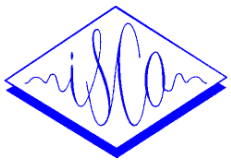

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Recent Technology Developments in Connected Digit Speech Recognition

B. H. Juang and *J. G. Wilpon*
AT&T Bell Laboratories, Murray Hill, New Jersey 07974

Abstract

The use of automatic speech recognition (ASR) over telephone lines has increased tremendously over the past several years.

One of the most important vocabularies for ASR technology today is the *digit* set. The ability to recognize a string of fluently spoken digits is essential to many applications. Over the past year, our experience with connected digit recognition in real-world settings has steadily increased. We have learned that current connected digit recognition algorithms are not robust to – (1) the increased variation in pronunciation and articulation, and (2) the variety of transducer, transmission and background conditions that exist in the telephone network. Hence, if improvements in recognition performance are to be achieved these issues have to explicitly considered.

In this paper, we report results of several directions of research aiming at increasing the robustness and performance of connected digit recognition over the telephone network.

1 Introduction

For many applications of speech recognition over the telephone, connected digit input is essential. Such applications include credit card and account number validation, catalog ordering and voice dialing by spoken digits. Until recently, technology developments pertaining to connected digit recognition have been focused on a particular database recorded by Texas Instruments (TI) over a decade ago under controlled, quiet studio conditions [5]. Various techniques have been reported to be able to achieve better than 99.7% word accuracy [2]–[3] on this particular database. However, when applied to more realistic databases recorded by AT&T over the telephone network, these techniques were found to perform no where near the previously reported accuracy. An obvious conclusion is that the techniques developed for the TI database fail to address several issues that become important under realistic conditions. These issues include, most notably, increased variation in pronunciation and articulation, unknown acoustic ambients and channel distortion. These issues have to be explicitly considered in the development of new techniques to improve the recognition accuracy.

Increase in variation in pronunciation and articulation often can be observed in databases collected over a widespread geographical area. The data set so recorded would be accordingly less homogeneous. In the context of statistical pattern recognition, a less homogeneous data set poses a need in more carefully choosing and estimating the prescribed probability model. We therefore pursued discriminative training and a model architecture which allows efficient incorporation of context dependency.

Another direction in the technology developments is the method of signal conditioning which aims at minimizing or equalizing the effects of ambient noise and the unknown telephone channels. We focus our attention on linear models of such ambient and channel conditions. In particular, we developed an iterative method which by way of maximum likelihood estimates the extraneous additive component (bias) in a given spectral representation. The estimated bias component is then removed from the spectral representation in order to eliminate the mismatch between the unknown data and the models which may have been trained in a different condition based on a different data set.

In this paper, we report on these main directions of technology developments and the performance improvements that these new techniques were shown to be able to offer in connected digit recognition applications.

2 Database

Four databases recorded over the telephone network were used in our investigation. These are: 1) *88-Mall*, 2) *91-Mall*, 3) *Teletravel*, and 4) *UCS* databases. Only the 88-Mall database was employed in the algorithm development while others were used in various evaluations. This would allow us to examine the cross-database robustness of a new algorithm. It also provides a test scenario for signal conditioning and normalization methods some of which are designed to equalize the signal discrepancies, if any, between different databases.

Among the various attributes of a speech database, particularly pertinent in our study are talker and dialectical (i.e. speech production) variations and network and channel (i.e. speech transmission) variations.

3 Discriminative Training

A speech recognizer stores the characteristics (models) of individual vocabulary classes for pattern matching. These characteristics are learned from a set of known examples via a training procedure. Traditionally, the characteristics manifest in the form of probability distribution based on the classical Bayesian formulation and the training procedure amounts to distribution estimation. Maximum likelihood and its variants have long been the preferred methods.

However, the optimality of these methods in pattern recognition is conditioned on the correct choice of the distribution form for the data. For speech signals, no known distribution has been verified to be the correct distribution. Thus the distribution estimation based approach using the maximum likelihood criterion is always suboptimal in terms of speech recognition.

Recently, a new approach called discriminative training was introduced [4] in that the goal of model training is set to directly minimize the error rate of the recognizer. The approach can work with any choice of discriminant functions, including hidden Markov models, and has been shown to produce substantially lower speech recognition error rate in various isolated word experiments. We therefore pursued a particular extension of this technique, called minimum string error rate training, of this approach for connected digit recognition applications.

