ELSEVIER

Speech Communication 34 (2001) 321–331

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Integration of utterance verification with statistical language modeling and spoken language understanding ☆

R.C. Rose *, H. Yao, G. Riccardi, J. Wright

AT&T Labs – Research, Speech and Image Processing Lab, 180 Park Ave., Room D129, Florham Park, NJ 07932, USA

Received 13 January 1999; received in revised form 10 May 2000; accepted 14 June 2000

Abstract

Methods for utterance verification (UV) and their integration into statistical language modeling and understanding formalisms for a large vocabulary spoken understanding system are presented. The paper consists of three parts. First, a set of acoustic likelihood ratio (LR) based UV 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 15 call-types that are accepted by the application.

1. Introduction

This paper is concerned with the acoustic modeling, language modeling, and understanding components of a large vocabulary spoken language understanding system. The goal of the work described in the paper is to develop systems where these individual components are more closely coupled, both in the methods that are used to configure them and in the manner in which they interact in the system's operation. The process of a configuring large vocabulary spoken language understanding system for a task from data has in the past been performed by training stochastic models for each of the individual components separately. Language models and spoken language understanding models have generally been defined over a fixed lexicon obtained from text or transcribed speech utterances, and have *not* themselves incorporated any acoustic knowledge. It will be shown here that automatic speech recognition (ASR) performance can be improved by incorporating representations of acoustic confidence directly into language model training. It will also be shown that spoken language understanding performance can be improved by incorporating acoustic confidence when associating utterances with semantic categories.

☆ This paper is based on a communication presented at ICASSP98, and has been recommended by the Editorial Board of Speech Communication.

* Corresponding author. Tel.: +1-973-3608524; fax: +1-973-3608091.

E-mail addresses: rose@research.att.com (R.C. Rose), s_yao@research.att.com (H. Yao), dsp3@research.att.com (G. Riccardi), jwright@research.att.com (J. Wright).

0167-6393/01/$ - see front matter
PII: S0167-6393(00)00040-6

Utterance verification (UV) techniques are applied in this work to an automated call routing task (Gorin et al., 1997; Riccardi et al., 1997). 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 out-of-vocabulary (OOV), 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 (Lleida and Rose, 1996; Young and Ward, 1993), automation of telecommunications services (Wilpon et al., 1990), and is certainly true in case for machine interpretation of human–human dialog (Cox and Rose, 1996; Rose et al., 1995). Utterance verification in this context implies the ability to detect vocabulary words in an utterance that may 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 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 (Rahim et al., 1996).

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 (Lleida and Rose, 1996). Each hypothesized word or phrase obtained from the ASR decoder is assigned a confidence measure which is 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 (Neti et al., 1997). The approach that is considered here attempts to extend the notion of an SA, which is currently used to describe an N-gram language model for speech recognition (Riccardi et al., 1996). In the simplest case, a state in an 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 modeling 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 semantic categories.

2. Automated call routing task

The utterances used for the experimental study described in this paper were taken from a database of 10,000 spoken transactions between customers and human telephone operators over the public switched telephone network. There were very few repeat callers. We focused on the first customer utterance, the unconstrained response to the greeting prompt of "How may I help you?" (Gorin et al., 1997; Riccardi et al., 1997). These utterances were orthographically transcribed and then labeled

with one or more of fifteen call-type labels. Fourteen of the labels correspond to specific actions that the customer may request, such as *billing-credit*, *collect-call* or *rate-request*, and the last (*other*) subsumes the remainder. Although the callers' spoken language varies widely, most of the time they are asking for one of a moderate number of services, so only a small proportion of the utterances are labeled *other*.

A subset of 2243 utterances was used for training subword acoustic hidden Markov models (HMM), and a subset of 7844 utterances for training language models for recognition and understanding. The overall vocabulary of the training data is approximately 3600 words. A separate subset of 1000 utterances was used for testing. The average utterance duration is 5.9 s, corresponding to an average sentence length of 18 words, and the longest utterance duration is 53 s. The OOV rate at the token level in the test sentences is 1.6%, and at the utterance level 30%. It is difficult to characterize the range of "typical" OOV words as coming from a particular part of speech. Typical utterances that were input to the system are shown in examples given in Sections 4 and 5.

