A USER-CONFIGURABLE SYSTEM FOR VOICE LABEL RECOGNITION

R. C. Rose[1], *E. Lleida*[2], *G. W. Erhart*[3], *and R. V. Grubbe*[3]

[1]AT&T Research, Murray Hill NJ
[2]University of Zaragoza, Spain
[3]Lucent Technologies, Columbus, OH

ABSTRACT

A set of techniques for configuring a speech recognition system to a particular user are described in the context of voice label recognition over the public switched telephone network. User-configurable vocabularies are provided through automatic acoustic baseform determination based on an inventory of speaker independent subword acoustic units. The tendency of input utterances to contain out-of-vocabulary or non-speech information is accounted for using likelihood ratio based utterance verification procedures. Mismatch between a given user's utterances and the HMM model is accounted for using a frequency warping approach to speaker normalization. The performance of these techniques was evaluated on utterances taken from a trial version 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 each individual user. 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.

An experimental study was carried out in the context of a trial name dialing service. This service is briefly described in Section 2. An evaluation speech corpus corresponding to a subset of the total utterances collected from actual users of the service is also described. The acoustic modeling performed as part of the subword acoustic model based HMM speech recognizer used for name dialing is described in Section 3. Baseline speech recognition performance is presented for a variety of different modeling scenarios. 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. Finally, in Section 5, a set of speaker normalization techniques intended to reduce the acoustic mismatch between the speaker independent acoustic models and new speakers' utterances are evaluated on the name dialing task.

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 a phonetic transcription, that must be derived automatically from the enrollment utterance. Separate phonetic transcriptions are derived from each enrollment utterance and 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 demonstrates that a large percentage of the calls received did not contain valid vocabulary words. Of a total of 12150 calls to the service, less than 85% of the calls contained words that had been entered into the users' vocabularies as part of the enrollment procedure. For most speakers, however, there was still a shortage of out–of–vocabulary (OOV) utterances available for evaluating the performance of utterance verification techniques. In order to increase the number of OOV speech utterances, 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 words in the active vocabulary are referred to below as the set of in–vocabulary (INV) words. The utterances corresponding to the remaining vocabulary words were then used to represent OOV speech. The total test set contained 3594 utterances.

3 ACOUSTIC MODELING

It is generally 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. Name dialing, in the context of the application described here, has no such prior information as to the pronunciation of the vocabulary words. Hence, there are two acoustic modeling problems that must be addressed in configuring a name dialing system. The first is the problem of training the initial set of subword acoustic units. The second problem is obtaining the phonetic transcription of the vocabulary words. This section will describe the analysis and modeling procedures used to deal with both of these problems.

The subword acoustic models were trained from a subset of a telephone based speech recognition corpus collected over the public switched telephone network [9]. The subset of the corpus consisted of 12146 general phrases read by 2004 speakers recruited from the population of AT&T employees in the continental United States. A set of 43 general context phoneme models, $\mathcal{P}$, were trained using the segmental k–means algorithm from this database.

Unless explicitly stated otherwise, all simulations in this paper were performed using twelve linear prediction derived cepstrum coefficients along with the first and second difference coefficients resulting in a 39 component observation vector. No channel or noise compensation strategies were employed, nor was there any processing applied that requiring multiple passes over the utterance.

For each voice label, the user speaks a set of M enrollment utterances which are 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 are produced as part of the phonetic baseform for class j are taken from the set of 43 vocabulary independent phones described above. The optimum phonetic baseform is obtained by

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

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 [10]. 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].

Speech Recognition Performance			
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.