3.1 Minimum String Error Rate Training

Consider a sequence of speech observation vectors $\mathbf{O}$ and a sequence of words $\mathbf{w} = (w_1, w_2, \ldots)$. We use w_i to denote the word identity and the word model parameter set interchangeably without ambiguity. The string (word sequence) model which best matches the observation sequence is determined by a dynamic programming algorithm (Viterbi algorithm, level building algorithm, etc.) according to

$$\overline{\mathbf{w}} = \arg\max_{\mathbf{w}} [\log f(\mathbf{O}, \overline{\mathbf{q}}_{\mathbf{w}} \mid \mathbf{w})] \quad (1)$$

where $f(\ )$ is the hidden Markov model probability measure parametrized by Λ which is omitted without ambiguity and

$$\overline{\mathbf{q}}_{\mathbf{w}} = \arg\max_{\mathbf{q}_{\mathbf{w}}} [\log f(\mathbf{O}, \mathbf{q}_{\mathbf{w}} \mid \mathbf{w})] \quad (2)$$

is the optimal state sequence associated with $\mathbf{w}$.

In segmental minimum string error rate training, the modeling is based on the string model corresponding to the (known) correct word sequence and the string models corresponding to the N most competitive word sequences according to the recognizer (which is to be refined in the training process). The N most competitive word sequences are obtained by an N-best search algorithm [7]. Following the minimum error rate training formulation, we define a misclassification measure, with Λ denoting the recognizer parameter set,

$$d(\mathbf{O};\Lambda) = -\log f(\mathbf{O}, \overline{\mathbf{q}}_{\mathbf{w}_0} \mid \mathbf{w}_0) + \left\{ \frac{1}{N-1} \sum_{\mathbf{w} \neq \mathbf{w}_0} \log f(\mathbf{O}, \overline{\mathbf{q}}_{\mathbf{w}} \mid \mathbf{w})^{\eta} \right\}^{1/\eta} \quad (3)$$

where η is a positive number and $\mathbf{w}_0$ is the correct (known during training) word sequence. The misclassification measure is then embedded in a smoothed 0-1 function $\ell(\ )$ to emulate the recognizer performance in terms of the error rate,

$$\ell(\mathbf{O}, \Lambda) = \ell(d(\mathbf{O}, \Lambda)) = [1 + e^{-\gamma d(\mathbf{O};\Lambda)}]^{-1} \quad (4)$$

where γ is a positive number specifying the smoothness of the function.

Parameters of the HMMs representing the unit models are then adjusted with a generalized probabilistic descent algorithm [1] to minimize the loss function, using the segmentation of the given speech utterance based on these models, until the process converges to a local minimum.

3.2 Evaluation

The minimum string error (MSTE) training method was applied to the training set of the 5-site 88-Mall data using 12 context independent 10-state whole word models. The number of mixture components in each state was set at 32. The whole word models were used to test various databases under the normal known-length (KL) and unknown-length (UL) decoding scenarios. Table 1 shows a comparison of error rate obtained from using ML and MSTE/GPD trained models for the four databases respectively. Error rate reduction in the range of 15% to 50% was achieved by the MSTE/GPD method in unknown-length decoding and 15% to 25% in known-length decoding.

	ML		MSTE/GPD	
Test Database	**KL**	**UL**	**KL**	**UL**
88-Mall	2.3	2.5	N/A	1.3
91-Mall	2.1	4.4	1.5	2.4
Teletravel	2.9	4.2	2.1	2.9
UCS	5.4	7.1	4.7	5.9

Table 1: Digit error rate comparison between ML and MSTE/GPD models for various databases.

4 Model Architecture

When words are spoken in a connected fashion, the acoustic variability increases tremendously at the word junctions, resulting in poor modeling and unsatisfactory recognition performances. One way to remedy the problem is to increase the "resolution" in modeling by enlarging the model size in terms of numbers of states and mixtures, particularly for the "silence" or "acoustic ambient" models. (The silence model is the most eminent to benefit from increased model resolution without the immediate difficulty of the sparse data problem due to the fact that it normally has much fewer states.) This model architecture revision is straightforward and has been shown to bring about performance improvements.

Another remedy can come from context dependent acoustic modeling which takes into account explicit variations in a word spoken in a certain context.

Context dependency may be introduced in the form of whole-word (digit) models (CDWW) as well as subword (sub-digit) models (CDSW). We focus on the latter in connected digit experiments.

