

ROBUST MULTIMODAL UNDERSTANDING

Srinivas Bangalore

AT&T Labs-Research
180 Park Avenue
Florham Park, NJ 07932
srini@research.att.com

Michael Johnston

AT&T Labs-Research
180 Park Avenue
Florham Park, NJ 07932
johnston@research.att.com

ABSTRACT

Contemporary multimodal prototypes provide an excellent proof of concept but are not sufficiently robust in their handling of user input to be adopted by real users engaged in complex tasks. The goal of this paper is to investigate techniques that improve the robustness of multimodal understanding to the point where effective multimodal interfaces can be made feasible. We present two different approaches - a pattern-matching based approach and a classification-based approach to robust multimodal understanding. We compare these approaches by evaluating them on data collected in the context of a multimodal conversational system.

1. INTRODUCTION

Multimodal interfaces that combine spoken and graphical interaction are a natural choice for a broad range of applications where personnel require rapid and effective access to information while mobile. However, while contemporary multimodal prototypes provide an excellent proof of concept, they are not sufficiently robust in their handling of user input to be adopted by real users engaged in complex tasks. They are prone to recognition errors from the component modalities and brittle with respect to unexpected or ill-formed inputs. The goal of this paper is to investigate techniques that improve the robustness of multimodal input processing and understanding to the point where effective multimodal interfaces can be made feasible.

Robustness in understanding refers to the property of the semantic interpretation component that irrespective of the grammaticality of an input utterance allows the interpreter to produce a (partial) meaning representation for the utterance. This issue has been of great interest in the context of speech-only conversational systems [1, 2, 3, 4, 5, 6, 7, 8, 9], but has received little attention in the context of multimodal systems. The robustness techniques adopted in these systems can be characterized by a few approaches: heuristic grammar-based partial parsing approach [1, 2, 4, 5], probabilistic grammar-based approach [3, 6, 9] and direct translation-based approach [7, 8]. [10] use a grammar-based partial parser [1] for speech understanding in a multimodal calendar application.

In this paper, we investigate approaches to achieve robust understanding in the context of a multimodal application designed to provide an interactive city guide: MATCH. In Section 2, we present the MATCH application, the architecture of the system and the apparatus for multimodal understanding. We discuss our approaches to robust understanding in Section 3. In Section 4, we present the data collection and performance results of experiments on the collected data set.

2. THE MATCH APPLICATION

MATCH (Multimodal Access To City Help) is a working city guide and navigation system that enables mobile users to access restaurant and subway information for New York City (NYC) [11, 12].

The user interacts with a graphical interface displaying restaurant listings and a dynamic map showing locations and street information. The interactions can be using speech, by drawing on the display with a stylus, or using synchronous multimodal combinations of the two modes. The user can ask for the review, cuisine, phone number, address, or other information about restaurants and subway directions to locations. The system responds with graphical callouts on the display, synchronized with synthetic speech output. For example, if the user says *phone numbers for these two restaurants* and circles two restaurants as in Figure 1 [a], the system will draw a callout with the restaurant name and number and say, for example *Time Cafe can be reached at 212-533-7000*, for each restaurant in turn (Figure 1 [b]). If the immediate environment is too noisy or public, the same command can be given completely in pen by circling the restaurants and writing *phone*.



Fig. 1. Two area gestures

2.1. MATCH Multimodal Architecture

The underlying architecture that supports MATCH consists of a series of re-usable components which communicate over sockets through a facilitator (MCUBE) (Figure 2). Users interact with the system through a Multimodal User Interface Client (MUI). Their speech and ink are processed by speech recognition [13] (ASR) and handwriting/gesture recognition (GESTURE, HW RECO) components respectively. These recognition processes result in lattices of potential words and gestures. These are then combined and assigned a meaning representation using a multimodal finite-state device (MMFST) [14, 11]. This provides as output a lattice encoding all of the potential meaning representations assigned to the user inputs. This lattice is flattened to an N-best list and passed to a multimodal dialog manager (MDM) [11], which re-ranks them in accordance with the current dialogue state. If additional information or confirmation is required, the MDM enters into a short information gathering dialogue with the user. Once a command or query is complete, it is passed to the multimodal generation component (MMGEN), which builds a multimodal *score* indicating a coordinated sequence of graphical actions and TTS prompts. This score is passed back to the Multimodal UI (MUI). The Multimodal UI coordinates presentation of graphical content with synthetic speech output using the AT&T Natural Voices TTS engine [15]. The sub-

way route constraint solver (SUBWAY) that identifies the best route between any two points in New York City.

