SPEECH RECOGNITION USING AUTOMATICALLY DERIVED ACOUSTIC BASEFORMS

R. C. Rose[1] *and E. Lleida*[2]

[1]AT&T Labs – Research, Murray Hill NJ
[2]University of Zaragoza, Spain

ABSTRACT

This paper investigates procedures for obtaining user-configurable speech recognition vocabularies. These procedures use example utterances of vocabulary words to perform unsupervised automatic acoustic baseform determination in terms of a set of speaker independent subword acoustic units. Several procedures, differing both in the definition of subword acoustic model context and in the phonotactic constraints used in decoding have been investigated. The tendency of input utterances to contain out-of-vocabulary or non-speech information is accounted for using likelihood ratio based utterance verification procedures. Comparisons of different definitions of the likelihood ratio used for utterance verification and of different criteria for estimating parameters used in the likelihood ratio test have been performed. The performance of these techniques has been evaluated on utterances taken from a trial of a voice label recognition service.

1 INTRODUCTION

There has been considerable interest in telecommunications based speech recognition services that provide user configurable vocabularies. Name dialing systems are a good example. These systems provide personalized voice controlled repertory dialers that can be easily configured by individual users. This paper describes a set of techniques that were investigated for voice label recognition over the public switched telephone network and an experimental study evaluating the performance of these techniques over a large population of users.

In previous implementations of name dialing systems, relatively simple unsupervised automatic acoustic baseform determination procedures were used for obtaining speaker dependent word pronunciations. These automatically derived phonetic baseforms are often quite different from baseforms that might be obtained from a pronunciation dictionary. However, speaker dependent recognition performance has been found to be very similar regardless of whether the baseforms are automatically derived or obtained from a prespecified lexicon [3, 8]. While one should not extrapolate these results beyond limited vocabulary speaker dependent voice label recognition tasks, these results are very encouraging when considering any task involving automatic acquisition of user-specific lexicons. They also encourage investigation of how linguistic constraints, acoustic–phonetic constraints, model robustness, and speaker dependence should be integrated into the process of identifying the optimum choice of the acoustic baseform. As a first step in evaluating these issues, a study was performed to investigate different linguistic constraints and different definitions of HMM subword acoustic model context for automatic acoustic baseform determination. This study is described in Section 3.

Utterance verification techniques are very important for user–configurable vocabularies because out–of–vocabulary input from users is especially common. To deal with these "unexpected" events, several techniques based on likelihood ratio based hypothesis testing criteria are applied to verifying word hypotheses produced by the speech recognizer and also to modifying the optimization criterion used in the decoder. Techniques and experiments relating to utterance verification in the name dialing service are discussed in Section 4.

2 NAME DIALING TASK

2.1 Description of Service

The vocabulary independent speech recognition techniques described in this paper are motivated by the development of a system which allows a user to associate voice labels with frequently dialed telephone numbers. Once the service has been accessed, the user can place a call by simply speaking the voice label that has been associated with the desired number. For the system evaluated in this study, a user adds new entries to the voice label inventory by speaking three utterances of the word during an enrollment procedure.

The name dialing service is based on a speaker independent subword acoustic model based HMM speech recognizer. Each vocabulary word is represented as a sequence of subword acoustic units, or phonetic transcription, that must be derived automatically from the enrollment utterance. Separate phonetic transcriptions are derived from each enrollment utterance and these are encorporated into the speaker specific lexicon.

2.2 Speech Corpora

The speech corpus used to conduct the experimental study described in Section 4 is composed of utterances collected from actual users of a trial version of the name dialing service described in Section 2.1. A 56 speaker subset of the total population of speakers participating in the field trial was chosen for the evaluation corpus. This subset was chosen primarily based upon their frequency of use of the service in order to provide a sufficient number of utterances per speaker. The total number of test utterances per speaker ranged from 70 to 200 utterances.

