RECOGNIZING CONNECTED DIGITS IN A NATURAL SPOKEN DIALOG

Mazin Rahim
AT&T Labs - Research
180 Park Avenue, Florham Park, NJ 07932.
mazin@research.att.com
www.research.att.com/info/mazin

ABSTRACT

This paper addresses the general problem of connected digit recognition in the telecommunication environment. In particular, we focus on a task of recognizing digits when embedded in a natural spoken dialog. Two different design strategies are investigated: keyword detection or word spotting, and large-vocabulary continuous speech recognition. We will characterize the potential benefits and describe the main components of each design method, including acoustic and language modeling, training and utterance verification. Experimental results on a subset of a database that includes customers responses to the open-ended prompt *"How may I help you?"* are presented.

1. INTRODUCTION

Connected digits play a vital role in many applications of speech recognition over the telephone. Digits are the basis for credit card and account number validation, phone dialing, menu navigation, etc.

Progress in connected digit recognition has been remarkable over the past decade (e.g.,[2, 1]). For databases recorded under carefully-monitored laboratory conditions such as the Texas Instrument database, speech recognizers have been able to achieve less than 0.3% word error rate [2]. Dealing with telephone speech has added a new dimension to this problem. Variations in the spectral characteristics due to different channel conditions, speaker populations, background noise and transducer equipment cause a significant degradation in recognition performance. This degradation, however, can be somewhat minimized through advances in acoustic modeling, discriminative training and robustness, enabling speech recognition systems to operate at about 1% digit error rate [3].

Without doubt, the ultimate objective in digit recognition is to accurately recognize digits embedded in a natural spoken dialog. For example users' response to the prompts *"What number would you like to call?"* or *"May I have your card number please?."* Clearly these types of prompts impose a new set of challenges to the problem of recognizing digits particularly when dealing with *naive* users of the technology. Unlike the more general problems in ASR, such as Switchboard, the difficulty here is *not only* to deal with fluent and unconstrained speech, but being able to accurately recognize an *entire* digit string that may be encoded by digits, natural numbers and/or alphabets. In addition, systems must be able to accommodate for out-of-vocabulary words, hesitation, false-start and other acoustic variations, such as background noise and regional accent.

In this paper, we address the general problem of digit recognition in the telecommunication environment. In particular, we focus on the task of recognizing users' responses when prompted to say their card or phone number. This task is part of an experimental study for customers responding to the open-ended prompt *"How may I help you?"* [5]. We first characterize the problem using a database of 2178 spoken utterances. The objective then is to study two rather different strategies to performing digit recognition; a keyword detection approach versus a large-vocabulary speech recognition approach. We will describe a prototype system and discuss the main components of each approach, including acoustic and language modeling, training and utterance verification. Finally we will report on experimental results comparing the performance of the two systems in terms of accuracy and speed.

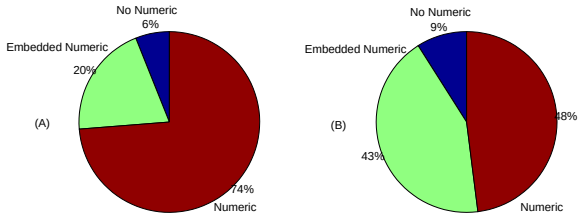

Figure 1. Pie charts of users' responses to (A) card number, and (B) phone number prompts based on spoken numeric.

2. DATABASE ANALYSIS

The experimental database included over 20,000 transactions of which a smaller subset of 2178 utterances represented customers' responses to card and phone number queries [4]. This speaker-independent sub-database, which we will refer to as the *digit database*, has been partitioned into 1552 utterances for training and 626 utterances for testing. Utterances ranged from 1 to 45 words in length, of which 15% of the words were non-digits.

To calibrate the difficulty of this task, we subdivided the digit database based on two sets of results. The first, which is displayed in Figure 1 in the form of pie charts, is a partitioning of the data according to three categories: (a) *numeric*, which includes the digits, natural numbers and alphabets, (b) *embedded numeric*, which includes those numeric that have been spoken among other vocabulary words, and (c) *no numeric*, which includes utterances not containing any numeric keywords. The pie charts indicate that the distribution of users' responses based on spoken numeric is different for the card and phone number prompts. Furthermore, a large proportion of users prefer not to respond with numeric keywords alone. In the case of the phone number prompt, 43% of the utterances contained embedded numeric and 9% included no numeric.