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

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

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 first was an unconstrained network of phonemes, referred to below as the "phone–net". The most likely phone sequence, $\hat{R}$, was obtained from the phone–net for the test utterance according to Equation 1 in Section 3. The alternate model probability is given as

$$\log P(Y \mid \lambda^I) = P(Y \mid \hat{R}) \; , \tag{3}$$

The second parameterization considered is based on a frame based likelihood ratio test. The alternate hypothesis probability is given in terms of the observation probabilities $b_i(y_t) = p(\vec{y}_t \mid s_t = i)$ at time t for state $s_t = i$, $i \in \mathcal{S}_t$

$$\log P(Y \mid \lambda^I) = \sum_{t=1}^{T} \log \sum_{i \in \mathcal{S}_t} b_i(y_t). \tag{4}$$

In Equation 4, $\mathcal{S}_t$ 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 4. This is very similar to utterance verification and word spotting procedures proposed elsewhere [2, 8]

Finally, the last parameterization considered for the alternate hypothesis model in Equation 2 was simply the sum of the log likelihoods of competing vocabulary word candidates

$$\log P(Y \mid \lambda^I) = \sum_{i=2}^{M} \log P(Y \mid W_i) \; . \qquad (5)$$

In Equation 5, $W_2, \ldots, W_M$ are the alternate word candidates sorted by their likelihood produced during recognition.

Figure 1 shows a comparison of utterance verification performance using three different definitions of the alternate model probability. These include the "phone-net" given by Equation 3, the "state-net" in Equation 4, and the "word-net" in Equation 5. 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.

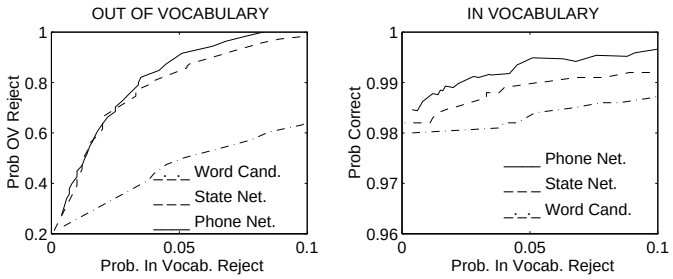

Figure 1: Performance comparison using different representations for the alternate hypothesis model. *a)* **ROC's for OOV word rejection vs. the INV rejection** *b)* **ROC's for INV word detection vs. the INV rejection.**

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 4. 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 in Equation 4 will have already been computed as part of the Viterbi search. The "word-candidate" alternate model corresponds to the combined word candidate likelihoods shown in Equation 5 with $M = 5$. It is quite surprising to note that the utterance verification performance for the word-candidate shown in Figure 1 is far worse than the other less constrained alternate models. The reason for this poor performance is the lexically constrained word-net does not adequately "cover" the acoustic space meant be represented by the alternate model.

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 2: EER performance for three UV models.

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, 5]. Similar training procedures have been developed and applied to a connected recognition task [6]. 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 [5], 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 [5] 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 2 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 2 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 2 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 SPEAKER NORMALIZATION

The name dialing task as described in Section 2 relies on the use of speaker independent subword acoustic hidden Markov models. While the enrollment procedure described in the previous section produces speaker dependent phonetic transcriptions, there may still be significant acoustic mismatch between the speaker independent HMM's and a new speaker's utterances. A procedure was investigated for reducing the effects of this mismatch without significantly increasing the amount of computation required during recognition or the amount of memory that must be dedicated to the storage of speaker dependent parameters. The procedure corresponds to a frequency warping approach to speaker normalization. The goal of this approach is to warp the frequency scale of speech utterances in order to maximize the likelihood of the utterance with respect to the HMM model. The use of a maximum likelihood criterion for selecting a frequency warping function in continuous speech

recognition was originally proposed in [1]. The frequency warping techniques discussed here are based on the work described in [4].

Li and Rose linearly expanded and compressed the frequency scale for an utterance by adjusting the positions and the shapes of the filters in a mel–scale filter bank implemented as part of mel–frequency cepstrum (MFC) analysis [4]. During recognition, an optimum linear warping factor was chosen which maximized the likelihood of the test utterance with respect to the speaker independent HMM's. During training, the frequency scale of the utterances from each training speaker were warped so that the resulting speaker independent HMM was defined over a normalized feature set. When applied to speaker normalization in a telephone based connected digit recognition task, frequency warping reduced the error rate by approximately twenty percent.