The vocabularies that were trained by each speaker ranged in size from 7 to 36 voice aliases, with an average over the entire population of 15 aliases per speaker. In order to evaluate the performance of techniques for verifying the occurrence of keywords in unconstrained utterances, it is

necessary to have a relatively large number of non–keyword utterances. An analysis of the distribution of calls to the name dialing service showed that over 15% of the calls received did not contain valid vocabulary words. However, since there was a shortage of out–of–vocabulary (OOV) utterances available for evaluating the performance of utterance verification techniques, an artificial scenario was constructed where a five word subset of each speaker's total vocabulary was chosen as the speaker's active vocabulary. The utterances corresponding to the remaining vocabulary words were then used to represent OOV speech. The words in the active vocabulary are referred to below as the set of in–vocabulary (INV) words. The total test set contained 3594 utterances.

3 ACOUSTIC MODELING

In most speech recognition applications it is assumed that a pronunciation dictionary exists for all words in the vocabulary or that some lexical representation of the vocabulary words exists from which a phonetic pronunciation can be derived. This implies that, for every lexical item that will be input to the speech recognizer, there is a set of rules which describes how that lexical item will be expanded according to the inventory of subword units. These rules might be defined explicitly by linguists as entries in a lexicon, or defined as probabilistic grapheme–to–phoneme rules that are derived from text corpora. Networks of pronunciations for lexical items may also be defined through prespecified phonological rules applied to a baseform pronunciation. However, we are interested in those applications where no prior information concerning the pronunciation of the vocabulary words is available. Our first problem is to determine a set of baseform pronunciations from unlabeled enrollment utterances.

3.1 Unsupervised Automatic Baseform Determination

We investigated the performance of several unsupervised maximum likelihood decoding procedures for extracting acoustic baseforms from enrollment utterances. The performance of unigram, bigram, and trigram phonotactic grammars trained from a six million word text corpus taken from the Associated Press News Wire were compared on this task.

For each voice label, the user spoke a set of M enrollment utterances which were analyzed into sequences of observation vectors $Y_1, \ldots, Y_M$. The jth observation vector Y_j is given by the T_j length sequence $Y_j = \vec{y}_{j,1}, \ldots, \vec{y}_{j,T_j}$, where $\vec{y}_{j,t}$ is the 39 component observation vector obtained at time t of the jth utterance. The enrollment procedure produces a set of M phonetic baseforms $R^1, \ldots, R^M$. The jth phonetic baseform R^j is given by the N_j length sequence $R_j = r_{j,1}, \ldots, r_{j,N_j}$. The phonemes $r_{j,n} \in \mathcal{P}$ that were produced as part of the phonetic baseform for class j were taken from the inventory of subword acoustic units. Several different sets of subword units have been defined, and are described below. The optimum phonetic baseform is obtained by

$$R_j = \arg\max_{R \in \mathcal{P}} P(Y_j \mid R) P(R) \,. \qquad (1)$$

The probability, $P(R)$, in Equation 1 represents the prior phone sequence probability. A comparison of statistical unigram, bigram, and trigram models was made for representing this probability. These language models were trained without back-off from six million words of text [6].

3.2 Definition of Subword Context

A second acoustic modeling problem that must be addressed in configuring a speaker dependent lexicon is the problem of training the initial set of HMM subword acoustic models. Different sets of subword acoustic units were investigated in terms of the definition of the context that each set of units represents. We compared the performance of context independent monophone models as well as context dependent tri–phone models. In all cases, the acoustic models were trained from a corpus of 12146 phrases collected over the public telephone network in the United States. The speaker population consisted of 2004 speakers, and there was a total of 4439 unique words in the training set.

Context dependent HMM models were trained using the forward–backward algorithm from the above "task–independent" corpus using a simple back–off procedure. Three state tri–phone models containing four Gaussian mixtures per state were trained for tri–phone contexts whose occurrence count in the training data exceeded a threshold N_t. Di–phone models were trained for di–phone contexts with occurrence counts greater than N_d. General context models were trained to cover contexts that occurred with frequency less than N_d.

The decoding network inferred by Equation 1 took the form of a single finite state network. The network was generated and interfaced with the maximum likelihood decoder [1] using a newly designed set of utilities for operating on finite state machines [5]. Descriptions of the statistical phonotactic grammar and the subword context constraints were compiled into separate finite state networks. These networks were composed to form the complete finite state network used for decoding the optimum phonetic baseform.