The second result is displayed in Figure 2 which shows a subdivision of the digit database according to ten different categories. This includes digits only (1-9, oh and zero), embedded digits (digits among other vocabulary words),

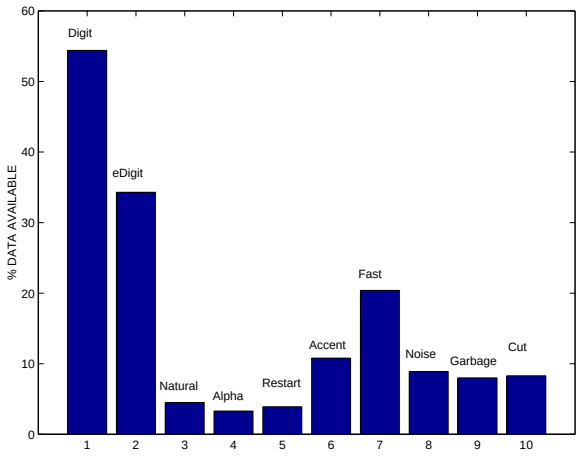

Figure 2. Classification of the digit database as a function of vocabulary and call type.

natural numbers (e.g., hundred, eleven, etc.), alphabets (e.g., A, H, etc.), restarts (false starts, hesitations and corrections), accent (distinct regional dialect and accent), fast speech (>1.5 times faster than average speech rate), noise (severe background speech, music and noise), garbage (extraneous or uninformative speech) and cuts (utterances with incomplete words). The statistics on this database are significantly different than most of the databases that we have previously encountered. Nearly half of the utterances included digits only as opposed to almost 100% for the databases reported in [3]. The new challenge this database presents is the need to accurately recognize embedded digits, natural numbers, alphabets, restarts and extraneous speech which collectively constituted about half of the data. There are also high proportions of fast speech, cuts and severe background noise.

3. SYSTEMS FOR CONNECTED DIGITS

There are two schools of thoughts when designing systems to recognize connected digits. The first is to consider the task as a *detection* problem where the objective is to *spot* digits in a spoken dialogue. This design strategy is similar to that pursued in our previous studies [3, 9]. The second method is to treat the task as a large-vocabulary continuous speech recognition problem where digits would be merely a small subset of the active vocabulary in the lexicon. Issues concerning acoustic and language modeling, training and utterance verification for the two design methods are described next.

3.1. Acoustic Modeling

For the first design method (**WS**), a set of left-to-right continuous density hidden Markov models (HMMs) were applied which captured all possible inter-digit coarticulation [7]. Each digit was modeled by three segments, a head, a body and a tail. A digit had one body and multiple heads and tails depending on the preceding and following context. HMMs consisted of three to four states, each having a mixture of four Gaussian distributions.

For the second design method (**LVASR**), two sets of context-independent phones were used; one was dedicated for the numeric and the other for the remaining vocabulary words. The choice of this design strategy was motivated by the need to maintain real-time operation while providing high accuracy on continuous digit recognition. Each phone set consisted of forty left-to-right continuous density HMMs, with three states each and twenty four Gaussian components per state.

To accommodate for extraneous speech and background noise events, both methods introduced three filler models having different state topologies, with twenty four Gaussian distributions per state. This resulted in 276 units for **WS** and 83 units for **LVASR**.

3.2. Language Modeling

The two design methods for acoustic modeling enable us to exploit two fairly different approaches to language modeling. The first approach, a *deterministic* grammar that is essentially knowledge-driven, was applied in **WS** to constrain the length and characteristic of the digit string. For the phone numbers, this grammar was designed to accommodate for local, long distance and international calls. For the card numbers, the grammar was designed to accommodate for most available credit cards in the market that ranges from ten to sixteen digits in length.

The second approach which was used in **LVASR** is based on a *stochastic finite state* machine. This language model, provided by Riccardi et. al [10], was automatically learned from the training data using a variable N-gram stochastic automaton. This particular design includes back-off mechanism and enables parsing of any arbitrary sequence of words sampled from a given vocabulary.