For the name dialing task, this speaker normalization paradigm was modified to minimize the added computational complexity during recognition. Frequency warping was performed as part of the training procedure as described above. However, instead of reestimating a new warping factor for each test utterance as described in [4], the optimal warping factor for a given speaker was estimated from the enrollment utterances for that speaker. A single warping factor is then stored as a speaker dependent parameter for reducing model mismatch.

The procedure for estimating a linear warping factor $\hat{\alpha}_i$ for speaker i can be summarized as follows. The warping factor for each speaker is estimated from all enrollment utterances for each voice label in the speaker's inventory of labels. Let $\mathbf{Y}_i = \{Y_{i,1}, \ldots, Y_{i,L}\}$ be the set of all enrollment utterances for speaker i. Also, assume that $\mathbf{R}_i = \{R_{i,1}, \ldots, R_{i,L}\}$ is the set of all phonetic baseforms or transcriptions for speaker i determined according to the maximum likelihood procedure described in Section 3. Then the optimum warping factor for speaker i is obtained by maximizing the likelihood of the warped utterances, $\mathbf{Y}_i^{\alpha}$ with respect to the HMM model set λ and the given transcription set $\mathbf{R}_i$:

$$\hat{\alpha}_i = \arg\max_{\alpha} P(\mathbf{Y}_i^{\alpha} \mid \lambda, \mathbf{R}_i) \ . \tag{6}$$

Assuming that the observation sequences, $Y_{i,j}$, are independent, the probabilities in Equation 6 can be updated as new enrollment utterances become available without having to access previous enrollment utterances.

The performance of this speaker normalization paradigm is described in Figure 2. The plots in the figure show ROC's obtained using MFC analysis which was computed using warped and unwarped filter bank coefficients. In comparing Figure 2 to the "state–net" curves in Figure 1, it is clear that the filter–bank and linear prediction based feature analysis obtain similar performance. It is also clear from Figure 2 that the frequency warped MFC analysis shows significant improvement over the unwarped MFC case.

6 SUMMARY AND CONCLUSIONS

A set of techniques and experiments for improving the performance of voice label detection 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. Speaker independent acoustic baseforms were created for each vocabulary word in terms of a set of vocabulary independent phoneme models. These baseforms were created through an enrollment procedure where the user of the service spoke an average of three utterances per voice label. 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

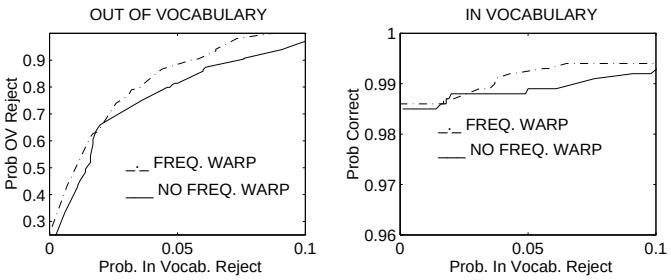

Figure 2: Word detection performance using speaker normalization based on speaker dependent frequency warping MFC and un–warped MFC. *a)* **ROC's for OOV word rejection vs. the INV rejection** *b)* **ROC's for INV word detection vs. the INV rejection.**

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%. The use of speaker normalization procedures based on frequency warping for reducing model mismatch in recognition were investigated. It was found that speaker normalization resulted in a 13% increase in OOV rejection performance at a given INV rejection rate with no extra computational complexity during testing.

REFERENCES

[1] A. Andreou, T. Kamm, and J. Cohen. Experiments in vocal tract normalization. *Proceedings of the CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[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] L. Lee and R. C. Rose. Speaker normalization using efficient frequency warping procedures. *To appear Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, May 1996.

[5] 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.

[6] M. Rahim, C. Lee, B. Juang, and W. Chou. Discriminative utterance verification using minimum string verification error training. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, May 1996.

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

[9] R. M. Sachs, M. Tikijian, and E. M. Roskos. 1993 U.S. english SUBWORD speech data set. *AT&T Bell Laboratories Technical Memorandum*, (BLO41221C-940504-01TM), 1994.

[10] Richard Sproat and Joseph Olive. Text to speech synthesis. *AT&T Technical Journal*, 74(2):35–44, 1995.