3.3 Baseline Performance

Baseline speech recognition performance was measured for several different configurations of a name dialing system and displayed in Table 1. Forty–three subword acoustic models were trained using three state left–to–right HMM's from the corpus described above. Table 1 describes each system in terms of the number of mixtures per state, the size of the speaker dependent speech recognition vocabulary, and the procedure used for obtaining phonetic baseforms for the vocabulary words. Since each of the 56 speakers in the trial created a separate recognition vocabulary, speech recognition performance was measured individually using a separate lexicon for each speaker and then averaged.

The first row of Table 1 gives recognition performance when the full vocabulary for each speaker is active during recognition and phonetic baseforms were obtained "automatically" from an average of three enrollment utterances per word. A level of 96.2% correct speech recognition performance was obtained. This fell only slightly to 96.0% when the number of mixtures was reduced to sixteen. In Section 4, several techniques are investigated for verifying the presence of a keyword within an utterance by defining a reduced vocabulary of five words per speaker, and considering the remaining words as "out–of–vocabulary." The third row of Table 1 shows that error–rate decreased over 60% when using the smaller vocabulary. A discussion of the word verification results will be given in Section 4. Finally, the last row of Table 1 describes the performance of a system which obtains phonetic baseforms for vocabulary words using the pronunciation engine from the Bell Labs text–to–speech system. A single phonetic expansion was obtained for each word. Since many vocabulary items were proper names, it was necessary to hand correct many of the pronunciations produced by the text–speech–system. It

was surprising to note that the error rate actually increased nearly 30% using the text–to–speech pronunciations. This is consistent with the findings of a similar study [3].

Word Accuracy			
System Configuration			
Mixtures per State	Vocabulary Size (words)	Enrollment Procedure	Percent Correct
64	15 (ave.)	automatic	96.2
16	15 (ave.)	automatic	96.0
64	5	automatic	98.4
64	5	tts	97.2

Table 1: Speech recognition performance for name dialing system under a variety of conditions.

3.4 Constrained Decoding

An additional set of experiments was performed to evaluate the importance of additional linguistic constraints and improved acoustic modeling on word accuracy for voice label recognition. Table 2 describes the word accuracy obtained when statistical phonotactic grammars and context dependent acoustic models were used for automatic acoustic baseform identification. The first column of Table 2 displays performance for context independent (CI) models and the second column displays performance for context dependent (CD) acoustic models. All of the results in Table 2 were obtained for speaker dependent vocabulary sizes averaging fifteen total words per speaker.

There are two important observations that can be made from these results. The first is that higher level phonotactic constraints in the form of statistical bigram and trigram grammars applied during the enrollment procedure have very little effect on word accuracy when CI models are used. The improvement obtained using phonotactic constraints is somewhat more pronounced when CD models are used. The second observation is that better modeling of acoustic context using tri–phone models trained from a task independent corpus has a significant effect on word accuracy. Context dependent acoustic models are shown to reduce the error rate by 32 percent over context independent models when a trigram grammar is used during the unsupervised enrollment procedure.

Word Accuracy		
Phonotactic Constraints	Subword Context	
	CI	CD
Unigram	96.0	96.9
Bigram	96.3	97.4
Trigram	96.3	97.5

Table 2: Voice label recognition word accuracy for a range of phonotactic constraints and for both context independent (CI) and context dependent (CD) subword acoustic units.

4 WORD HYPOTHESIS VERIFICATION

To deal with the problem of non–keyword input utterances, word hypothesis testing procedures have been developed for the purpose of verifying the presence of a vocabulary word in an utterance. This section describes several techniques that have been applied to utterance verification (UV) for the name dialing problem. Utterance verification procedures are investigated which are based on a likelihood ratio (LR) based hypothesis testing criterion.