3.3. Training

Training is carried out in two phases using all the available training corpus, $\mathbf{X}$. Maximum likelihood estimation (MLE) is performed followed by minimum classification error (MCE) training [6]. While in MLE, the objective is to compute a new set of recognition models, $\hat{\Lambda}$, through maximizing a log likelihood function,

$$\hat{\Lambda} = \arg\max_{\Lambda} g(\mathbf{X}, W_0; \Lambda), \tag{1}$$

in MCE training, we aim to compute $\hat{\Lambda}$ by minimizing a smoothed error function:

$$(\hat{\alpha}, \hat{\Lambda}) = \arg\min_{\alpha, \Lambda} \{1 + e^{-\alpha d(\mathbf{X}, \Lambda)}\}^{-1}, \quad \alpha > 0, \tag{2}$$

where $d(\mathbf{X}, \Lambda)$, the misclassification distance, is defined as

$$d(\mathbf{X}, \Lambda) = -g(\mathbf{X}, W_0; \Lambda) + \log\{\frac{1}{N}\sum_{n=1}^{N} e^{g(\mathbf{X}, W_n; \Lambda)}\}. \tag{3}$$

W_0 is the word sequence for the "true" events while W_n are considered as competing hypotheses and can be generated by an N-best search.

For MLE, the training process for **WS** is conceptually similar to that for **LVASR** and differs strictly in the mechanics. In the case of **WS**, non-digit and extraneous speech segments are treated as out-of-vocabulary events for optimizing the filler models. On the other hand, the filler models for **LVASR** are trained using 0.4% of the words in the transcription of W_0 that do not appear in the training dictionary, as well as other extraneous speech events.

For MCE training, the challenge is to optimize the acoustic models while dealing with task-specific language models. Clearly for the case of **WS**, dealing with a deterministic non-weighted grammar simplifies the training process significantly. In fact, training in this framework provides not only improved discrimination among digits, but also between digits and other vocabulary words. For **LVASR**, the use of a stochastic language model in this large vocabulary framework, presents several new challenges. First, training needs to be relatively fast. Second, training should be selective, namely, numeric keywords should benefit the most.

Third, though the objective function in Equation 2 needs to accommodate for the language model, it should also enable the acoustic model parameters to be trained relatively freely of the constraints imposed by the grammar. Unlike the study reported in [8], replacing the objective function with a phone error objective did not yield any improvement in performance.

In our framework, an efficient implementation of MCE along with fast N-best search enabled models to be trained about twice real time. Assigning dedicated sets of units for modeling numeric versus other vocabulary words was crucial in providing selective discriminative training. Furthermore, our framework enabled the acoustic and language models to be tightly integrated, a feature that was essential in improving the overall system performance. Detail information on the extension of MCE training to large vocabulary recognition will be published at a later date.

3.4. Utterance Verification

An important component of a successful spoken dialog system is the ability to identify out-of-vocabulary utterances and utterances that are poorly recognized. This is particularly important for digit recognition since it provides the system with a verification measure of confidence that determines whether or not to automate the call. In [9], we considered the problem of utterance verification for connected digits as a testing statistical hypothesis where the task is to test the *null* hypothesis that a given digit exist in a segment of speech against the *alternative* hypothesis which assumes the digit or digit string does not exist within the speech segment. A dedicated set of verification models was introduced which provided each digit string with a confidence score.

In this study, a confidence score, $\mathcal{CS}(\cdot)$, is computed as a normalized likelihood ratio measure such that

$$\mathcal{CS}(\mathbf{X}, \Lambda) = -d(\mathbf{X}, \Lambda). \quad (4)$$

This formulation is consistent with our discriminative training paradigm since minimizing the misclassification distance implies maximization of $\mathcal{CS}(\cdot)$. In practice, $\mathcal{CS}(\cdot)$ was computed using two best candidates.

4. EXPERIMENTAL RESULTS

The objectives of the experiments presented in this section are two folds:

1. Contrast the two different approaches to digit recognition: word spotting (**WS**) versus large vocabulary recognition (**LVASR**)
2. Characterize the major sources of errors for connected digits within natural spoken dialog.

All experiments have been performed using the AT&T's Watson speech recognition system [11]. For feature extraction, an input signal, sampled at 8 kHz, was pre-emphasized and grouped into frames of 30 msec durations at every interval of 10 msec. Each frame was Hamming windowed, Fourier transformed and then passed through a set of 22 triangular band-pass filters. Twelve mel cepstral coefficients were computed by applying the inverse discrete cosine transform on the log magnitude spectrum. To reduce channel variations while still maintaining real-time performance, each cepstral vector was normalized using cepstral mean subtraction with an operating look-ahead delay of 30 speech frames. To capture temporal information in the signal, each normalized cepstral vector along with its frame *log* energy were augmented with their first and second order time derivatives. The energy coefficient, normalized at the operating look-ahead delay, was also applied for endpointing the speech signal. Recognition was performed through standard Viterbi beam search over a dictionary of 3.6K words and perplexity 14 [5].

