How may I help you?

Conference Paper · May 2000

DOI: 10.1109/IVTTA.1996.552741 · Source: IEEE Xplore

Allen L. Gorin
MIT

Jay G. Wilpon
Interactions LLC

HOW MAY I HELP YOU?

A.L. Gorin, B.A. Parker[(1)], R.M. Sachs[(2)] and J.G. Wilpon

AT&T Research
Murray Hill, New Jersey
algor@research.att.com

ABSTRACT

We are interested in providing automated services via natural spoken dialog systems. There are many issues that arise when such systems are targeted for large populations of non-expert users. In this paper, we describe an experimental vehicle to explore these issues, that of automatically routing calls based on a user's fluently spoken response to open-ended prompts such as '*How may I help you?*' A spoken dialog system for call-routing has been constructed, with subsequent processing for information retrieval and form-filling. To enable experimental evaluations, a database has been generated of 10,000 fluently spoken transactions between customers and human agents. We report on preliminary experimental results for that database.

INTRODUCTION

There are a wide variety of interactive voice systems in the world, some residing in laboratories, many actually deployed. Most of these systems, however, either explicitly prompt the user at each stage of the dialog, or assume that the person has already learned the permissible vocabulary and grammar at each point. While such an assumption is conceivable for frequent expert users, it is dubious at best for a general population on even moderate complexity tasks. In this work, we describe progress towards an experimental system which responds appropriately to what people actually say, in contrast with what the system designer expects them to say.

The problem of automatically understanding fluent speech is difficult, at best. There is promise of solution, however, within constrained task domains. In particular, we focus on a system whose initial goal is to understand its input sufficiently to route the caller to an appropriate destination. Such a *call router* need not solve the user's problem, but only transfer the call to someone or something which can. For example, if the input is "*Can I reverse the charges on this call?*", then the caller should be connected to an existing automated subsystem which completes collect calls. Another example might be "*How do I dial direct to Tokyo*?", whence the call

(1) AT&T, Basking Ridge, N.J.
(2) AT&T, Holmdel, N.J.

should be connected to a human agent who can provide dialing instructions. Such a call router should be contrasted with traditional telephone switching, wherein a user must know the phone number of their desired destination, or in recent years navigate a menu system to self-select the desired service. In the method described here, the call is instead routed based on the *meaning* of the user's speech.

This paper proceeds as follows. First, an experimental spoken dialog system is described for call-routing plus subsequent automatic processing of information retrieval and form-filling functions. Second, a database is described of 10K fluently spoken transactions between humans for this task. In particular, we describe the language variability in the first customer utterance, responding to the prompt of '*AT&T. How may I help you?*'. Finally, we report on preliminary experimental results for classifying these utterances from speech. This work extends an earlier system and results described in [G94].

A SPOKEN DIALOG SYSTEM

The goal of a call-router is to recognize and understand the user's speech sufficiently to determine the call-type. Dialog is necessary since, in many situations, the call type cannot be determined from a single input. This can be due to an ambiguous request or to imperfect performance of our spoken language understanding (SLU) algorithms.

One important component of dialog is *confirmation*, wherein the machine proposes its understanding of the user's input, receiving reinforcement feedback as to the appropriateness of its response. In previous work, we have described a dialog model for integrating these user inputs based on a feedback control system [G95]. In case the machine's interpretation is in error, the subsequent user input can provide both disconfirmation and *clarifying* information. Finally, once the call-type has been successfully negotiated and agreed upon, there is often a subsequent form-filling dialog to *complete* the information necessary to provide a service. To

To appear in the Proc. of IVTTA
Philadelphia, October 1996

illustrate these issues, we now present several transcripts of spoken human/machine dialogs with this system. We remark that in the current prototype, there has been little effort as yet to optimize the prompts and messages that the machine generates. This is the subject of ongoing research [A96][B96] which will be integrated at a later date. The first dialog demonstrates the basic call-routing function using *explicit confirmation*, (where *M* denotes machine and *U* denotes a human user).

M: How may I help you?
U: Can you tell me how much it is to Tokyo?
M: You want to know the cost of a call?
U: Yes, that's right.
M: Please hold on for rate information.

The second dialog illustrates an information retrieval stage after call-type classification. Observe that in addition to the call-type, the SLU algorithm must also extract the place-name *Chicago*. Further observe the use of an *implicit confirmation* request [B96], wherein the user's silence is interpreted as affirmation.