A likelihood ratio score

$$S(Y, \lambda^C, \lambda^I) = \log P(Y \mid \lambda^C) - \log P(Y \mid \lambda^I) \qquad (2)$$

is computed in order to test the hypothesis that utterance Y was generated by the most likely model, λ^C, obtained during speech recognition versus Y having been generated by an alternate hypothesis model λ^I. Two issues relating to LR based UV are investigated for verifying the existence of vocabulary words in the name dialing application. The first is the effect of different parameterizations for the alternate hypothesis model in the LR test of Equation 2. The second issue relates to the criterion used to estimate the parameters of the models λ^C and λ^I when used for UV.

4.1 Definition of Alternate Hypothesis Model

Several very simple parameterizations for the alternate hypothesis model were investigated. The best trade–off between performance and computational complexity was obtained using a frame based likelihood ratio test. The alternate hypothesis probability was given in terms of the observation probabilities $b_i(y_t) = p(\vec{y}_t \mid s_t = i)$ at time t for state $s_t = i$, $i \in \mathcal{S}_t$

$$\log P(Y \mid \lambda^I) = \sum_{t=1}^{T} \log \sum_{i \in S_t} b_i(y_t). \qquad (3)$$

In Equation 3, $\mathcal{S}_t$ is the set of states used for forming the alternate hypothesis model probability. Several definitions of this set were evaluated. The best performing of these was the simplest. For each observation frame y_t, the probabilities $b_{s_t}(y_t)$ for all active states are computed, and the M most likely states are used to form the set $\mathcal{S}_t$ in Equation 3. This is very similar to utterance verification and word spotting procedures proposed elsewhere [2, 9]

Figure 1 shows a comparison of utterance verification performance using three different definitions of the alternate model probability. These include the "state–net" given by Equation 3, a "phone–net" consisting of an unconstrained network of phonemes obtained according to Equation 1, and a "word–net" consisting of alternate word candidates obtained during recognition. The performance of the different utterance verification procedures is described using a set of receiver operating characteristic (ROC) curves. The performance was measured separately for each speaker in the 56 speaker corpus described in Section 2.2 and each word in that speaker's reduced five word vocabulary. A separate speaker dependent, word dependent threshold was applied to the likelihood ratio scores for each word. Each curve represents an average of individual speaker dependent, word dependent ROC curves. It is clear from Figure 1 that the phone-net and state–net alternate hypothesis models achieved the best performance. The plot on the left in Figure 1 shows that, in both cases, over 80% of the out–of–vocabulary utterances were rejected at an operating point where only 5% of the within–vocabulary utterances were rejected. The plot on the right shows that nearly 99% of in–vocabulary words were correctly detected by the two methods at the same operating point.

The performance shown for the state-net alternate model was obtained using the twenty most likely states in the network as the set $\mathcal{L}_t$ of states in Equation 3. This alternate model performs only slightly worse than the phone-net. This is very important because the state-net alternate model requires very little additional complexity during decoding. This is because the probabilities $b_i(y_t)$ that are used

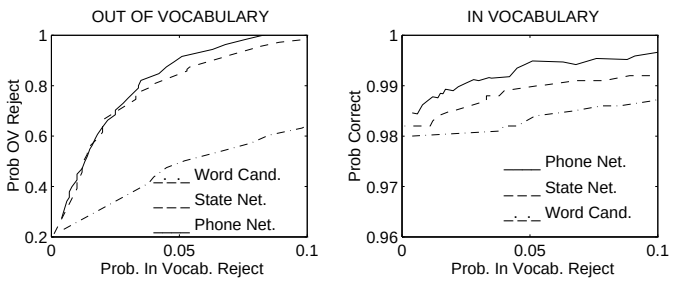

Figure 1: Performance comparison using different representations for the alternate hypothesis model. *a)* **ROC's for OOV word rejection vs. the INV rejection** *b)* **ROC's for INV word detection vs. the INV rejection.**

UV Equal Error Rate Performance		
Alternate Hypothesis Model λ^I	Decision Thresholds	
	Word Dep.	Word Indep.
State-Net	5.1	14.9
ML HMM	4.5	14.6
LR Trained HMM	2.8	11.2

Table 3: EER performance for three UV models.

in Equation 3 will have already been computed as part of the Viterbi search.