4.1 Context Dependent Subword Models

We consider and propose hybrid context dependency that combines the strength of a phone or subword based method

and that of a word based method seems appropriate. The hybrid context dependent modeling method defines context in words but acoustically models the word by subword units. Since the beginning and ending parts of a spoken digit are more likely to be affected by the neighboring digits or silence (context), it is reasonable to use interword subword units (conditioned on the context) particularly for the beginning and ending parts. Thus, each digit j is divided into three parts, namely the *head* unit, h_j, the *body* unit, b_j and the *tail* unit, t_j. The body unit represents the more stable part of a digit which is less affected by the inter-digit coarticulation effect. The head part is modeled according to the word context, i.e. the digit or silence that precedes it. The head model for digit j is therefore specified as $t_i h_j b_j$ if it follows digit i or $\#h_j b_j$ if it follows a pause or silence. Similarly, the tail models can be represented according to the digit or silence that follows. Figure 1 illustrates the network that comprises a digit (j) based on concatenation of these context dependent subword models. For our tasks, there are $(12 + 1 + 12) \times 11 = 275$ such subword models. The head-

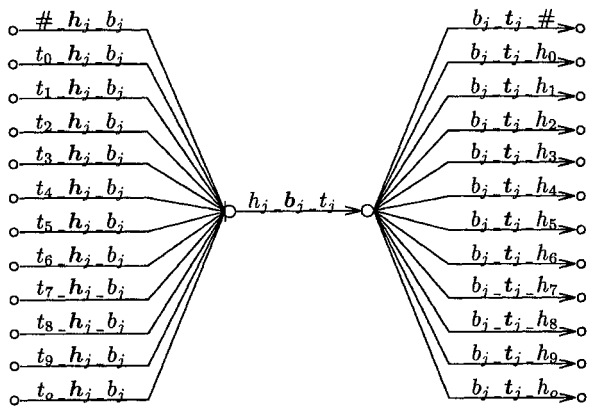

Figure 1: A subword representation of word j based on context dependent head-body-tail units.

body-tail decomposition was not applied to the silence background or pause, modeling of which remained unchanged from the traditional arrangement. The total number of subword models is thus 276. These subword models, in general, have much fewer states than a whole word model and hence do not necessarily increase the system complexity.

4.2 Evaluation

The hybrid technique was evaluated on the 88-Mall database using 5-site, 24623 endpointed utterances as the training set and a separate 5-site 11909 endpointed utterances as the test set. The results are tabulated in Table 2 together with the original baseline performance of a whole-word based context independent system and a system using 56 CDWW models. Incorporation of context dependency is observed to lead to about 25% reduction in errors.

Further evaluation of the hybrid context dependent modeling technique was conducted on a variety of databases. Table 3 summarizes the recognition performance in terms of

Unit Type	# of Mixture	Word Error Rate
Whole Word	64	2.5
276 CDSW	16	1.1

Table 2: Comparison of whole word modeling and subword modeling techniques in connected digit recognition.

the digit error rate, in comparison with the original context independent modeling results. The conditions for evaluation were much richer than the pilot study that produced Table 2. The hybrid technique was shown to be able to achieve 40-50% reduction in digit error rate.

	CIWW		CDSW	
Database	KL	UL	KL	UL
88-Mall	n/a	2.5	n/a	1.1
91-Mall	2.1	4.4	1.2	2.4
Teletravel	2.9	4.2	1.3	1.9
UCS	5.4	7.1	3.9	4.7

Table 3: Digit error rate achieved by a context independent whole word system and a context dependent subword system in known-length (KL) and unknown-length (UL) decoding for various databases.

5 Signal Bias Removal

In a simplified signal bias model, if we denote the speech feature vector sequence by $X = (x_1, x_2, \ldots, x_T)$, the observed feature vector sequence $Y = (y_1, y_2, \ldots, y_T)$ is assumed to contain a constant additive term b,

$$y_t = x_t + b \ . \tag{5}$$

The formulation of the bias removal method is based on maximizing the likelihood of a speech model in which the bias is considered as the unknown parameter.

With an additive bias, b, we can write

$$p(Y|b) = \prod_t \max_i \ p(y_t - b \mid \lambda_i) \tag{6}$$

as the likelihood function for the unknown bias b. The maximum likelihood estimate of b, denoted by b_{ML}, is

$$b_{\mathrm{ML}} = \arg\max_b \prod_{t=1}^{T} \max_i \ p(y_t - b \mid \lambda_i) \tag{7}$$

which results in

$$b_{\mathrm{ML}} = \overline{b} = \frac{1}{T} \sum_{t=1}^{T} (y_t - z_t) \ , \tag{8}$$

where

$$z_t = \mu_j = \arg\max_i \ p(y_t - b \mid \lambda_i) \ . \tag{9}$$

A special case of the bias removal method is "cepstral mean subtraction". For many real time applications, the above bias removal algorithm can be implemented in a sequential manner.

