Multi-Modal Conversational Search and Browse

Larry Heck[1]*, Dilek Hakkani-Tür*[1]*, Madhu Chinthakunta*[1]*,Gokhan Tur*[1]
Rukmini Iyer[2]*, Partha Parthasarathy*[2]*, Lisa Stifelman*[2]*, Elizabeth Shriberg*[1]*, Ashley Fidler*[2]

[1]Microsoft Research, Mountain View, CA
[2]Microsoft, Sunnyvale, CA
larry.heck@ieee.org

Abstract

In this paper, we create an open-domain conversational system by combining the power of internet browser interfaces with multi-modal inputs and data mined from web search and browser logs. The work focuses on two novel components: (1) dynamic contextual adaptation of speech recognition and understanding models using visual context, and (2) fusion of users' speech and gesture inputs to understand their intents and associated arguments. The system was evaluated in a living room setup with live test subjects on a real-time implementation of the multimodal dialog system. Users interacted with a television browser using gestures and speech. Gestures were captured by Microsoft Kinect skeleton tracking and speech was recorded by a Kinect microphone array. Results show a 16% error rate reduction (ERR) for contextual ASR adaptation to clickable web page content, and 7-10% ERR when using gestures with speech. Analysis of the results suggest a strategy for selection of multimodal intent when users clearly and persistently indicate pointing intent (e.g., eye gaze), giving a 54.7% ERR over lexical features.

Index Terms: spoken dialog systems, spoken language understanding, multi-modal fusion, conversational search, conversational browsing.

1. Introduction

Spoken dialog (conversational) systems have seen considerable advancements over the past two decades [1]. A variety of practical goal-oriented conversational systems have been built and deployed. The goal of these systems is to automatically identify the intent of the user as expressed in natural language, extract associated arguments or slots, and take actions accordingly to satisfy the user's requests.

A major limitation of conversational systems is their narrow scope; conversational systems are constrained to operate over a small number of narrowly defined, known domains, with hand-crafted domain-dependent schemas (ontologies). As a result, there has been an increased level of interest by the research community to create open-domain conversational systems. These systems utilize very broad vocabularies, grammars, and intent models. However, the breadth of domain coverage comes at the cost of lower accuracy; without the constraints of limited tasks, speech-enabled systems are often unable to cope with the complexity open-domain speech recognition and understanding.

Advances in hand-held devices, touch displays and vision processing technology provide an opportunity for the speech community to increase the domain coverage of conversational systems. Rather than relying on spoken input only, systems can exploit the *visual constraints* introduced by touch, gesture, and eye gaze. For example, pointing gestures can be used to narrow the focus of attention to sub-region of the visual presentation, giving the conversational system useful priors on what to expect the user to say (e.g., selecting an item by pointing at it and saying "that one"). Since Bolt's seminal work on voice and gesture at the graphics interface [2], several studies investigated use of multi-modality for conversational interactions with a machine. Previous studies investigated the use of pointing gestures [3], touch gestures (including selection of items or an area on the screen for example with a remote control [4, 5], with finger [6] or with a pen [7]), and gaze and head-pose [8].

Another promising source of constraints for open-domain conversational systems is data from web search and internet browsers [9]. Web search engines and browsers are perhaps the most pervasive, ubiquitous open-domain tools available to people today to find information and complete transactions. In many ways, search and browse have elements of automated conversational interactions, or the "interactive, spontaneous communication between two or more [agents] who are following rules of etiquette" [10]. Search and browse conversations are interactive because the system responds to what has previously been communicated. The conversations are spontaneous because the user is not constrained by domain. Developers of search engines and browsers place considerable emphasis on the design of interactions. These interaction models in many ways are patterned after rules of etiquette of human-human conversations, with designs considering how to maximize information flow while minimizing unpleasant interruptions (e.g., relevance versus monetization).

Early work on leveraging search engines and browsers focused on exploiting offline information in the user logs: queries and corresponding clicks on links (documents) from search engines and browsers capturing interactions over many hundreds of millions of users and sessions. Work on exploiting the query-click graphs include [11–16]. More recent work has focused on human-computer addressee detection for conversational browsing [17], as well as methods to exploit the combination of search logs and semantic graphs [18–21].