The performance of a fully automatic spoken language system with customer utterances as input and call-type labels as output was evaluated on the test subset. The test utterances were passed through the speech recognizer and then classified using matched salient grammar fragments exploiting the UV weighting as described in Section 5.

3. Acoustic measures for UV

This section presents an LR based procedure for generating word level acoustic confidence measures (Lleida and Rose, 1996, 2000). First, UV is presented in a hypothesis testing framework, and the form of the densities used in the LR based hypothesis test for UV is described. Second, the procedures for training a dedicated set of UV–HMM models is described. Finally, UV performance is presented for the utterances in the call routing task described in Section 3.3.

3.1. UV models and hypothesis testing

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

In the context of UV, it is assumed that subword HMM models are given for each subword unit u in subword set $\mathscr{P}$ for both target hypotheses, $\{\lambda_u^C, u \in \mathscr{P}\}$, and alternative hypotheses, $\{\lambda_u^A, u \in \mathscr{P}\}$. The UV score, S_k, for a given word, w_k, is obtained by combining the LR scores for the acoustic subword units, $u_{k,j}$, $j = 1, \ldots, N_k$, that make up that word. Given Y_u, a sequence of observation vectors that have been decoded as HMM subword unit, u, a log LR score

$$\log \mathscr{LR}_u = \log P(Y_u|\lambda_u^c) - \log P(Y_u|\lambda_u^a) \tag{1}$$

can be computed. The purpose of Eq. (1) is to provide a measure of the degree to which the subword unit explains the data.

A log LR score like the one in Eq. (1) can exhibit undesirable behavior as a result of the large dynamic range that is characteristic of all LRs. In order to deal with this large dynamic range, the following definition is used for the alternate hypothesis probability:

$$P(Y_{u_j}|\lambda_{u_j}^a) = \alpha P(Y_{u_j}|\lambda^{\mathrm{bg}}) + (1-\alpha)P(Y_{u_j}|\lambda_{u_j}^{\mathrm{im}}), \tag{2}$$

where $0 \leqslant \alpha \leqslant 1$. In Eq. (2), the alternate hypothesis probability for subword unit u is a weighted linear combination of probabilities computed from

a subword dependent "imposter" model $\lambda_{u_j}^{\text{im}}$, and a "background" model λ^{bg}, which is shared across all units. The purpose of $\lambda_{u_j}^{\text{im}}$, 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. Each of the subword models $\lambda_{u_j}^{\text{im}}$ and $\lambda_{u_j}^{c}$ are represented as three state left-to-right models. The purpose of λ^{bg}, referred to here as the background alternative model, is to provide a broad representation which "covers" the entire space of acoustic features. This broad representation serves to reduce the dynamic range and to reduce the influence of out-liers on the value of the LR. A single one state, 64 mixture background alternate hypothesis model is shared amongst all "target" HMM models. While a method for estimating the interpolation coefficient α in Eq. (2) was described in (Lleida and Rose, 1996, 2000), α was empirically determined for the experimental evaluation described in Section 3.3.

A word level score S_k is computed as a non-linear weighting of the subword level log LR scores and used to form a decision rule

$$S_k = f(\mathscr{LR}_{u_1}, \ldots, \mathscr{LR}_{u_{N_k}}) \underset{\mathscr{H}_1}{\overset{\mathscr{H}_0}{\gtrless}} \tau, \tag{3}$$

where $f()$ is a geometric mean. The effect of the geometric mean in Eq. (3) is to assign greater weight to units with lower LR scores. As a result, an individual unit with a particularly low LR score can cause the word score S_k to be low. This would allow for the rejection of word w_k if a single subword unit LR score were low. Eq. (3) also describes a decision criterion, where w_k is either accepted or rejected as a correctly decoded word hypothesis by comparing S_k to a decision threshold τ. $\mathscr{H}_0$ represents the null hypothesis corresponding to a portion of the utterance being correctly decoded as word w_k, and $\mathscr{H}_1$ represents the alternate hypothesis.