4.2 LR Based Training

A parameter estimation procedure for estimating both null hypothesis and alternate hypothesis model parameters for UV according to a likelihood ratio criterion was proposed in [7, 4]. The goal of the training procedure is to obtain model parameters λ^C and λ^I which increase the log LR, $S(Y, \lambda^C, \lambda^I)$, in Equation 2 for correctly hypothesized keywords and decrease $S(Y, \lambda^C, \lambda^I)$ for false alarms. This is accomplished by applying an iterative discriminative training algorithm. As in [4], the alternate hypothesis model, λ^I, includes a 3 state HMM with 8 mixtures per state trained for each subword HMM. The reader is referred to [4] for a more detailed discussion of the phone based alternate hypothesis models.

A single set of alternate hypothesis models was trained for the entire population of 56 speakers using the utterances collected during enrollment. The iterative training procedure was initialized from a single three state maximum likelihood (ML) trained HMM. The UV performance is summarized in Table 3 as the equal error rate (EER), the error probability obtained when the probability of false word acceptance and false word rejection are equal. Separate EER figures are given for word dependent and word independent decision thresholds.

The first row of Table 3 displays the UV performance for the state–net alternative hypothesis model whose receiver operating characteristic curves are shown in Figure 1. The second row represents the EER performance of a single three state HMM model obtained from maximum likelihood training on the enrollment utterances. Finally, the last row of Table 3 gives the word verification performance after the discriminative LR training procedure had been performed on the enrollment utterances. It is clear from the table that the LR training procedure had a significant effect on word verification performance for the case of word dependent and word independent decision thresholds.

5 SUMMARY AND CONCLUSIONS

A set of techniques and experiments for automatic acoustic baseform determination and utterance verification in a name dialing task have been described. These techniques were evaluated on utterances that were collected during an actual trial of a name dialing service over the telephone network. The effect of higher level phonotactic constraints and improved acoustic modeling for obtaining speaker dependent acoustic baseforms from unlabeled enrollment utterances were evaluated. The application of these techniques resulted in a reduction in error rate of 37%. Several different likelihood ratio based hypothesis testing procedures were evaluated for word verification. The procedures differed both in the manner in which the alternate hypothesis model was defined and in the criterion used for estimating model parameters in training. A likelihood ratio based training procedure was found to improve UV performance by over 25%.

6 ACKNOWLEDGEMENTS

The authors wish to express their appreciation to Enrico Bocchieri for assistance in integrating HMM models and ML decoder, Giuseppe Riccardi for assistance with Ngram modeling tools, and Mike Riley for his help with tools for operating on finite state machines.

REFERENCES

[1] E. Bocchieri, G. Riccardi, and J. Anantharaman. The 1994 at&t atis chronus recognizer. *Proc. Spoken Language Syst. Tech. Workshop*, pages 265–268, January 1995.

[2] J. M. Boite, H. Bourlard, B. D'hoore, and M. Haesen. A new approach to keyword spotting. *Proc. European Conf. on Speech Communications*, September 1993.

[3] R. Haeb-Umbach, P. Beyerlein, and E. Thelen. Automatic transcription of unknown words in a speech recognition system. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pages 840–843, April 1995.

[4] E. Lleida and R. C. Rose. Efficient decoding and training procedures for utterance verification in continuous speech recognition. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, May 1996.

[5] M. Mohri, F. Pereira, and M. Riley. Weighted automata in text and speech processing. *ECAI Workshop*, pages 46–50, 1996.

[6] G. Riccardi, E. Bocchieri, and R. Pieraccini. Nondeterministic stochastic language models for speech recognition. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pages 247–250, April 1995.

[7] R. C. Rose, B. H. Juang, and C. H. Lee. A training procedure for verifying string hypotheses in continuous speech recognition. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pages 281–284, April 1995.

[8] R. C. Rose, E. Lleida, G. W. Erhart, and R. V. Grubbe. A user–configurable system for voice label recognition. *Proc. Int. Conf. on Spoken Lang. Processing*, October 1996.

[9] R. C. Rose and D. B. Paul. A hidden Markov model based keyword recognition system. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, April 1990.