M: How may I help you?
U: What's the area code for Chicago?
M: You want an area code.
U: (silence)
M: The area code for Chicago is 312.

In some cases, the input is ambiguous or misunderstood, so the device must then engage in a clarifying dialog. There are several possible strategies for such clarification in this task [B96]. In the example below, the device proposes the most likely interpretation, expecting the response to be a combination of (dis)confirmation and/or clarification.

M: How may I help you?
U: Can I charge this call please?
M: You want to charge this to another number?
U: No, to my Universal card.
M: You want to charge this call to a credit card.
U: (silence)
M: Please speak your card number.

The last example illustrates a call-type classification followed by a form-filling subdialog.

M: How may I help you?
U: I tried calling my Aunt and got a wrong number.
M: You need a billing credit.
U: (silence)
M: Please speak the number that you dialed.
U: 908 582 2778
M: Was the call billed to the phone that you're calling from?
U: Yes it was.
M: You will be given credit, thank you for calling.

There are many user interface issues in such dialogs, in particular those of selecting the best prompts to stimulate maximally consistent responses from large populations of non-expert users [A96][B96]. In the next section, we discuss the variations in response to a particular open-ended prompt, in the situation where users are talking to a person.

DATABASE

In order to enable experimental evaluation, we generated a database of 10K spoken transactions between customers and human agents. First, both channels of the dialog were recorded from the agents' headset jacks onto a digital audio tape (DAT). At the end of each transaction, a control key was manually depressed (by the human agent) to generate a DTMF code, serving both as a segmentation marker and a call-type label. These recordings were then automatically segmented, filtered and downsampled to generate a stereo speech file for each transaction.

We then focused on the first customer utterance, responding to the greeting prompt of "*AT&T. How may I help you*?" These utterances were endpointed, orthographically transcribed and then labeled as to call-type and quality of the speech and channel. We remark on the distinction between the call-action labels provided by the agents and by the labelers. The agent's DTMF tag comprised an on-the-spot single label for the entire transaction. The labelers, however, based their decision on the initial customer utterance only, plus were allowed to select more than one call-label per utterance. We observed that 84% of the utterances were labeled with a single call-type, 16% with two (e.g. DIAL_FOR_ME and COLLECT), then a small remainder (0.6%) with 3 labels. It is possible for the agent-generated call-type to not match any of the labeler's, since sometimes the first utterance is ambiguous, with things becoming clear only after some dialog. An issue for future study is the correlation between these labeling methods, plus an analysis of the reasons for their mismatches.

Several samples of (first) utterances follow, where digits are replaced with the symbol *x*.

yes I need to make a long distance phone call and charge it to my home phone number please

yes how much is it to call the number I just dialed

yes where is area code x x x

yes what time is it in area code x x x right now I'm trying to gauge the time difference

I just I'm trying to get a number from information

Although people's spoken language varies widely, most of the time they are asking for one of a moderate number of services. We selected a subset of 14 services plus an *OTHER* class to subsume the remainder. This distribution is highly skewed, as illustrated in the rank-frequency plot in Figure 1.

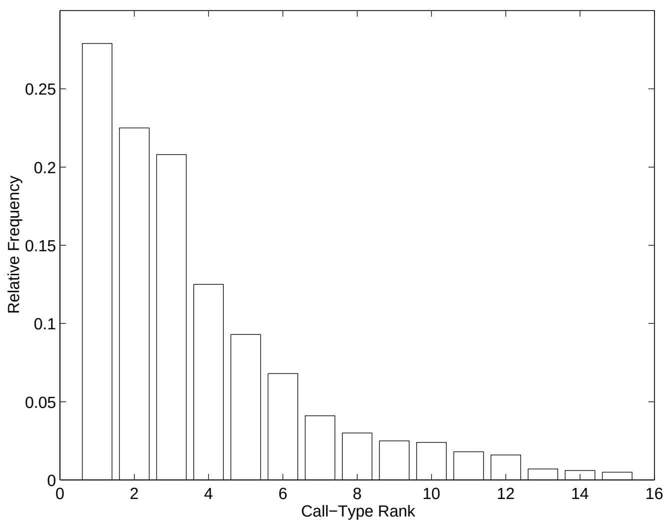

Figure 1. Rank Frequency Distribution of Call-Types

We now discuss the vocabulary in this database. Of the 10K utterances, 8K are used for training language models for recognition and understanding, the remaining 2K reserved for testing. Figure 2 shows the increase in vocabulary size accumulated over the 8K utterances, with a final value of ~3600 words. Even after 8K utterances, the slope of the curve is still significantly positive. We examined the tail of the lexicon, i.e., the last 100 vocabulary words accumulated. Approximately half were names (either people or places), but the other half were regular words (e.g. *authorized*, *realized*, *necessary*, …). The out-of-vocabulary rate on the test sentences is 1.7%, and the test-set perplexity using a statistical bigram model is 21 [R96].