In incorporating acoustic confidence measures into the SLU system, it is not assumed that individual words have been classified according to the above hypotheses. Instead, a simple non-parametric approach is used for converting the word scores, S_k, to the a posteriori probabilities of the word being correctly decoded given its confidence measure, $P(C = 1|S_k = s)$. Here, $C = 1$ corresponds to the event that w_k is correctly decoded and $S_k = s$ corresponds to the event that the word score S_k takes on the value s. This is the measure that is presented to the SLU system. Empirical distributions were obtained for the above a posteriori probabilities by partitioning the range of word level scores into B discrete bins and computing bin occupancy counts for the scores from the training data to obtain

$$P\left(C = 1 \middle| S_k \in \left[\frac{i-1}{B}, \frac{i}{B}\right], \quad i = 1, \ldots, B\right). \tag{4}$$

A single "back-off" distribution of the same form was trained for those words that had insufficient occurrences in the training data for dedicated distributions to be estimated.

3.2. UV model training

Two training procedures are implemented here for estimating UV model parameters. The first is based on a maximum likelihood (ML) optimization criterion, the second procedure is based on a discriminative procedure that will be referred to here as LR training. More detailed discussion of both procedures for UV model training can be found elsewhere (Lleida and Rose, 1996; Rose et al., 1995).

ML training of the subword unit dependent models, λ_u^c and λ_u^{im}, is performed in two steps. First, speech recognition is performed on a set of development utterances, and subword units corresponding to correct decodings, insertions and substitutions are labeled in the output stream. Second, ML based forward–backward training of the UV HMM models is performed. This is done by updating the conditional expectations for imposter model parameters using decoded units that were labeled as false alarms in the hypothesized strings, and updating the target model conditional expectations using decoded units labeled as correctly decoded. As a result of this process, each subword model $\lambda_{u_j}^{\text{im}}$ is trained to represent the events that are frequently confused with subword unit u_j. It is important to note that the insertions

and substitutions are labeled at the subword level. This implies that if a word substitution occurs in speech recognition where the decoded word differs from the actual word in only a single subword unit, then only that single unit will be recorded as a substitution.

In LR training, the cost function used for estimating HMM model parameters is very similar to the LR criterion that is used for UV (Lleida and Rose, 1996; Rose et al., 1995). A procedure similar to that used above for ML training is performed to obtain a string of decoded subword units that have been labeled as being correctly decoded and false alarms. These labeled units are used in a gradient update procedure for updating the target and alternative hypothesis model parameters. The goal of the procedure is to increase the average LR for those observations corresponding to correctly decoded units and to decrease the average LR for those observations corresponding to false alarms.

It should be noted that there exists a training corpus from the task domain described in Section 2 for all of the acoustic modeling techniques that are investigated here. This is also true for the language and understanding models discussed in Sections 4 and 5. Furthermore, the LR training procedure that is used for UV model training attempts to optimize what can be considered as a discriminative training criterion. Discriminative training algorithms are generally considered appropriate in situations where it is difficult to pose the exact structure of the model and also where a fairly large, labeled task specific speech training corpus is available. Informal experiments have suggested that a reduction in the number of acoustic training utterances result in a reduction in the level of UV performance relative to that presented in Section 3.3. As a result, the issues of portability of these techniques to a new domain where there are limited task specific resources are not directly addressed here.

3.3. Acoustic UV performance

In order to evaluate UV performance, confidence measures of the form described in Section 3.1 were calculated for each word in the decoded word strings obtained from the 1000 utterance test set described in Section 2. These confidence measures were evaluated in terms of their ability to distinguish between correctly and incorrectly decoded words appearing in the output word strings. Only a subset of the total vocabulary words were included in the measure. These were words that were both considered to be "salient" according to the spoken language understanding unit (SLU) and were also not on a list of short duration words whose acoustic realization were considered to be highly dependent on the surrounding context. Examples of words that were used for evaluating acoustic UV performance include *area*, *charge*, *credit*, *direct* and *long*. The 134 words that were chosen accounted for approximately 30% of the total word occurrences in the test utterances.