In this paper, we create an open-domain conversational system by combining the power of internet browser interfaces with multi-modal inputs and data mined from web search and browser logs. We focus on two input modes, speech and gesture, and combine them to interact with browser and web page interfaces and page elements (e.g., links, drop-down menus, forms). By utilizing the pre-existing interaction mechanisms of web pages, we are able to by-pass the requirement to craft interactive user experiences for each domain of interest. In this way, the system inherits the open-domain designs and protocols of internet searching and browsing.

2. Conversational Scenario

In the conversational search and browse scenario, a user is free to navigate and interact with any page on the web through natural conversations with the machine. The user can speak with no constraints on their vocabulary, grammar, or choice of intent. As the user is browsing, they may choose to refer to content on the current page or not. Users may select links of the page contents in at least 3 ways:

1. **Explicit clicks:** User utterance refers to a link on the page, such as "show me Il Fornaio" or "Il Fornaio" in Figure 1. The utterance may be accompanied with hand gestures and eye focus.
2. **Location referrals:** User's utterance may include the relative position of the hyper-link on the page, such as "click on the top one". These may again be accompanied with gestures and eye focus.
3. **Gesture and speech:** Users may click on a link by gesturing in combination with speech, for example, pointing to the link and saying "that one", where the spoken utterance does not overlap with the anchor text of the hyper-link.

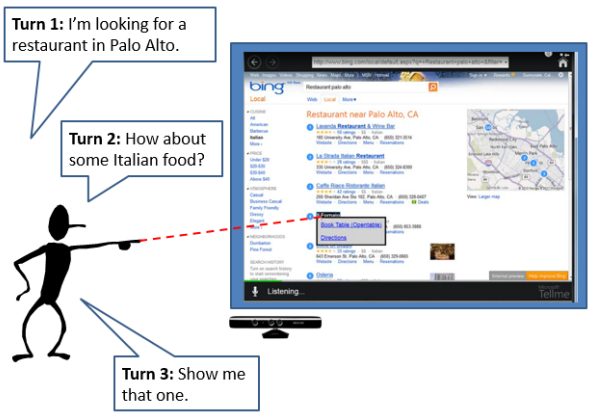

Figure 1: *Example multimodal (speech + gesture) scenario.*

Developing a system to enable the above scenario presents several technical challenges. First, the system must decide whether the user is referring to content on the current page or another page. If the user refers to the current page, the system must capture the intent: click, fill a form, scroll up/down, etc. In Figure 1, the user's intent in Turn 1 (from another web page) was to navigate to this current page by saying "I'm looking for a restaurant in Palo Alto". Turn 2 refines the content on the page to only show Italian restaurants. And finally, in Turn 3, the intent was to select the restaurant link they gestured towards. If the user had said "now show me what's playing at the closest theater", the system would need to recognize the shift in user intent/task as well as understand that the user is not referring to any content on the page, but rather wants to navigate to a movie theater listings web page.

3. Context Adaptation

A particularly effective method to reduce complexity of conversational systems is adapting to context. The context is in multiple forms. Some of the more common examples include:

- **Visual Context**: used to increase the prior likelihood the user will refer to entities/relations on the page
- **Dialog Context**: used for grounding, co-reference resolution, as well as potentially more complex inference and reasoning
- **Personal Context**: used to increase the prior of choices based on personal preferences from histories, geographically, etc.

In this paper, we leverage visual context. We use maximum-a-posteriori (MAP) unsupervised adaptation to adapt the statistical language model (SLM) of the speech recognizer to the content on the page [22]. The adaptation text can either be extracted from the page links (anchor text/titles) and/or landing page content. The extraction can be completed at either run-time or during an offline web page crawl procedure. For the example in Figure 1, the listed restaurant names, street names, food genre are all extracted from the scrape of the link/anchor text. The SLM probabilities and lexicons for names such as "Il Fornaio" can be increased to reflect the given visual context. The details of the restaurant found on the landing page of the link can be included in the adaptation data as well.

In addition to adapting the speech recognizer, the visual context can be used to adapt the semantic components of the system. For the example of Figure 1, priors of intents related to restaurants would be increased (reservations, reviews, etc.). Each of the links represents a new intent that can be dynamically added to the system.

An advantage of adapting to the page content at run-time versus crawl-time is the scalability of the solution: the system is always "fresh" and able to support conversational interaction on a page even if its content has recently changed. This is particularly important for dynamic pages (restaurant/movie reviews, breaking news, sporting results).

By following the above procedure to dynamically adapt to the visual context, the system in effect scales to the breadth of the web. By adjusting priors based on the visual content of the page, as well as related/connected pages (landing pages), the system can achieve this scale *robustly*, as will be demonstrated in the experiments of Section 5.