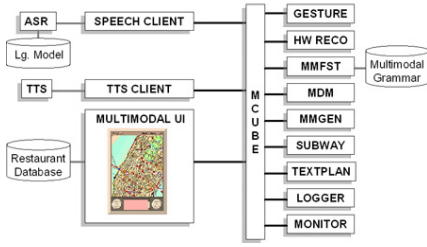


Fig. 2. Multimodal Architecture

2.2. Multimodal Integration and Understanding

Our approach to integrating and interpreting multimodal inputs [11, 12] is an extension of the finite-state approach proposed by Johnston and Bangalore [16, 14]. In this approach, a declarative multimodal grammar representation captures both the structure and the interpretation of multimodal and unimodal commands. The grammar representation consists of a series of context-free rules. The multimodal aspects of the grammar become apparent in the terminals, each of which is a triple $W:G:M$, consisting of speech (words, W), gesture (gesture symbols, G), and meaning (meaning symbols, M). The multimodal grammar encodes not just multimodal integration patterns but also the syntax of speech and gesture, and the assignment of meaning. The meaning is represented in XML, facilitating parsing and logging by other system components. The symbol SEM is used to abstract over specific content such as the set of points delimiting an area or the identifiers of selected objects. In Figure 3, we present a small simplified fragment from the MATCH application capable of handling information seeking commands such as *phone for these three restaurants*. The epsilon symbol (ϵ) indicates that a stream is empty in a given terminal.

CMD	→	$\epsilon:\epsilon:\langle cmd \rangle$ INFO $\epsilon:\epsilon:\langle /cmd \rangle$
INFO	→	$\epsilon:\epsilon:\langle type \rangle$ TYPE $\epsilon:\epsilon:\langle /type \rangle$ for: $\epsilon:\epsilon$ $\epsilon:\epsilon:\langle obj \rangle$ DEICNP $\epsilon:\epsilon:\langle /obj \rangle$
TYPE	→	phone: $\epsilon:\epsilon:phone$ review: $\epsilon:\epsilon:review$
DEICNP	→	DDETPL $\epsilon:\epsilon:area:\epsilon$ $\epsilon:\epsilon:sel:\epsilon$ NUM HEADPL
DDETPL	→	these: $G:\epsilon$ those: $G:\epsilon$
HEADPL	→	restaurants: $rest:\langle rest \rangle$ SEM: $SEM:\epsilon$ $\epsilon:\epsilon:\langle /rest \rangle$
NUM	→	two: $2:\epsilon$ three: $3:\epsilon$... ten: $10:\epsilon$

Fig. 3. Multimodal grammar fragment

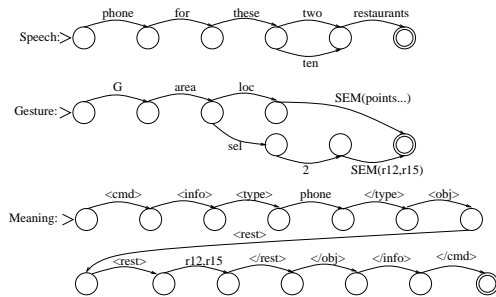


Fig. 4. Multimodal Example

In the example above where the user says *phone for these two restaurants* while circling two restaurants (Figure 1 [a]), assume the speech recognizer returns the lattice in Figure 4 (Speech). The

gesture recognition component also returns a lattice (Figure 4, Gesture) indicating that the user’s ink is either a selection of two restaurants or a geographical area. The multimodal grammar (Figure 3) expresses the relationship between what the user said, what they drew with the pen, and their combined meaning, in this case Figure 4 (Meaning). The meaning is generated by concatenating the meaning symbols and replacing SEM with the appropriate specific content: $\langle cmd \rangle \langle info \rangle \langle type \rangle$ phone $\langle /type \rangle \langle obj \rangle \langle rest \rangle [r12,r15] \langle /rest \rangle \langle /obj \rangle \langle /info \rangle \langle /cmd \rangle$. For the purpose of evaluation of concept accuracy, we developed an approach similar to Boros et al [17] in which computing concept accuracy is reduced to comparing strings representing core contentful concepts. We extract a sorted flat list of attribute value pairs that represents the core contentful concepts of each command from the XML output. The example above yields:

$$cmd : info \ type : phone \ object : selection. \quad (1)$$

The multimodal grammar is compiled into a finite-state device using standard approximation techniques ([18]). The result is used for creating language models for ASR, aligning the speech and gesture results from the respective recognizers and transforming the multimodal utterance to meaning. All these operations are achieved using finite-state transducer operations (See [16, 14] for details).