The performance was evaluated in terms of receiver operating characteristic (ROC) curves as shown in Fig. 1. Each of the curves in Fig. 1 were generated by sweeping a single decision threshold over the a posteriori scores obtained for all decoded occurrences of the set of 134 words. The use of empirical distributions for the estimation of a posteriori probabilities from LRs as described in Section 3.1 serves to reduce the variability of these scores across words in the vocabulary. The vertical axis in Fig. 1 represents the probability of detecting a correctly decoded word hypothesis and the

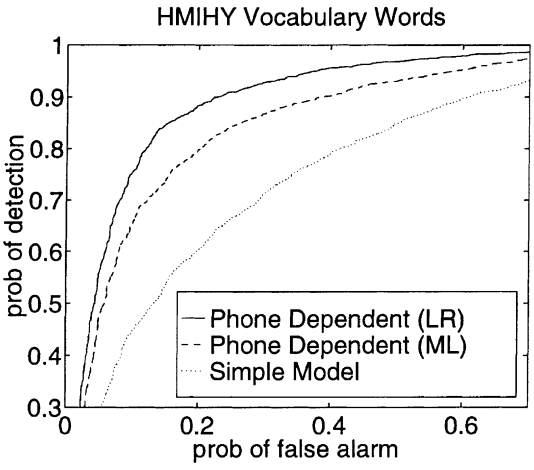

Fig. 1. ROC curves plotted over the confidence measures obtained for three separate sets of UV models.

horizontal axis represents the probability of false acceptance of an incorrectly decoded word hypothesis. The three experiments that were performed correspond to different definitions of the alternate hypothesis model probability densities in Eq. (2), and different UV model training criteria. The "simple model" curve was obtained using only the subword independent background model to define the alternate hypothesis model density so that $P(Y_u|\lambda_u^A) = P(Y_u|\lambda^{\mathrm{bg}})$. The other two "subword dependent" curves were obtained using the linear combination of background and imposter models in Eq. (2) to obtain the subword level log LR scores with $\alpha = 0.2$. Anecdotal experiments showed that performance is not very sensitive to the exact setting of α. For the "ML" curve, UV models were trained using a ML criterion. For the "LR" curve, the UV models were trained using the discriminative LR criterion.

It is clear from Fig. 1 that the use of subword dependent units for representing the alternate hypothesis probability significantly improves the word level detection performance on this task. At a probability of false alarm equal to 20%, the detection probability increases by over 30% when subword dependent units are used. Fig. 1 also shows an increase in performance by 10% when parameters are trained using an LR criterion for training subword dependent parameters.

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.

4.1. Incorporating measures of acoustic confidence

Our approach to integrating acoustic level confidence with the language model is to augment the word n-gram event space, which currently defines linguistic context, with encoded values of acoustic confidence. A stochastic language model 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. In Section 3 we have shown how to estimate UV scores for each word in a speech utterance. Since these two information sources are synchronous, the acoustic and lexical information can be coupled in order to learn a stochastic model based on their joint distribution. Thus, the word string can be replaced by a symbol-pair sequence and 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. In particular, the acoustic and lexical context for word w_i in a third-order statistical n-gram 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 $P(w_i = v|w_{i-1} = u, c_{i-i} = 1)$ 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 n-gram 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 (Riccardi et al., 1996).

4.2. Stochastic transducers for language modeling and UV

Our approach to language modeling is based on stochastic finite state machines (SFSM) learning. The SFSM are a tool for providing a compact and efficient model to represent a wide class of probability distributions over sequences drawn from a finite set (Riccardi et al., 1996). In this work, we consider the variable N-gram stochastic automaton (VNSA) learning algorithm (Riccardi et al., 1996). The VNSA is a non-deterministic SA that allows for parsing any possible symbol sequences $W = w_1, w_2, \ldots, w_N$ drawn from an input language L_i. In its simplest implementation, the state s_i in the VNSA encapsulates the word sequence prefixes observed in the training data. Each state recognizes a symbol w_i and the probability of going from state s_i to state s_{i+1} is $p_i = p(s_{i+1}|s_i)$. The probability of a word sequence W is then associated with the state sequences ξ_j^W and computed as $P(W) = \sum_j P(\xi_j^W)$.

