, January 1991.

[2] W. Chou, C.-H. Lee, B.-H. Juang, "Minimum error rate training of inter-word context dependent acoustic model units in speech recognition," *Proceedings International Conference on Spoken Language Processing,* pp. 439-442, 1994.

[3] W. Chou, B.-H. Juang, C.-H. Lee, "Minimum error rate training based on N-best string models," *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 652-655, 1993.

[4] C. N. Jacobsen, J. G. Wilpon, "Automatic recognition of Danish natural numbers for telephone applications," *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 459-462, 1996.

[5] K. Ma, N. Morgan, "Scaling down: Applying large vocabulary hybrid HMM-PLP methods to telephone recognition of digits and natural numbers, " *Proc. 1995 IEEE Workshop on Neural Networks for Signal Processing,* pp 223-232, 1995.

[6] P. Ramesh, J. G. Wilpon, "Modeling state durations in hidden markov models for automatic speech recognition," *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 381-384, 1992.

[7] C. de la Torre, L. Hernandez-Gomez, F. J. Caminero, C. Martin del Alamo, "Recognition of spontaneously spoken connected numbers in Spanish over the telephone line," *Proceedings EUROSPEECH-95,* pp. 2123-2126, 1995.

[8] C. de la Torre, L. Hernandez-Gomez, F. J. Caminero-Gil, C. Martin del Alamo, "On-line garbage modeling for word and utterance verification in natural numbers recognition, " *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 845-848, 1996.