3. ROBUST MULTIMODAL UNDERSTANDING

The grammar-based interpreter uses composition operations to transduce multimodal strings (gesture,speech) to an interpretation. The set of speech strings that can be assigned an interpretation are exactly those that are represented in the grammar. It is to be expected that the accuracy of meaning representation will be reasonable, if the user’s input matches one of the multimodal strings encoded in the grammar. But for those user inputs that are not encoded in the grammar, the system will not return a meaning representation. In order to improve the usability of the system, we expect it to produce a (partial) meaning representation, irrespective of the grammaticality of the user’s input and the coverage limitations of the grammar. It is this aspect that we refer to as robustness in understanding.

3.1. Pattern Matching Approach

In order to overcome the possible mismatch between the user’s input and the language encoded in the multimodal grammar (λ_g), we use an edit-distance based pattern matching algorithm to coerce the set of strings (\mathcal{S}) encoded in the lattice resulting from ASR (λ_S) to match one of the strings that can be assigned an interpretation. The edit operations (insertion, deletion, substitution) can either be word-based or phone-based and are associated with a cost. These costs can be tuned based on the word confusions present in the domain. The edit operations are encoded as a transducer (λ_{edit}) as shown in Figure 5 and can apply to both one-best and lattice output of the recognizer. We are interested in the string with the least number of edits ($argmin$) that can be assigned an interpretation by the grammar. This can be achieved by composition (\circ) of transducers followed by a search for the least cost path through a weighted transducer as shown below.

$$s^* = argmin_{s \in \mathcal{S}} \lambda_S \circ \lambda_{edit} \circ \lambda_g \quad (2)$$

This approach is akin to example-based techniques used in other areas of NLP such as machine translation. In our case, the set of examples (encoded by the grammar) is represented as a finite-state machine.

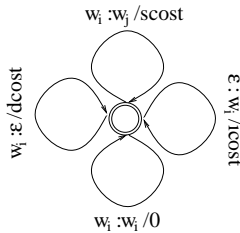


Fig. 5. Edit transducer with insertion, deletion, substitution and identity arcs. w_i and w_j could be words or phones. The costs on the arcs are set up such that $\text{scost} < \text{icost} + \text{dcost}$.

3.2. Classification-based Approach

A second approach is to view robust multimodal understanding as a sequence of classification problems in order to determine the *predicate* and *arguments* of an utterance. The meaning representation shown in (1) consists of an predicate (the command attribute) and a sequence of one or more argument attributes which are the parameters for the successful interpretation of the user’s intent. For example, in (1), `cmd : info` is the predicate and `type : phone` `object : selection` is the set of arguments to the predicate.

We determine the predicate (c^*) for a N token multimodal utterance (S_1^N) by maximizing the posterior probability as shown in Equation 3. We use a large set of features of S to determine the posterior probability and hence resort to a classification algorithm that is robust to large set of features for this purpose.

$$c^* = \underset{c}{\operatorname{argmax}} Pr(c | S_1^N) \quad (3)$$

We view the problem of identifying and extracting arguments from a multimodal input as a problem of associating each token of the input with a specific tag that encodes the label of the argument and the span of the argument. These tags are drawn from a tagset which is constructed by extending each argument label by three additional symbols I, O, B , following [19]. These symbols correspond to cases when a token is inside (I) an argument span, outside (O) an argument span or at the boundary of two argument spans (B) (See Table 1).

User Utterance	cheap thai upper west side
Argument Annotation	<price> cheap </price> <cuisine> thai </cuisine> <place> upper west side </place>
IOB Encoding	cheap_price thai_cuisine upper_place<I> west_place<I> side_place<I>

Table 1. The $\{I,O,B\}$ encoding for argument extraction.