The automatic learning of stochastic finite state automata that incorporate ACC probabilities fits very nicely under the frame-work of the VNSA (Riccardi et al., 1996). 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 n-gram contexts, but can also be invoked to deal with lack of statistical robustness in the estimation of ACC probabilities. Furthermore, the VNSA formalism can be extended to learn joint probability distributions for stochastic transducers. Transducers recognize strings from an input language L_i (e.g. set of all word sequences) and map into strings of an output language L_o (e.g. set of all possible c_i sequences). The class of transducers we consider in this work is called *sequential* transducers. For further details on this literature see (Berstel, 1979). In Fig. 2, we draw an example of a stochastic transducer for the input sequences (*collect call please*, *collect*) and the paired binary score sequence $(1\,0\,1), (1)$. In Fig. 2, the arcs carry the triplet $w_i : c_i/p_i$, where $w_i \in V$, $c_i \in \{0, 1\}$ and p_i are the state transition probability (final states are double-circled). Note that, in the case of stochastic transducers, p_i is the probability of going from state s_i to state s_{i+1} and emitting the symbol c_i in the language L_o. Hence, stochastic transducers allow for local modeling (i.e. state level) of joint probabilities.

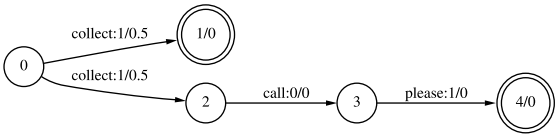

Fig. 2. Example of stochastic transducer. Each arc in the transducer carries the triple $w : c/p$ corresponding to the input and output label and transition probability, respectively.

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 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 (Riccardi et al., 1996).
5. Prune states in VNSA network (Riccardi et al., 1996).

Using this algorithm, the SFSM 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 each speech input utterance with a sequence of word/symbol pairs (w_i, c_i). The next section describes results obtained when a finite state machine for large vocabulary speech recognition was trained using a training set of these word/symbol pair sequences.

4.3. Performance of UV/LM integration

The class of stochastic transducers described in Section 4.2 has been incorporated into the large vocabulary speech recognizer and tested on the 1000 utterance test set for the task described in

Section 2. The speech decoding algorithm performs a standard Viterbi search over all hypothesized word/symbol pairs $W, C = (w_1, c_1), (w_2, c_2), \ldots, (w_N, c_N)$ in order to maximize

$$(\hat{W}, \hat{C}) = \underset{W,C}{\operatorname{argmax}} (A|W)P(W, C), \tag{5}$$

where we assumed that the acoustic likelihood of a word is conditionally independent of the confidence scores, C, i.e. $P(A|W, C) = P(A|W)$. Thus, the prediction on the *best* quantized scores c_i does not involve any on-line computation of LR scores. 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%. To assess the performance of the c_i labeling over the word sequences, we have compared the system just described (c_i scoring based on the stochastic transducers) with the system presented in Section 3 (on-line computation of quantized LR scores). There are two useful figures of merit we have computed to evaluate the accuracy of the UV coding scheme. They are the probability of correctly labeling words for the case of $c_i = 0$ (i.e. misrecognized word) and $c_i = 1$ (correctly recognized word), and they correspond to the probability $P(c_i = \hat{c}_i = 0)$ and $P(c_i = \hat{c}_i = 1)$, respectively. In Table 1, the two figures of merit are shown for the stochastic transducer-based (ST) and on-line LR score computation (OLLR) systems. It is very interesting to note that the ST system provides a good indication as to whether or not a given word was correctly decoded.

Table 1
Figures of Merit I ($P(c_i = \hat{c}_i = 0)$) and II ($P(c_i = \hat{c}_i = 1)$) for the two systems ST and OLLR

System	Figures of Merit	
	I	II
ST	0.470	0.930
OLLR	0.395	0.852

The goal in developing the expanded finite state network described in Section 4.2 was to make it possible to implement a system that can extract word level measures of acoustic confidence during decoding and use coded representations of these confidence measures as the network is expanded. Clearly, the experimental results described in this section suggest that the combination of the ACC probability computation with the on-line LR scores has the potential for improving the ASR performance. There are a number of important issues involved in implementing this fully integrated ACC system. The first is the implementation of a single-pass CSR decoder/UV system that can produce confidence scores frame synchronously and make them available to the finite state network. A low complexity single-pass decoder, designed to directly optimize a LR criterion, has been proposed and evaluated for this purpose in (Lleida and Rose, 1996). A second issue involves exactly how the frame synchronous UV scores are passed to the finite state network, and what heuristics must be implemented in combining the new acoustic and language scores as paths are propagated in the network. These issues are addressed in more detail in (Rose and Riccardi, 1999).