Utterance length varies greatly, from a minimum of one word (e.g. "*Hello*?") to 183, with an average of 20 words/utterance. The distribution of these lengths for the 10K transcriptions is shown in Figure 3. In that same figure, the cumulative distribution is also shown. Observe that almost all of the sentences have length less than 60. Recall that these utterances are the *initial* user response to the greeting prompt.

PRELIMINARY EXPERIMENTAL RESULTS

We report on preliminary results using an early subset of the database, in particular focusing on the two-class problem of distinguishing billing credit requests from the others. A statistical bigram grammar was trained from 2200 transcriptions. 'Off-the-shelf' acoustic subword models for telephone speech were utilized in AT&T's BLASR speech recognizer, with a single dictionary pronunciation for each word. Salient phrase fragments for the 'billing credit' class were acquired from those training transcriptions plus associated call-type labels. The algorithm for phrase fragment induction is described in [G96]. These were then exploited in a simple pattern-matching classifier, spotting for salient phrase fragments in the ASR output. This 'billing credit' detector was then evaluated on a separate test set of 1800 utterances. For comparison, it was also evaluated on the transcriptions of those test utterances.

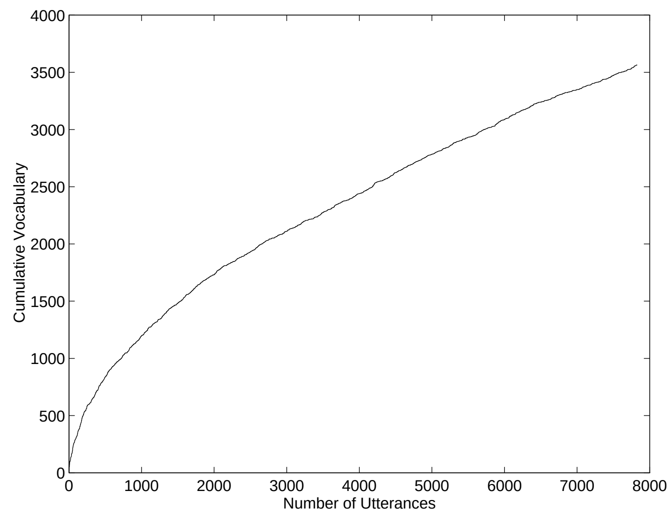

Figure 2. Vocabulary Growth in Database

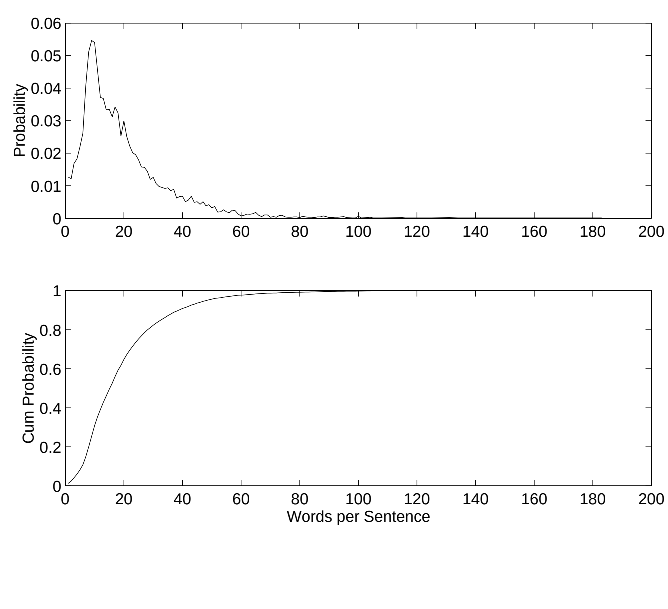

Figure 3. Words per Utterance in the Initial Responses

In such a system, one typically measures two performance figures. The first performance figure is the probability of *correct detection*, i.e., that a caller desired a billing credit and was properly understood. Two such examples are shown below, where the substrings matching a salient phrase fragment are highlighted by capitals and internally connected by underscores.