Given this encoding, the problem of extracting the arguments is a search for the most likely sequence of tags (T^*) given the input multimodal utterance S_1^N as shown in Equation (4). We approximate the posterior probability $Pr(T | S_1^N)$ using independence assumptions as shown in Equation (5).

$$T^* = \underset{T}{\operatorname{argmax}} Pr(T | S_1^N) \quad (4)$$

$$\approx \underset{T}{\operatorname{argmax}} \prod_i Pr(t_i | S_{i-n}^i, S_{i+1}^{i+n-1}, t_{i-1}, t_{i-2}) \quad (5)$$

Owing to the large set of features that are used for predicate identification and argument extraction, we estimate the probabilities using a classification model. In particular we use the Adaboost classifier [20] wherein a highly accurate classifier is build by combining many “weak” or “simple” base classifiers f_i , each of which

may only be moderately accurate. The selection of the weak classifiers proceeds iteratively picking the weak classifier that correctly classifies the examples that are misclassified by the previously selected weak classifiers. Each weak classifier is associated with a weight (w_i) that reflects its contribution towards minimizing the classification error. The posterior probability of $Pr(c | x)$ is computed as in Equation 6.

$$Pr(c | x) = \frac{1}{(1 + e^{-2 \cdot \sum_i w_i \cdot f_i(x)})} \quad (6)$$

4. EXPERIMENTS AND RESULTS

In this section, we describe a set of experiments to evaluate the two robust multimodal understanding approaches presented in Section 3. The corpus of multimodal data used for this study was collected in a laboratory setting from a set of sixteen first time users (8 male, 8 female). The subjects were AT&T personnel with no prior knowledge of the system and no experience building spoken or multimodal systems. A total of 833 user interactions (218 multimodal / 491 speech-only / 124 pen-only) resulting from six sample task scenarios involving finding restaurants of various types and getting their names, phones, addresses, or reviews, and getting sub-way directions between locations were collected and annotated.

For the purpose of our experiments we use the subset of 709 utterances that involved speech only and multimodal exchanges. We use concept token accuracy and concept string accuracy as evaluation metrics for the entire meaning representation in these experiments. These metrics correspond to the word accuracy and string accuracy metrics used for ASR evaluation. We also report the accuracy of identifying the predicates and arguments using string accuracy metrics. All results presented in this section are based on 10-fold cross-validation experiments run on the 709 utterances.

We used a class-based trigram model trained on the collected corpus as the language model for the ASR in all these experiments. We defined different classes such as areas of interest (e.g. riverside park, turtle pond), points of interest (e.g. Ellis Island, United Nations Building), type of cuisine (e.g. Afghani, Indonesian), price categories (e.g. moderately priced, expensive), and neighborhoods (e.g. Upper East Side, Chinatown). The ASR performed at a word-accuracy of 73.8% and a sentence accuracy of 57.1%. In other work [21], we have used the speech component of the multimodal grammar to construct a language model.

The baseline multimodal understanding system composes the input multimodal string with the grammar to produce an interpretation. Thus an interpretation can be assigned to only those multimodal strings that are encoded in the grammar. However, the result of ASR and gesture recognition may not be one of the strings encoded in the grammar, and such strings are not assigned an interpretation. This fact is reflected in the low concept string accuracy shown in Table 2.

The pattern-matching based robust understanding approach mediates the mismatch between the strings that are recognized and the strings that can be assigned an interpretation. We experimented with word based pattern matching as well as phone based pattern matching on the one-best output of the recognizer. As shown in Table 2, the pattern-matching robust understanding approach improves the concept accuracy significantly. Furthermore, the phone-based matching method outperforms the word-based matching method.

For the classification-based approach to robust understanding we used a total of 10 predicates such as *help*, *assert*, *inforequest*, and 20 argument types such as *cuisine*, *price*, *location*. We use unigrams, bigrams and trigrams appearing in the multimodal utterance as weak classifiers for the purpose of predicate classification. In order to predict the tag of a word for argument extraction, we use

	Predicate String Accuracy(%)	Argument String Accuracy(%)	Concept Token Accuracy(%)	Concept String Accuracy(%)
Baseline	65.2	52.1	53.5	45.2
Word-based Pattern-Matching	73.7	62.4	68.1	59.0
Phone-based Pattern-Matching	73.7	63.8	67.8	61.3
Classification-based	84.1	59.1	73.5	56.4

Table 2. Performance results of robust multimodal understanding

the left and right trigram context and the tags for the preceding two tokens as weak classifiers. The results are presented in Table 2.

Both the approaches to robust understanding outperform the baseline model significantly. However it is interesting to note that while the pattern-matching based approach has a better argument extraction accuracy, the classification based approach has a better predicate identification accuracy. We see two possible explanations for this difference. First, the pattern-matching based approach attempts at a globally consistent meaning representation and hence more conducive for argument extraction while the classification-based approach relies on local information which is more conducive for identifying the simple predicates in MATCH. Second, the pattern-matching approach uses the entire grammar as a model for matching while the classification approach is trained on the training data which is significantly smaller when compared to the number of examples encoded in the grammar.