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 (Gorin et al., 1997). These are acquired automatically from the training data by searching the space of phrase

fragments guided by two criteria: the mutual information between the words within a phrase, and the mutual information between the phrase and the set of call-type labels. More recently we use salient grammar fragments for classification (Wright et al., 1997). These are also acquired automatically, by clustering the salient phrases (using a combined string-distortion and semantic-distortion measure) and forming the phrase clusters into finite-state machines. The salient grammar 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 non-terminal symbols representing auxiliary data within the sentence, such as a telephone or calling card number, which can be of value in determining the call-type.

A simple example of a grammar fragment is shown in Fig. 3(a), 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, see Fig. 3(b). In general, there may be multiple occurrences of salient fragments within an utterance, and occurences may also overlap. First, a confidence score is associated with each detected event, given by the geometric mean of UV scores for the individual words in the matched phrase. For each grammar fragment, there exists an associated posterior distribution over the call-types (derived from the training data) and this is scaled by this confidence score. For each call-type, the lattice of detected events is then parsed to find the highest cumulative scaled posterior probability along a path through non-overlapping detections. These cumulative scores are then passed through a single-layer 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 by applying a similar procedure to the transcribed training data and using the manually-assigned labels to determine the required output for each utterance. No UV is involved in classifier training. We test the utility of UV in classification by comparing the results obtained in this way with those obtained when the UV scores are ignored, i.e. no scaling of the posterior distributions.

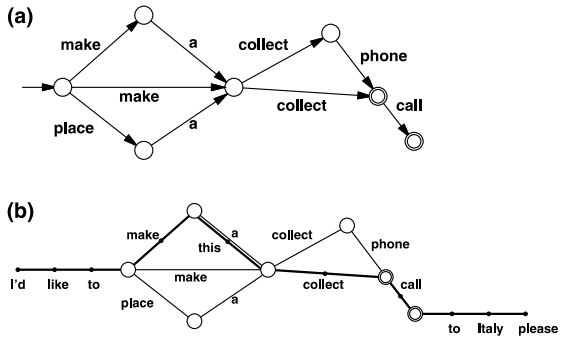

Fig. 3. (a) A simple grammar fragment; (b) approximate match of grammar fragment to an utterance.

In the call-routing task, one of the 15 call-types is called *other* and these are utterances that don't fall into any of the specific categories. The intention is that these calls be transferred immediately to a human agent, so this establishes a criterion for rejection. We can measure the true and false rejection rates for a labeled test set, as well as the true classification rate. A call is rejected either if the decision is "other" or if the score is below a given threshold. By varying the threshold we can generate ROC curves of the type shown in Fig. 4, which displays the percentage of utterances in the 1000 utterance test corpus that were correctly classified according to call-type versus the percentage of utterances that were incorrectly rejected by the system. For the rank 2 curve, a correct classification means that either of the two highest-ranked call-types is correct. Incorporating UV into call-type classification clearly results in a significant improvement in performance at both rank 1 and rank 2.

There follow two examples of test sentences where UV helps to resolve a semantic conflict caused by a recognition error. Salient phrases within the recognizer output are shown in uppercase together with the confidence score. In the first example, the phrase "number for me" (associated with the call-type *dial-for-me* (Gorin et al., 1997) is an error, and receives a low score compared with the correctly-recognized phrase "to get credit" (associated with the actual call-type *billing-credit*). In the second, the phrase *home phone* (associated with the call-type *third-number*) is again an error, and receives a low score compared with the word *collect* (associated with the actual call-type "collect").

