INTEGRATION OF UTTERANCE VERIFICATION WITH STATISTICAL LANGUAGE MODELING AND SPOKEN LANGUAGE UNDERSTANDING

R. C. Rose, H. Yao, G. Riccardi, and J. Wright
AT&T Labs - Research, 180 Park Ave., Florham Park, NJ 07932
(rose,s_yao,dsp3,jwright)@research.att.com

ABSTRACT

Methods for utterance verification (UV) and their integration into statistical language modeling and spoken language understanding formalisms for a large vocabulary spoken understanding system are presented. The paper consists of three parts. First, a set of acoustic likelihood ratio based utterance verification techniques are described and applied to the problem of rejecting portions of a hypothesized word string that may have been incorrectly decoded by a large vocabulary continuous speech recognizer. Second, a procedure for integrating the acoustic level confidence measures with the statistical language model is described. Finally, the effect of integrating acoustic level confidence into the spoken language understanding unit (SLU) in a call-type classification task is discussed. These techniques were evaluated on utterances collected from a highly unconstrained call routing task performed over the telephone network. They have been evaluated in terms of their ability to classify utterances into a set of fifteen semantic actions corresponding to call-types that are accepted by the application.

1 INTRODUCTION

In this paper, utterance verification techniques are applied to an automated call routing task [1, 6]. The distinguishing aspect of this task is that it attempts to derive a small number of semantic actions from utterances spoken by users who may have little or no knowledge of the limitations of the system. It is often the case that the utterances that are presented to the system have no relevance at all to the domain in question, contain words or phrases that are not legitimate vocabulary words, or were not correctly recognized by the ASR component of the system. The call routing task and the characteristics of the utterances derived from the task are briefly described in Section 2.

It is often the case that human-machine interfaces are configured so that a large percentage of the input utterances are ill-formed. This is the case for user-initiated human-machine dialog [11, 3], automation of telecommunications services [9], and is certainly true in case for machine interpretation of human human dialog [2, 8]. Utterance verification in this context implies the ability to detect legitimate vocabulary words in an utterance that is assumed to contain words or phrases which are not explicitly modeled in the speech recognizer. However, even when input utterances tend to be well-formed and contain relatively few out-of-vocabulary (OOV) words, UV techniques can be applied to determine when decoded word hypotheses are correct. These procedures have been shown to improve performance in a number of applications where OOV utterances are relatively rare including telephone based connected digit and command word recognition [5].

A set of acoustic likelihood ratio (LR) based confidence measures for UV are defined in Section 3, and preliminary UV results for these measures on the call routing task are described. These measures are similar to a set of techniques that were developed and applied to a "movie locator" dialog task [3]. Each hypothesized word or phrase obtained from the ASR decoder is associated with a confidence measure and passed along to the natural language back end to weight decisions in classifying utterances according to call-type.

A mechanism by which acoustic and linguistic information can be combined through incorporating the notion of acoustic confidence in a stochastic automaton (SA) is discussed in Section 4. There are a number of examples of confidence measures that incorporate both acoustic and language level scores [4]. The approach that is considered here attempts to extend the notion of a stochastic automaton which is currently used to describe an N-gram language model for speech recognition [7]. In the simplest case, a state in a SA may correspond to a word context for some word w_i, and the weight on an arc extending from the state would correspond to the probability of producing w_i given the previous state. There are a number of ways in which acoustic confidence could be incorporated into this framework. In Section 4, we investigate a method where the definition of a state in the language model can be expanded to include not only the word context but also a discrete, coded representation of the acoustic confidence obtained for the word history. By including an additional state variable corresponding to acoustic confidence we thereby expand the state space of the associated SA.

Classification of spoken utterances into a small number of semantic categories by the SLU involves searching through a lattice of grammar fragments that have been extracted from the input speech. Section 5 describes how word level acoustic confidence scores are used in the process of obtaining the a posteriori probabilities that are associated with these grammar fragments.