5. CONCLUSION

Robust multimodal understanding is essential for improving the usability of a multimodal interface. In this paper, we have presented two approaches to achieve robust multimodal understanding - a pattern-matching based approach and a classification-based approach. We have evaluated these approaches in the context of a multimodal prototype application and demonstrated that both these approaches significantly outperform the baseline understanding system.

6. ACKNOWLEDGMENTS

We thank Patrick Ehlen, Helen Hastie, Candy Kamm, Amanda Stent, Guna Vasireddy, and Marilyn Walker for their contributions to the MATCH system. We also thank Allen Gorin, Mazin Rahim, Giuseppe Riccardi, and Juergen Schroeter for their comments on earlier versions of this paper.

7. REFERENCES

- [1] W. Ward, "Understanding spontaneous speech: the phoenix system," in *ICASSP*, 1991.
- [2] J. Dowding, J. M. Gawron, D. E. Appelt, J. Bear, L. Cherny, R. Moore, and D. B. Moran, "GEMINI: A natural language system for spoken-language understanding," in *Proceedings of ACL*, 1993, pp. 54–61.
- [3] S. Seneff, "A relaxation method for understanding spontaneous speech utterances," in *Proceedings, Speech and Natural Language Workshop*, San Mateo, CA, 1992.
- [4] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, "An architecture for a generic dialogue shell," *JNLE*, vol. 6, no. 3, 2000.
- [5] A. Lavie, *GLR*: A Robust Grammar-Focused Parser for Spontaneously Spoken Language*, Ph.D. thesis, Carnegie Mellon University, 1996.
- [6] Scott Miller, David Stallard, Robert Bobrow, and Richard Schwartz, "A fully statistical approach to natural language interfaces," in *Proceedings of ACL*, June 1996.
- [7] Klaus Macherey, Franz Josef Och, and Hermann Ney, "Natural language understanding using statistical machine translation," in *Proceedings of Eurospeech*, 2001.
- [8] Kishore Papineini, Salim Rukous, and Todd Ward, "Feature-based language understanding," in *Proceedings of the European Conference on Speech Communication and Technology*, Rhodes, Greece, September 1997, pp. 1435–1438.
- [9] M. Rayner and B. A. Hockey, "Transparent combination of rule-based and data-driven approaches in speech understanding," in *In Proceedings of the EACL 2003*, 2003.
- [10] M. T. Vo and C. Wood, "Building an application framework for speech and pen input integration in multimodal learning interfaces," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996.
- [11] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor, "MATCH: An architecture for multimodal dialog systems," in *Proceedings of ACL*, Philadelphia, 2002.
- [12] M. Johnston, S. Bangalore, A. Stent, G. Vasireddy, and P. Ehlen, "Multimodal language processing for mobile information access," in *In Proceedings of ICSLP*, Denver, CO, 2002.
- [13] R.D. Sharp, E. Bocchieri, C. Castillo, S. Parthasarathy, C. Rath, M. Riley, and J. Rowland, "The Watson speech recognition engine," in *In Proceedings of ICASSP*, 1997, pp. 4065–4068.
- [14] M. Johnston and S. Bangalore, "Finite-state multimodal parsing and understanding," in *Proceedings of COLING*, Saarbrücken, Germany, 2000.
- [15] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T next-generation TTS," in *In Joint Meeting of ASA; EAA and DAGA*, 1999.
- [16] S. Bangalore and M. Johnston, "Tight-coupling of multimodal language processing with speech recognition," in *Proceedings of ICSLP*, Beijing, China, 2000.
- [17] M. Boros, W. Eckert, F. Gallwitz, G. Görz, G. Hanrieder, and H. Niemann, "Towards Understanding Spontaneous Speech: Word Accuracy vs. Concept Accuracy," in *Proceedings of ICSLP*, Philadelphia, 1996.
- [18] M-J. Nederhof, "Regular approximations of CFLs: A grammatical view," in *Proceedings of the International Workshop on Parsing Technology*, Boston, 1997.
- [19] Lance Ramshaw and Mitchell P. Marcus, "Text chunking using transformation-based learning," in *Proceedings of the Third Workshop on Very Large Corpora*, MIT, Cambridge, Boston, 1995.
- [20] R.E. Schapire, "A brief introduction to boosting," in *Proceedings of IJCAI*, 1999.
- [21] Srinivas Bangalore and Michael Johnston, "Balancing data-driven and rule-based approaches in the context of a multimodal conversational system," in *Proceedings of Automatic Speech Recognition and Understanding*, 2003.