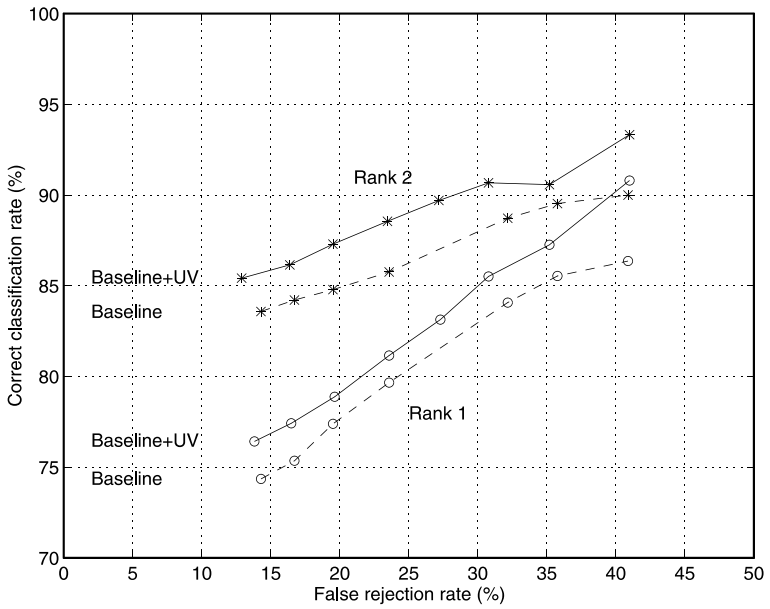

Fig. 4. ROC curves describing the effect of UV on the call-type classification performance for the HMIHY task. The solid curves correspond to the case where UV scores are integrated into SLU, and the dashed curves correspond to the real-time baseline system implemented without UV.

Example 1.

Transcription: "I like to get credit for a misdialed number please"

Recognition: "Hi do you have TO_GET_CREDIT (0.78) for a mister dial this NUMBER_FOR_ME (0.30)"

Example 2.

Transcription: "Well I was trying to make a collect call to a mobile phone that's not possible I guess"

Recognition: "Hi I was trying to make a COLLECT (0.99) call to my HOME_PHONE (0.31) but I'd like to"

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% when UV 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.

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

Berstel, J., 1979. Transductions and Context-Free Languages. Teubner Studienbuchner Informatik.

Cox, S., Rose, R.C., 1996. Confidence measures for the switchboard database. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, May, pp. 511–514.

Gorin, A.L., Riccardi, G., Wright, J.H., 1997. How may I help you? Speech Communication 23, 113–127.

Lleida, E., Rose, R.C., 1996. Efficient decoding and training procedures for utterance verification in continuous speech recognition. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, May, pp. 507–510.

Lleida, E., Rose, R.C., 2000. Utterance verification in continuous speech recognition-decoding and training procedures. IEEE Trans. Speech Audio Process. 8 (2), 126–139.

Neti, C.V., Roukos, S., Eide, E., 1997. Word-based confidence measures as a guide for stack search in speech recognition. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, April, pp. 883–886.

Rahim, M., Lee, C., Juang, B., Chou, W., 1996. Discriminative utterance verification using minimum string verification error training. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, May, pp. 3585–3588.

Riccardi, G., Pieraccini, R., Bocchieri, E., 1996. Stochastic automata for language modeling. Comput. Speech Language 10, 265–293.

Riccardi, G., Gorin, A.L., Ljolje, A., Riley, M., 1997. A spoken language system for automated call routing. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, April, pp. 1143–1146.

Rose, R.C., Riccardi, G., 1999. Automatic speech recognition using acoustic confidence conditioned language models. In: Proceedings of the European Conference on Speech Communications, September, pp. 303–306.

Rose, R.C., Juang, B.H., Lee, C.H., 1995. A training procedure for verifying string hypotheses in continuous speech recognition. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, April, pp. 281–284.

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

Wright, J.H., Gorin, A.L., Riccardi, G., 1997. Automatic acquisition of salient grammar fragments for call-type classification. In: Proceedings of the European Conference on Speech Communications, September, pp. 1419–1422.

Young, S.R., Ward, W.H., 1993. Recognition confidence measures for spontaneous spoken dialog. In: Proceedings of the European Conference on Speech Communications, September, pp. 1177–1179.