Correct Detections

yes I_JUST_DIALED AN_INCORRECT_NUMBER.

I_WAS_CUT_OFF when trying to call this number.

The associated error of this type is a *missed detection*, leading to missed opportunities for automation. The resulting action from a missed detection would be to transfer the caller to a human agent. Two such examples follow, where no salient substrings were detected.

Missed Detections

I am trying to call wooster and the number I have rings to a different number.

I'm going to blame this one on my wife I misread her handwriting.

The second performance figure is the probability of *correct rejection,* i.e., an input was *not* a billing credit request and was appropriately rejected. We remark that the associated errors, i.e., *false detections,* lead to misinterpretations which must then be resolved by dialog. Two such examples follow, where salient fragments were detected in utterances that were not billing credit requests.

False Detections

yes I have a number here and I don't know if it's a WRONG_NUMBER.

I was trying to get xxx xxx xxxx and it said it WAS_DISCONNECTED.

The performance curve of Figure 4 was generated by varying the salience threshold for the phrase fragments in the pattern-matching classifier. For details of this method, we refer the reader to [G96]. The costs of the two types of errors are very different within a call-routing task. For example, consider the operating point where the correct detection rate is 40% and the correct rejection rate is 90%. In that case, 40% of the billing credit requests are routed automatically, while the remaining 60% are transferred to a human agent. Of the callers who did not want a billing credit, 90% are correctly rejected while the remaining 10% of false detections require further dialog for correct routing.

CONCLUSIONS

We have described an experimental system for call routing via natural spoken dialog, targeted at large populations of non-expert users. A database was generated of 10,000 human/human transactions for such a task, on which preliminary experimental results have been reported.

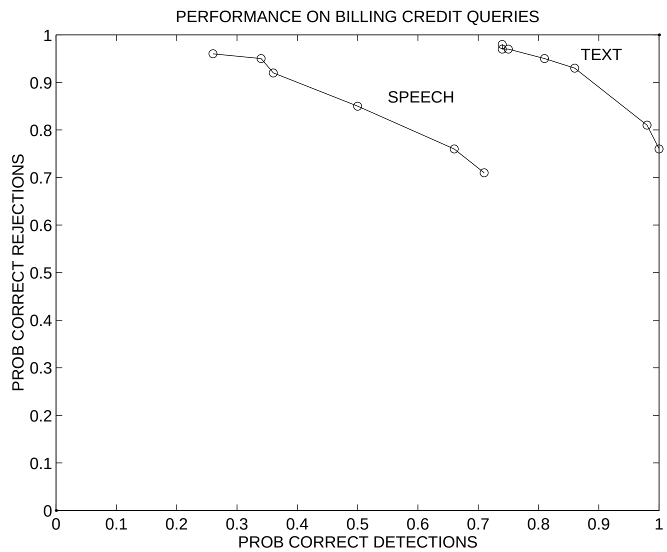

Figure 4. Detecting Billing Credit Queries from Speech

REFERENCES

[G94] A.L. Gorin, H. Hanek, R. Rose and L. Miller, "*Automated Call Routing in a Telecommunications Network*," Proc. IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA), pp. 137-140, Sept. 1994, Kyoto, Japan.

[G95] A.L. Gorin, "*On Automated Language Acquisition*," 97(6), pp. 3441-3461, Journal of the Acoustical Society of America (JASA) , June 1995.

[A96] A. Abella, M. Brown, B. Buntschuh, "*Developing Principles for Dialog-Based Interfaces*," Proc. ECAI Spoken Dialog Systems Workshop, Budapest. August 1996, to appear.

[B96] S. Boyce and A.L. Gorin, " *User Interface Issues for Natural Spoken Dialog Systems*," Proc. Intl. Symp. on Spoken Dialog (ISSD) , Oct. 1996, to appear.

[G96] A.L. Gorin, "*Processing of Semantic Information in Fluently Spoken Language*," Proc. of Intl. Conf. on Spoken Language Processing (ICSLP), Oct. 1996, to appear.

[R96] G. Riccardi, *private communication*, May 1996.