4. Multimodal Click Intent Detection

In addition to expressing intent verbally, a user may find it more natural in certain situations to express their intent visually. The simultaneous combination of two modes of intent expression is referred to as multimodal intent. This paper focuses on the combination of speech and hand gesture. Specifically, we study the effect of speaking while pointing at an object, such as a link on a web page. The scenario in Figure1 shows a user pointing at a restaurant link and saying "Show me that one." Multimodal interactions such as these are powerful, saving time by reducing dialog turns as well as intent/speech recognition errors.

In the following, we discuss each mode of intent capture separately, and then how they are combined. Then we show experimental results that illustrate the power of the resulting multimodal user interface.

4.1. Lexical Intent

Given the dynamic nature of web pages, we seek an effective lexical intent similarity measure that can be implemented without the requirement for supervised training. For this purpose, we utilize the well known *term frequency-inverse document*

frequency (TF-IDF) similarity measure from web search relevance [23].

For our purposes, we treat the k-th actionable element on the web page, p_k, (e.g., link, drop-down menu, form) as a document. We will refer to the user's utterance as a query, q. The TF-IDF similarity between the query, q, and the page element, p_k, is given as

$$\text{TF-IDF}(q, p_k) = \sum_{t \in q} \text{tf-idf}_{t,p_k} \quad (1)$$

where t denotes each term (word) in the query, TF is the number of occurrences of the term in p_k, and IDF is the log inverse of the number of page elements that contain the term t. The IDF factor is especially important for our task, since many terms on a given page will have little or no semantic discriminating power. For example, anchor text from links on a restaurant web page are likely to have the term *restaurant* in almost every link.

4.2. Gesture Intent

As with lexical intent, we seek a measure to capture the simultaneous voice and visual gesture intent of the user. For both speech recognition and hand movement detection, we use the popular and low-cost sensor Kinect. Kinect is a microphone array and a skeletal tracking motion sensing input device by Microsoft for the Xbox video game console and Windows PCs. The sensor has adequate resolution and software to accurately track hand movements. However, additional processing is required to discriminate intentional hand gestures such as pointing from unintentional hand movements.

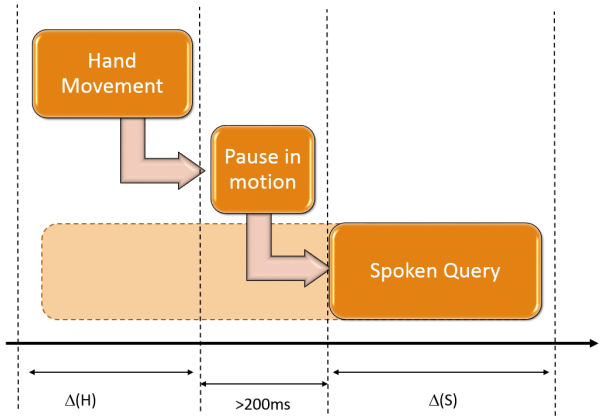

Figure 2: *Pointing gesture intent model*

For this work, we employ a simple model of pointing intent; a sequence starting with the hand motion, a brief pause with the hand still, followed by a spoken query. Figure 2 represents this model. Typically (as with Kinect) the hand gesture controls a cursor. The simplest method to determine the intended object selection is to compute the shortest straight line distance from the cursor to the (bounding box around) the page element. To decrease the chance of false positives, a *gesture focus region* may also be used. Gesture focus regions can be implemented with a weighting function, typically based on the inverse distance of the cursor to the object. Figure 3 shows the family of exponential inverse distance weighting (IDW) functions used in the experiments

$$\text{Gesture Score} = \exp\left(\frac{-|d|^a}{10^b}\right). \quad (2)$$

The IDW is used to specify the region of focus around the gesture's cursor. The goal of the IDW is to help balance the precision-recall of the gesture detection: a narrower region around the cursor (e.g., $a = 2, b = 0$) decreases the false alarms by reducing the affect of nearby objects, while a wider region ($a = 1, b = 1$) decreases the chance of incorrectly missing the intended object. The distance is measured in pixels from the gesture cursor to the object on the screen (e.g., web link, drop-down menu, form region). The IDW functions map the distances to $[0, 1]$, with a distance of 0 (the user is pointing directly at the object of interest) mapped to the maximum weight of 1.