2 AUTOMATED CALL ROUTING TASK

The utterances used for the experimental study described in this paper were taken from spoken transactions between customers and human telephone operators over the public switched telephone network. The utterances correspond to customers responses to the open-ended prompt "How may I help you?" [6]. The first utterance from the customer in this transaction was transcribed and labeled according to one of fifteen call-types. The call-types themselves correspond to a set of actions relating to the routing of the incoming call. Examples of these call-types include *collect*, *calling card*, and *third party billing*, with an additional "*other*" type to handle calls that do not correspond to those that have been defined. A set of 2243 utterances were used for training sub-

word acoustic acoustic hidden Markov models (HMM), and a set of 1000 utterances were used as a test set. The utterances were an average of 5.3 seconds in duration, with the longest utterance being 52.7 seconds. There is an average of nineteen words per utterance with an out-of-vocabulary rate in the test data at the utterance level of thirty percent. The best ASR performance on this test set using context dependent acoustic models and a variable phrase N-gram stochastic automaton (VNSA) language model is relatively low (approximately 55 - 60 percent word accuracy) [6]. Furthermore, the base line performance on this test set for a real-time system that was used in these experiments is approximately 45 percent word accuracy.

3 ACOUSTIC MEASURES FOR UTTERANCE VERIFICATION

This section presents a likelihood ratio (LR) based procedure for generating word level acoustic confidence measures. First, the UV problem is presented in a hypothesis testing framework. Second, the form of the densities used in the LR based hypothesis test for UV is described and the methods used for training the model parameters associated with these densities is presented. Finally, a simple non-parametric approach is presented for converting these LR scores to a posteriori probabilities for use in the SLU system.

It is assumed that the input to the speech recognizer is a sequence of feature vectors $Y = \{\vec{y}_1, \vec{y}_2, ..., \vec{y}_T\}$ representing a speech utterance containing both within-vocabulary and out-of-vocabulary words. The within-vocabulary words will be referred to here as belonging to the class of "target" hypotheses and the out-of-vocabulary words will be referred to as "imposters" or belonging to the class of alternate hypotheses. Incorrectly decoded vocabulary words appearing as substitutions or insertions in the output string from the recognizer will also be referred to as belonging to the class of alternate hypotheses. It is also assumed that the output of the recognizer is a single *word* string hypothesis $\mathcal{W} = w_1, ... w_K$ of length K. Of course, all the discussion in this section can be easily generalized to the problems of verifying one of multiple complete or partial string hypotheses produced as part of an N-best list or word lattice as well.

Under the assumptions of the Neyman-Pearson hypothesis testing framework, both the target hypothesis and alternate hypothesis densities are assumed to be known. In the context of UV, it will be assumed that the target or correct hypothesis model λ^c and the alternate model λ^a corresponding to a hypothesized vocabulary word are both hidden Markov models. A LR based hypothesis test can then be defined under the assumptions that the null hypothesis, $\mathcal{H}_0$, corresponds to Y being generated by the target model λ^c, and alternative hypothesis, $\mathcal{H}_1$, corresponds to Y being generated by the alternative model λ^a.

$$\log P(Y|\lambda^c) - \log P(Y|\lambda^a) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \tau \qquad (1)$$

where τ is a decision threshold. Given the target hypothesis probability $P(Y|\lambda^c)$ which models correctly decoded hypotheses for a given recognition unit and the alternate hypothesis probability $P(Y|\lambda^a)$ which models the incorrectly decoded hypotheses, Equation 1 describes a test which accepts or rejects the hypothesis that the observation sequence Y corresponds to a legitimate vocabulary word by comparing the LR to a threshold.

The UV score for a given word, w_k, is obtained by combining the LR scores for the acoustic subword units, $u_{k,i}, i = j, \ldots, N_k$, that make up that word. The log of the likelihood ratio given in Equation 1 is computed for each subword unit using null hypothesis model and alternate hypothesis model probabilities that are defined below. For all of the simulations described in this paper, the set of 53 acoustic subword HMMs that were originally trained for speech recognition using the forward-backward algorithm were used for the null hypothesis model, λ^c. The alternate hypothesis probability for a given subword unit, u_j, is actually formed as the linear combination of two different models

$$P(Y|\lambda^a(j)) = \alpha_{bg} P(Y \mid \lambda^a_{bg}) + \alpha_{im} P(Y \mid \lambda^a_{im}(j)) \qquad (2)$$

where α_{bg} and $\alpha_{im} = 1 - \alpha_{bg}$ are linear weights.