Table 1 presents the performance of **WS** and **LVASR** as a function of call type. Two different measurements are reported. The first reflects the performance on digits and the second on the entire vocabulary set. In the case of **WS** only the former measurement is reported since the system was set to recognize digits only. For either measurement, both the word and string error rates are displayed.

	Card Number		Phone Number	
	Digit	**All**	**Digit**	**All**
WS	7.4/42.4	-	14.8/48.8	-
LVASR	5.3/35.7	9.0/43.4	7.9/34.8	15.7/48.1

Table 1. Performance (word/string) of **WS** and **LVASR** for users' responses to card and phone number queries.

The three most striking results in Table 1 are the following: (a) The large-vocabulary approach (i.e., **LVASR**) does significantly better on digit recognition than the detection approach (i.e., **WS**); (b) The performance on users' responses to card number is generally better than that for phone number, a result that can be implied from Figure 1 due to the higher number of utterances with numeric only; (c) The digit error rate is substantially lower than the average word error rate which can be attributed to better acoustic modeling and more data availability for digits.

The performance on digits only for both call types combined are shown in Figure 3. The graphs display the digit and string error rates as a function of processing speed on an SGI R10000 machine. Varying the speed of the decoder has been obtained by changing the operating beam width.

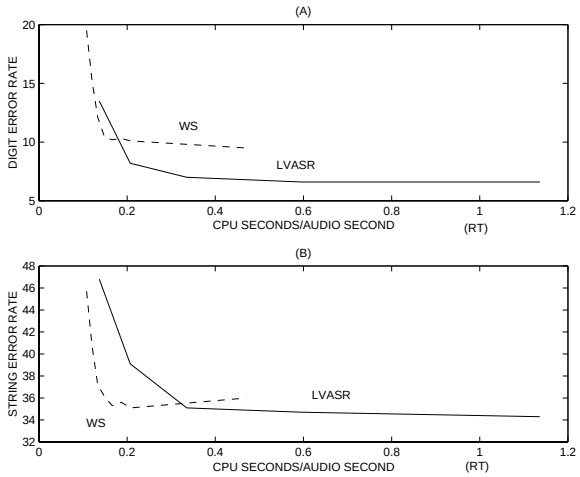

Figure 3. Digit and string error rates versus processing speed per audio second for **WS** and **LVASR**.

For either method, the "knee" points on the curves are well below real-time. One interesting observation, however, is that the large-vocabulary approach is only a factor of two slower than the detection approach.

One possibility for improving system performance is through utterance verification. Figure 4 shows the string error rate as a function of rejection rate for **WS** and **LVASR**. These measurements have been recorded by performing a likelihood ratio test, where each string is assigned a confidence score (see Equation 4), and strings whose scores that exceed a given verification threshold are rejected. Figure 4 has been generated by varying the value of this threshold and tabulating the string error rate following rejection.

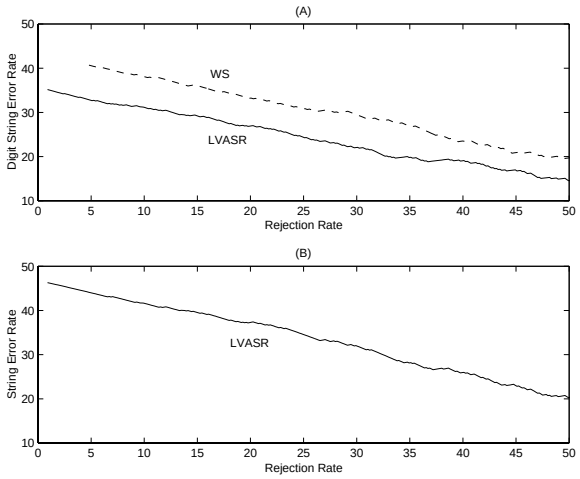

Figure 4. Utterance verification performance for (A) digits only, and (B) entire vocabulary set.

Figure 4(A) shows the verification performance for both methods on the digit vocabulary. This figure demonstrates that utterance verification is able to perform a reasonable job at identifying out-of-vocabulary word strings and incorrectly recognized strings. Figure 4(B) shows the "true" overall performance of **LVASR** upon rejection. From this figure, it is possible to extract possible operating points. For example, at 25% and 50% rejection rates, the string error rate reduces from 48% to 34% and 20%, respectively.