Database	Baseline				SBR				Sequential SBR			
	Ins	Del	Sub	Wd-err	Ins	Del	Sub	Wd-err	Ins	Del	Sub	Wd-err
88-Mall	0.23	0.42	1.29	1.94	0.21	0.26	1.15	1.62	0.24	0.32	1.33	1.89
91-Mall	0.36	1.57	1.80	3.73	0.59	0.56	0.97	2.12	0.48	0.76	1.10	2.35

Table 4: Percentage digit error rates of insertion, deletion and substitution for the baseline system with and without SBR.

5.1 Evaluation

The signal bias removal (SBR) method was implemented in a discrete density HMM recognizer [6] for evaluation, although the method can be equally applied to other HMMs. A 5-site 88-Mall data set, 14629 strings for training and 7073 strings for testing, was used in the evaluation. Also used for cross database evaluation (test only) was a 2-site 91-Mall data set of 2842 strings of 10, 14, 15 and 16 digits.

Table 4 compares the recognition performances in terms of the digit error rate of a baseline (VQ based) system and a system that incorporates SBR. The digit error rate is broken down into separate error categories, namely insertion, deletion and substitution. For the same database but in an open test, SBR reduced the word error rate from 1.9% to 1.6%. In cross database tests (i.e. 88-Mall models applied to 91-Mall data), the baseline digit error rate without SBR nearly doubled from 1.9% to 3.7%. With SBR, the digit error rate was reduced back to 2.1%, only slightly higher than that in the same database test. The last main column shows the digit error rate reduction from the baseline performance when the sequential SBR method was applied in all test conditions.

6 Summary

We have presented the results of several directions of technology development aimed at improving the accuracy of connected digit recognition. In the area of new training methods, we showed that minimum string error discriminative training was able to provide a 30-40% reduction in error rate compared to the traditional maximum likelihood method, without altering the model architecture. A versatile hybrid subword modeling technique was demonstrated to be very effective in allowing context dependent modeling for performance improvements, bringing about a reduction of more than 50% in error rate in some cases. We have also devised a new signal bias removal technique which achieved 16% reduction in digit error rate in a matched training condition and 43% reduction in mismatched training (cross database) conditions. A flexible N-best decoding (search) scheme that permits use of inter-word context dependent models was also implemented and shown to provide great potential for further reduction in error rate. In an experiment, it was shown that the digit error rate could be reduced from 1.1% to 0.2% if the top 4 candidates were properly considered in the final decision. We further outlined in this paper several extensions of these techniques for completeness as well as potential, additional improvements in accuracy. These new techniques will be properly integrated in the final system for best efficiency and performance.

Acknowledgements

We would like to attribute the work reported here to the individuals who studied, proposed or investigated the techniques: Wu Chou for discriminative training, Chin Lee for subword-based modeling and Mazin Rahim for the signal bias removal methods. Separate in-depth technical reports covering individual techniques are available.

References

[1] W. Chou, B. H. Juang, and C. H. Lee. Segmental GPD training of an HMM based speech recognizer. In *Proc. ICASSP '92*, pages I–473–476, March 1992.

[2] W. Chou, C. H. Lee, and B. H. Juang. Minimum error rate training based on N-best string models. In *Proc. ICASSP '93*, pages II–652–655, April 1993.

[3] R. Haeb-Umbach, D. Geller, and H. Ney. Improvements in connected digit recognition using linear discriminant analysis and mixture densities. In *Proc. ICASSP '93*, pages II–239–244, April 1993.

[4] B. H. Juang and S. Katagiri. Discriminative learning for minimum error classification. *IEEE Trans. Signal Processing*, SP-40(12):3043–3054, Dec. 1992.

[5] R. G. Leonard. A database for speech independent digit recognition. In *Proc. ICASSP '84*, pages 42.11.1–4, March 1984.

[6] M. Rahim and B. H. Juang. Signal bias removal for robust speech recognition in adverse environments. In *Proc. ICASSP '94*, Adelaide, Australia, April 1994.

[7] F. K. Soong and E. F. Huang. A tree-trellis based fast search for finding the N-best sentence hypotheses in continuous speech recognition. In *Proc. ICASSP '91*, pages I–705–708, May 1991.