The purpose of $\lambda^a_{im}(j)$, referred to here as the imposter alternate hypothesis model for subword unit u_j, is to provide a description of the speech segments that are frequently decoded incorrectly as u_j. The purpose of λ^a_{bg}, referred to here as the background alternative model, is to provide a broad representation of the feature space. This broad representation serves to reduce the dynamic range and to reduce the influence of outliers on the value of the likelihood ratio. A single one state, 64 mixture background alternate hypothesis model is shared amongst all "target" HMM models. The subword level LR scores are combined using a sigmoid weighting to form word level scores so as to mitigate the effects of the large dynamic range that is typical of any likelihood ratio based measure.

The subword unit dependent models, $\lambda^a_{im}(j)$, are three state HMM models. Maximum likelihood training of these models is performed in two steps. First, speech recognition is performed on a set of development utterances, and subword units corresponding to insertions and deletions are labeled in the output stream. Second, forward-backward training of the set of imposter models is performed by updating the conditional expectations for decoded units that were labeled as false alarms in the hypothesized strings. As a result, each subword model $\lambda^a_{im}(j)$ is trained to represent the events that are frequently confused with subword unit u_j.

In previous work on other tasks, both null hypothesis models and alternative hypothesis models were trained using a LR criterion which is similar to the LR used in UV. This modified training criterion resulted in significant improvement in utterance verification performance [3, 8]. Efforts are currently under way to train models according to this new training procedure on this task.

Figure 1 displays an example of where the occurrence of the hypothesized phrase "calling card" is verified in the recognition output for the 1000 utterance test corpus in the call routing task. This example was chosen because of the high semantic association of the phrase with the "*calling card*" call-type in the call routing task and its high frequency of occurrence in the recognized strings for the test set [1]. It is clear from the figure that the likelihood ratio based confidence measure provides reasonably good detection characteristics.

4 INTEGRATION INTO LANGUAGE MODEL

Using localized measures of acoustic confidence by themselves can be misleading when the effects of linguistic context are significant, as is true in the case of large vocabulary speech recognition. Stochastic language models for speech recognition are usually trained from text transcriptions and thus assume that the speech recognition is error-free. The goal here is to exploit acoustic confidence measures derived from the actual speech utterance to account for an imperfect decoder.

Our approach to integrating acoustic level confidence with the language model is to augment the Ngram word histories, which currently define linguistic context, with encoded values of acoustic confidence. A statistical language model

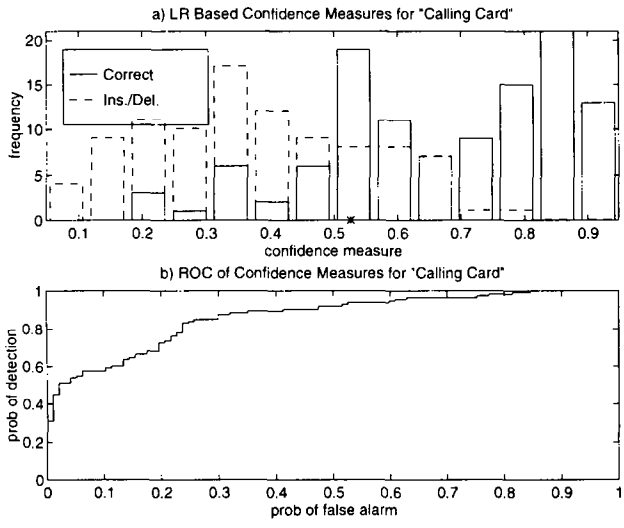

Figure 1: *a)* **Histogram of LR based confidence scores obtained for correctly and incorrectly decoded occurrences of the phrase "calling card" in the 1000 utterance test corpus (210 total occurrences).** *b)* **A receiver operating curve plotted over the confidence measures obtained for "calling card".**

is generally defined over the elements of a K length word sequence, $w_1, \ldots, w_K$, for an utterance where $w_i \in V$, and V is the lexicon for the task. The word string can be replaced by a symbol-pair sequence an utterance is represented by $(w_1, c_1), (w_2, c_2), \ldots, (w_K, c_K)$, where, $c_i \in [0, \ldots, Q-1]$, is a discrete, Q level encoding of the acoustic confidence for word w_i. Hence, the linguistic context for word w_i in a third order statistical Ngram language model would be augmented from $\{w_{i-1}, w_{i-2}\}$ to $\{(w_{i-1}, c_{i-1}), (w_{i-2}, c_{i-2})\}$.