To characterize the sources of errors in both methods, Figure 5 displays the digit error rate based on the classification strategy set forth in Figure 2. Each bar which is associated with a different class of data represents the errors made on digits only. Not surprisingly, **WS** outperforms **LVASR** in both the categories digits only and background noise. The major source of errors on digits is when they are embedded among other vocabulary words, especially natural numbers. For example, the digit error rate for **WS** rises from 2.8% to 14.0% when other vocabulary words are present, and to 30% when natural numbers are present.

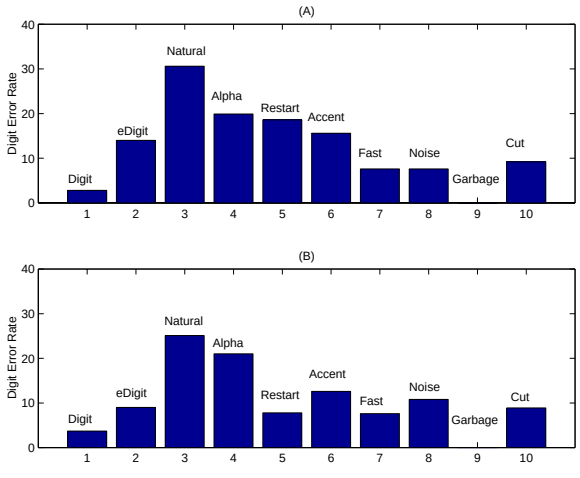

Figure 5. Digit error rate for (A) **WS** and (B) **LVASR** as a function of the vocabulary type and the acoustic characteristics of the speech signal.

5. SUMMARY

A characterization of the problem of recognizing digits embedded in a spoken dialog has been presented in this paper. On a database of users' responses to card and phone number prompts, we found only half of the utterances to include strictly connected digits. Two design methods have been investigated. The first considered the problem as a digit spotting task where as the second treated it as a large-vocabulary speech recognition task. For each method we addressed several aspects of the recognition module including acoustic and language modeling, training and utterance verification. Our results demonstrate that (a) **LVASR** performs better than **WS** for all classes of data except when strictly digits are present, and (b) a large proportion of the errors are attributed to when digits are embedded with other vocabulary words, especially natural numbers and alphabets. For these classes of data, **WS** was inferior to **LVASR** even though it operated twice as fast. At a 25% string rejection rate, the string and digit string error rates when using **LVASR** were 34% and 24%, respectively.

Acknowledgments

The author would like to thank G. Riccardi and E. Bocchieri for their technical contributions, and acknowledges fruitful discussions with L. Saul, A. Ljolje, A. Gorin, C. Lin, R. Rose and the Watson development team.

REFERENCES

[1] E. Buhrke, R. Cardin, Y. Normandin, M. Rahim, and J. Wilpon. Application of vector quantized hidden markov models to the recognition of connected digit strings in the telephone network. In *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1994.

[2] R. Cardin, Y. Normandin, and E. Millien. Inter-word coarticulation modeling and MMIE training for improved connected digit recognition. In *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 243–246, 1993.

[3] W. Chou, M. G. Rahim, and E. Buhrke. Signal conditioned minimum error rate training. In *Proc. European Conf. on Speech Communication and Technology*, pages 495–498, 1995.

[4] A. Gorin and G. Riccardi. Language variation over time and state in natural spoken dialog. *submitted to Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1999.

[5] A. L. Gorin, G. Riccardi, and J. H Wright. How May I Help You? *Speech Communication*, 23:113–127, 1997.

[6] B.-H. Juang and S. Katagiri. Discriminative learning for minimum error classification. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 40:3043–3054, 1992.

[7] C.-H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini, and A. E. Rosenberg. Improved acoustic modeling for large vocabulary continuous speech recognition. *Computer Speech Language*, 6(2):103–127, 1992.

[8] C.-H. Lee, B.-H. Juang, W. Chou, and J.J. Molina-Perez. A study on task-independent subword selection and modeling for speech recognition. In *Proc. ICSLP '96*, pages 1820–1823, 1996.

[9] M. Rahim, C.-H. Lee, and B.-H. Juang. Discriminative utterance verification for connected digits recognition. *IEEE Transactions on Speech and Audio Processing*, 5(3):266–277, May 1997.

[10] G. Riccardi, R. Pieraccini, and E. Bocchieri. Stochastic automata for language modeling. *Computer Speech and Language*, 10:265–293, 1996.

[11] R.D. Sharp, E. Bocchieri, C. Castillo, S. Parthasarathy, C. Rath, M. Riley, and J. Rowland. The Watson speech recognition engine. In *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 4065–4068, 1997.