NATURAL NUMBER RECOGNITION USING DISCRIMINATIVELY TRAINED INTER-WORD CONTEXT DEPENDENT HIDDEN MARKOV MODELS

Malan B. Gandhi

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

ABSTRACT

Many automatic speech recognition telephony applications involve recognition of input containing some type of numbers. Traditionally, this has been achieved by using isolated or connected digit recognizers. However, as speech recognition finds a wider range of applications, it is often infeasible to impose restrictions on speaker behavior. This paper studies two model topologies for natural number recognition which use minimum classification error (MCE) trained inter-word context dependent acoustic models. One model topology uses triphone context units while another is of the head-body-tail (HBT) type. The performance of the models is evaluated on three natural number applications involving recognition of dates, time of day, and dollar amounts. Experimental results show that context dependent models reduce string error rates by as much as 50% over baseline context independent whole-word models. String accuracies of about 93% are obtained on these tasks while at the same time allowing users flexibility in speaking styles.

1. INTRODUCTION

In many telephony applications or services that use automatic speech recognition, the ability to recognize numbers is often a major component. Traditionally, recognition of speech input containing numbers has been achieved with high accuracy using isolated or connected digit recognizers. As speech recognition gains visibility and finds a wider range of applications, it is often not feasible to expect users to provide input in the form of isolated or connected digits. For example, the number "847" is likely to be spoken as the sequence of digits "eight four seven" in a context such as a U.S. telephone number, but in a different context such as a monetary amount, the same number is more likely to be spoken in a natural way as "eight hundred and forty seven." In this paper, the latter case is referred to as a natural number.

Previous work on natural number recognition has been reported in the literature. In [5], the authors proposed the use of syllable-like units as the basic units of recognition, and tested the approach on a database consisting of the months of the year. Results on spotting a "time of day" event in telephone conversations were reported in [7]. The concept of head-body-tail models was used for connected digit recognition in [2], and extended to natural number recognition in [4]. In [10], the authors found an improvement in Castilian Spanish connected number recognition by using techniques such as tied-state modeling, multiple candidates, spectral normalization, gender modeling, and noise spotting. Results on recognition of Danish telephone numbers were reported in [6].

This paper seeks to address three potential applications of natural number recognition, namely, recognition of time of day, date, and monetary amounts. These types of natural number recognition can be used for developing products and services such as schedulers, travel planners, and for banking applications such as bill payments, fund transfers, etc. To make such services usable, the recognizer must have a very high string accuracy for input containing "naturally" spoken numbers.

Two different model topologies for recognition of natural numbers are proposed using inter-word context dependent models. Their performance is evaluated on the natural number recognition applications mentioned above and compared to a baseline context independent whole-word model set.

The organization of this paper is as follows. Section 2 describes the databases and vocabulary used for training and evaluation. The feature extraction process is described in Section 3 while Section 4 describes the proposed model topology and training procedure. Experimental results are given in Section 5, followed by conclusions in Section 6.

2. DATABASES AND VOCABULARY

The five 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. This database was collected over the U.S. telephone network through 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 collected by SRI and distributed by Linguistic Data Consortium (LDC). The data was collected in 8-bit mu-law digital format over T1 telephone lines. A total of 17,373 strings consisting of people saying the date, time of day, and strings of dollar amounts 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. 2,341 strings of spontaneously spoken natural numbers, telephone numbers, and dollar amounts 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.
- DB5: This database consists of natural numbers, date, and time of day strings collected over the U.S. telephone network as part of a data collection effort. 5,562 strings were used for training.

Of the 39,465 strings used for training, 14,083 strings contain dollar amount data, 7,795 strings contain dates, and 2,043 strings contain time of day data. The remaining strings contain digits, phone numbers, and other natural numbers. With this training data, the vocabulary of the recognizer consists of 95 words which are listed in Table 1.

zero, oh, one, ..., nine	AM, PM
ten, eleven, ..., nineteen	january, ..., december
twenty, thirty, ..., ninety	sunday, ..., saturday
first, second, ..., twelfth	noon, midnight
thirteenth, ..., nineteenth	morning, afternoon
twentieth, thirtieth	evening
hundred, thousand	quarter, half
dollar, dollars	past, 'til, of
cent, cents	in, the
and, a, point	next, last, this

Table 1: Natural Number Vocabulary

A testing corpus of 5,966 strings from databases DB2, DB3, and DB5 is used for evaluating the performance of the natural number models. Approximately 13% (784 strings) of the data used for evaluation is noisy data. None of the strings in the testing corpus are used for training the models.