The obvious advantage of this scheme is to reduce the probability that an inserted or substituted word u in the recognition output will result in additional errors. A very high observed co-occurrence of this word with another word v in the text training corpus may result in the word bigram probability $P(w_i = v|w_{i-1} = u)$ being very high. However, the word context could also be conditioned on, for example, a binary random variable representing acoustic confidence. As a result, $P(w_i = v|w_{i-1} = u, c_{i-1} = 0)$, corresponding to the case when there is low acoustic confidence at word $w_{i-1} = u$ might be much lower than $P(w_i = v|w_{i-1} = u, c_{i-1} = 1)$ corresponding to high acoustic confidence.

Of course, these probabilities must be estimated from a limited corpus of acoustic training utterances, which is generally over an order of magnitude smaller than the text corpus for training language models. With this small amount of data for training, the issue of dealing with the robustness of these acoustic confidence conditioned (ACC) probability estimates becomes critical. Our approach in the paper is to deal with this issue in a manner similar to that used in estimating language model probabilities. When an Ngram context occurs infrequently or not at all with a given acoustic confidence level in the acoustic training data, one of many possible back-off mechanisms may be invoked [7].

The automatic learning of finite state automata that incorporate ACC probabilities fits very nicely under the frame-work of the VNSA [7]. As described above, the notion of a state in the VNSA can be expanded to include encoded acoustic confidence measures along with word history. The notion of backing off to null states need not correspond strictly to proceeding from higher order to lower order Ngram contexts, but can also be invoked to deal with lack of statistical robustness in the estimation of ACC probabilities.

The following procedure has been investigated for training a stochastic language model that incorporates acoustic confidence:

1. Estimate word level UV scores for words in training-development data sets (4317 utterances).
2. Quantize UV scores into Q levels ($Q = 2$).
3. Estimate ACC word counts from data.
4. Learn VNSA state transition function and probabilities from word and quantized UV score sequences [7].
5. Prune states in VNSA network [7].

Using this algorithm, the stochastic finite state machine can be learned from two independent information sources: the lexical word sequence and the sequence of quantized acoustic scores. The stochastic transducer is designed by associating with each speech input utterance a sequence of word/symbol pairs (w_i, c_i). We have incorporated this class of stochastic transducers in the *WATSON* AT&T large vocabulary recognizer and tested on the 1000 utterance test set. The output produced by the recognizer is a hypothesized string of word/symbol pairs, providing an indication of the confidence associated with each word. An excerpt from this experiment is shown below:

ASR I'm/0 dialing/0 use/1 my/1 credit/1 card/1

REF I wanna use my credit card

ASR yes/1 I'm/0 trying/1 the/0 calling/1 card/1 call/1

REF yes I'm trying to make a calling card call

ASR hi/0 I'm/0 calling/0 the/0 number/1

REF hi I'm having trouble getting through to the number

where for each transcribed (**REF**) sentence is given the decoded (**ASR**) sequence of word-quantized-confidence-score pairs. The value of the quantized confidence score predicts the confidence on the decoded word. The recognition accuracy improved only slightly from 45% to 46.5%. However, it is very interesting to note that the decoded confidence level obtained during recognition provides a good indication as to whether or not a given word was correctly decoded.

5 UV IN CALL-TYPE CLASSIFICATION

Spoken utterances are classified as to call-type by recognizing and spotting the occurrences of salient events within them. Previously we have used salient phrase fragments for classification [1]. More recently we use grammar fragments [10] that can be regarded as clusters of semantically similar phrases, with a single posterior distribution over the call-types. These fragments have good coverage of the task and reasonably robust statistics, and tend to be less ambiguous than individual words. Moreover they can contain embedded nonterminal symbols representing salient data within the sentence, such as a telephone or calling-card number, which can be of value in determining the call-type. The grammar fragments are automatically acquired from transcribed and labeled training data. Each grammar fragment is represented as a finite-state machine, and a successful match of a path through the finite state machine to a substring of the utterance generates a detection from which call-type classification can follow.

In general there may be multiple occurrences of salient fragments within an utterance, and occurrences may also overlap. First, a confidence score is associated with each occurrence, given by the geometric mean of confidence scores

for the individual words in the phrase. The posterior distribution over the call-types is then scaled by this mean confidence score. For each call-type, the lattice of detections is then parsed to find the highest cumulative scaled posterior probability along a path through non-overlapping detections. These totals are then passed through a simple feed-forward neural network in order to generate an output for each call-type in the range (0,1), which we interpret as a set of probabilities. The network is trained using the transcribed and labeled training data from which the fragments are acquired.

In the call routing task, one of the 15 call types is called "*other*" and the intention is that these calls be transferred immediately to a human agent. There is therefore a criterion for rejection and we can measure the true and false rejection rates for a labeled test set, as well as the true classification rate. At each rank, a call is "rejected" either if the decision for that rank is "*other*" or if the score for that rank is below a given threshold. By varying the threshold we can generate ROC curves of the type shown in Figure 2 which displays the percentage of utterances in the 1000 utterance test set that were correctly classified according to call-type versus the percentage of utterances that were incorrectly rejected by the system. It should be noted that the systems represented by the curves in Figure 2 do not perform as well as the best performing system described in [10]. These differences are attributable to several factors including the fact that both a lower complexity language model and a lower complexity acoustic HMM model are used here. By comparing the solid curves, labeled "baseline+UV", with the dashed "baseline" curves, it is clear that incorporating UV in call-type classification results in a significant improvement in performance.

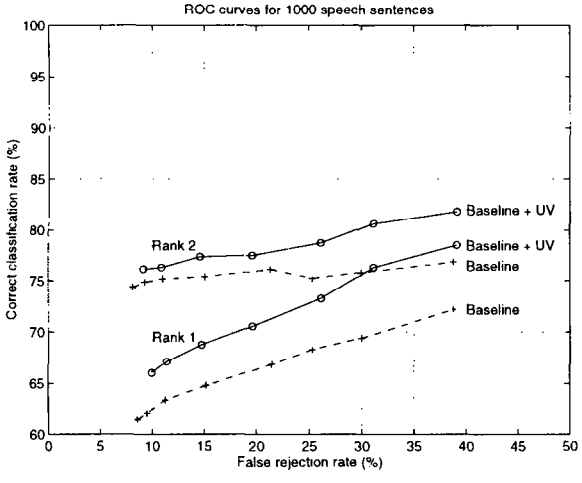

Figure 2: Receiver operating characteristic curves describing the effect of UV on the call-type classification performance for the HMIHY task. The solid curves corresponds to the case where UV scores are integrated into SLU, and the dashed curves correspond to the real-time baseline system implemented without UV.

6 SUMMARY AND CONCLUSIONS

This paper makes three major contributions to the general problem of continuous speech recognition from unconstrained speech utterances. The first contribution is a demonstration of the fact that UV techniques based on acoustic modeling procedures can by themselves help to detect words hypothesized by the speech recognizer that were correctly decoded. The second contribution is a statistically robust method for integrating acoustically derived UV measures with stochastic language models. Finally, a third contribution is the demonstration of how spoken language understanding performance can be improved when acoustic UV measures are integrated into the SLU. Call-type classification error was reduced by as much as 23 percent when utterance verification was used over an equivalent system that did not incorporate UV. The implementation of the techniques and the experimental results presented here represent a first attempt at developing formalisms that result in more closely coupled acoustic, language, and semantic modeling components of spoken language understanding systems.

7 ACKNOWLEDGEMENTS

The authors would like to express their appreciation to A. L. Gorin at ATT Labs Research for his contribution to formulating the HMIHY task and collecting and organizing the speech and text corpora associated with the task.

REFERENCES

[1] A.L.Gorin, G.Riccardi, and J.H.Wright. How may i help you? *Speech Communication (to appear)*, 1997.

[2] S. Cox and R. C. Rose. Confidence measures for the Switchboard database. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, May 1996.

[3] 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.

[4] C. V. Neti, S. Roukos, and E. Eide. Word-based confidence measures as a guide for stack search in speech recognition. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pages 883 886, April 1997.

[5] 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.

[6] G. Riccardi, A. L. Gorin, A. Ljolje, and M. Riley. A spoken language system for automated call routing. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pages 1143–1146, April 1997.

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

[8] 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.

[9] J. G. Wilpon, L. R. Rabiner, C. H. Lee, and E. R. Goldman. Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Trans on Acous. Speech and Sig. Proc.*, 38(11):1870–1878, 1990.

[10] J. H. Wright, A. L. Gorin, and G. Riccardi. Automatic acquisition of salient grammar fragments for call-type classification. *Proc. European Conf. on Speech Communications*, pages 1419–1422, September 1997.

[11] S. R. Young and W. H. Ward. Recognition confidence measures for spontaneous spoken dialog. *Proc. European Conf. on Speech Communications*, pages 1177–1179, September 1993.