3. FEATURE EXTRACTION

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. Each cepstral feature is processed using the hierarchical signal bias removal method [9] to reduce effects of channel distortion.

4. MODEL TOPOLOGY AND TRAINING

Two model topologies for recognition of natural numbers are proposed in this paper: (i) context dependent triphone (CDT) models, and (ii) context dependent head-body-tail (HBT) models. For comparison purposes, a context independent whole-word (CIWW) model set is also trained and tested.

All three model sets use speaker independent, continuous density left-to-right Hidden Markov Models (HMMs), with varying number of states and mixtures. The models are trained in two steps. The first step involves maximum likelihood (ML) training using the segmental k-means training method [8]. This is followed by iterative discriminative minimum classification error (MCE) training as in [3]. MCE training improves the discriminating ability of the ML training method by considering competing models and aims at minimizing the recognition error rate of the training data. Since the vocabulary of the evaluation data is fixed, this training is a task dependent training. Therefore, word recognition is used during MCE training when generating the N-best (N = 4 in these experiments) competing strings for each utterance.

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

4.1. Context Independent Whole-Word Models

This model set consists of whole-word units, one for each word in the lexicon (vocabulary). Each word is represented by a 10 state Gaussian mixture model containing 8 mixtures per state. The model for the word "point" contains 4 mixtures per state since there aren't enough occurrences of this word in the training data. The models for "AM "and "PM" contain 16 mixtures per state since the additional mixtures are found to give better discrimination between the two models. In addition, the model set contains a single state, 32 mixture silence model.

4.2. Context Dependent Triphone Models

Context dependent subword models are often used to model inter-word dependencies. One such model set is based on the occurrence of triphone contexts in the training data. The triphone contexts are obtained by forced segmentation of the training data using a model set of 41 context independent units, allowing silence to occur between words. A total of 1,073 triphone contexts are obtained from the training data. Only those triphone contexts which had more than 200 occurrences in the training data are modeled. There were 426 such contexts. The remaining contexts are lumped into the set of 41 context independent units. The resulting model set consists of 467 units, modeling intra-word and inter-word context dependencies. Each triphone context is represented by a 3 state model containing 8 mixtures per state. The silence model contains a single state with 32 mixtures.

4.3. Context Dependent Head-Body-Tail Models

Another context dependent subword model set often used to model inter-word dependencies, 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 96 words, including silence. To model all possible HBT contexts would require a large number of models. 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, most non-number words in the lexicon and the numbers "first," "second," "third," "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 days of the week have the common ending "day" and would therefore share a common tail context. The sharing of contexts is done by lexical rules. The model set is further reduced by using generalized contexts which lump head contexts that are underrepresented in the training data set. Similarly, for the tail contexts.

This inter-word context dependent model set consists of 58 bodies, 184 heads and 170 tails. The model set also includes 37 whole-word models and a silence model. 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 models have 10 states with most states containing 8 mixtures per state. The whole-word context independent model set described in Section 4.1 is used to bootstrap the context dependent models. Initial bootstrap segmentation for the HBT models is obtained by uniformly dividing the whole-word segments into 10 states to obtain head, body and tail segmentation.

5. EXPERIMENTAL RESULTS

The performance of the different model sets is evaluated on three natural number recognition tasks, namely, recognition of time of day, dates, and dollar amounts. Performance between the two training methods is also compared. The string error rate metric is used for evaluating performance.

Since these natural number tasks differ considerably in the vocabulary used, task dependent constraints are imposed during evaluation of performance. These constraints are defined using a Backus Naur Form (BNF) grammar compiler [1].

5.1. Dollar Amount Recognition

For the dollar amount task, the grammar recognizes any amount between zero and up to, but not including, a million dollars. The amount can be phrased in one of the following ways:

- 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. Performance of the models on the dollar amount task is shown in Table 2.

Model Set	Baseline ML Trained	MCE Trained	Error Reduction
CIWW	13.54%	8.88%	34.43%
CDT	9.54%	7.05%	26.16%
HBT	10.29%	7.24%	29.65%

Table 2: String Error Rate for Dollar Amount (3605 strings)

It is observed that the MCE trained models reduce the string error rate by up to 34.43% over ML trained models. CDT models show a reduction in string error rate of 2.68% over HBT models and a reduction of 20.62% over the baseline CIWW models.

5.2. Date Recognition

The grammar for date recognition is defined to allow flexibility in the way in which a date is spoken. For example, the date December 2, 1998 can be spoken in some of the following ways:

- December two, nineteen ninety eight,
- The second of December,
- Next Wednesday the second, etc.

For the task of date recognition, substitutions between words such as "fourteen" and "fourteenth" are not counted as errors since they do not change the meaning of the sentence. Table 3 shows the performance of the model sets on the date recognition task.

Model Set	Baseline ML Trained	MCE Trained	Error Reduction
CIWW	7.63%	6.80%	10.91%
CDT	6.93%	6.73%	3.00%
HBT	7.07%	6.31%	10.78%

Table 3: String Error Rate for Date (1442 strings)

MCE training reduces string error rate by up to 10.91% in the CIWW case. The HBT models give a 6.18% reduction in string error rate over the CDT models and a 7.14% reduction over the CIWW models.

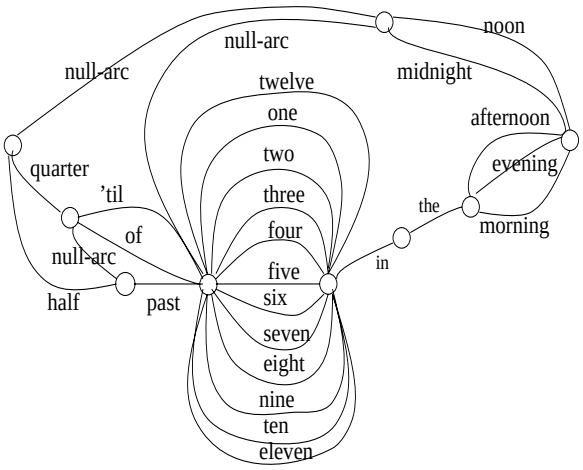

Figure 1: A Subset of Time of Day Grammar

5.3. Time of Day Recognition

The grammar for time of day recognition includes all possible times in the 24 hours of a day. Figure 1 shows a subset of the time of day grammar.

Model Set	Baseline ML Trained	MCE Trained	Error Reduction
CIWW	17.85%	13.93%	21.95%
CDT	8.05%	7.40%	8.10%
HBT	9.14%	6.96%	23.81%

Table 4: String Error Rate for Time of Day (919 strings)

Table 4 shows the performance of the model sets on the time of day recognition task. MCE training reduces string error rate by up to 23.81% in the HBT case. In this task, the HBT models give a 5.88% reduction in string error rate over the CDT models and a 50% reduction over the CIWW models.

The performance on the evaluation data for the three recognition tasks shows that MCE training significantly reduces string error rates for all tasks, and that modeling inter-word context dependencies improves performance over context independent whole word models.

6. CONCLUSIONS

In this paper, two inter-word context dependent model topologies have been proposed for natural number recognition. The applications considered for evaluation of performance are recognition of dates, the time of day, and spoken dollar amounts. Experimental results indicate that string error rate reductions of up to 50% can be achieved by using inter-word context dependent models over baseline context independent whole-word models. MCE training reduces string error rates by up to 34% over ML trained models. The two inter-word context dependent model sets gave similar performance. The methods described in this paper make it possible for a speech recognition system to accept naturally spoken input of dates, time of day, and dollar amounts with high recognition accuracies of about 93%. Since applications that incorporate these techniques allow users flexibility in speaking styles, they are more usable and likely to be more widely accepted.

7. ACKNOWLEDGEMENTS

The author acknowledges John Jacob for help with HBT modeling.

8. 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] M. B. Gandhi, J. Jacob, "Natural number recognition using MCE trained inter-word context dependent acoustic models," *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 457-460, 1998.

[5] Z. Hu, J. Schalkwyk, E. Barnard, R. Cole, "Speech recognition using syllable-like units," *Proceedings International Conference on Spoken Language Processing,* pp. 1117-1120, 1996.

[6] 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.

[7] P. Jeanrenaud, M. Siu, J. R. Rohlicek, M. Meteer, H. Gish, "Spotting events in continuous speech," *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 381-384, 1994.

[8] B.-H. Juang, L. Rabiner, "The segmental K-means algorithm for estimating parameters of hidden markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing,* Vol. 38, No. 9, pp. 1639-1641, 1990.

[9] M. Rahim, B.-H. Juang, W. Chou, E. Buhrke, "Signal conditioning techniques for robust speech recognition," *IEEE Signal Processing Letters,* Vol. 3, No. 4, pp. 107-109, 1996.

[10] 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.