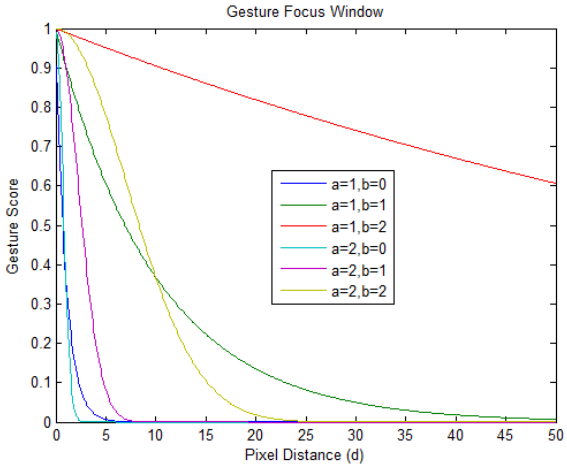

Figure 3: *Gesture focus windows around the cursor.*

4.3. Combining Intents

To form a single multimodal score for the kth page element, S_{M_k}, we use linear interpolation to combine the lexical score and the gesture score

$$\begin{aligned} S_{M_k} &= (1-\alpha)\ \text{TF-IDF}(q, p_k) + \alpha \cdot \text{Gesture Score} \\ &= (1-\alpha) \sum_{t \in q} \text{tf-idf}_{t,p_k} + \alpha \cdot \exp\left(\frac{-|d|^a}{10^b}\right) \end{aligned} \quad (3)$$

The values for α, a, and b are determined experimentally. Once the multimodal intent score is computed, it can be used for detection of intent by thresholding the score

$$\Lambda(S_{M_k}) \gtrless_{reject}^{accept} \theta. \quad (4)$$

The threshold θ can be optimized to minimize the cost of the error types: *false accept*, where the system incorrectly detected the presence of a user intent, and *miss*, where the system failed to detect the intent of the user.

5. Experiments and Results

5.1. Data Sets

We collected two data sets from 8 speakers during 25 sessions (set 1) and 7 speakers during 14 sessions (set 2). Both collections were performed in a living room set up, where users were seated on a couch approximately 5-6 feet away from a television screen. At the beginning of their first session, users were shown a short tutorial video demonstrating how the system can

be used, and were asked to improvise open-domain usage scenarios. In the first collection, the tutorial video included example usage scenarios with explicit (voice) clicks (as defined in Section 2) whereas in the second collection, the tutorial video included examples of using gesture with speech to click on a link. In both collections, users searched and browsed the web over open-domain tasks (e.g., shopping).

The total number of user turns in the first collection is 2,868, and 917 (31.9%) of these have a click intent. The second set includes 1,101 user turns, and 284 (25.8%) of these have a click intent. For the second set, we also computed the number of different types of clicks: 87.3% of the clicks are explicit clicks, 1.1% are location referrals and 11.6% include combined gesture and speech.

While hand pointing gestures were used for "click" intents, the collected data also includes cases of false gestures and false alarms by the system, such as a user lifting their arm to reach something on the coffee table. Hence, we further analyzed the usage of multi-modal input on a subset of the second collection. First, we separated out all utterances that control the display, such as "scroll down", as these can be captured with a high precision using the user's spoken utterances. Table 1 summarizes this analysis. 558 user turns are split into two: ones that are accompanied with a hand gesture and no gesture. The intent of these turns are categorized into click intents and non-click intents. In this analysis, we merged the gesture and speech clicks with location clicks as the second group is very infrequent, and named them "Click other". In this data subset, 22.8% of user utterances did not have a click intent, and yet a gesture was captured falsely. Similarly, 18.3% of the click utterances (excluding the explicit clicks) did not include a pointing gesture.

Table 1: *Statistics of user turns with/without hand gestures.*

	Gesture Found	No Gesture Found	TOTAL
Click "that one"	15 (2.7%)	1 (0.1%)	16 (2.8%)
Click other	25 (4.5%)	102 (18.3%)	127 (22.8%)
Non-Click	127 (22.8%)	288 (51.6%)	415 (74.4%)
TOTAL	167 (30.0%)	391 (70.0%)	558

5.2. Results

To examine the effectiveness of contextual adaptation for ASR, we used 2,868 utterances (9,346 words) from the first collection and completed tests on statistical language model (SLM) adaptation. While the average utterance length in this set looks short (3.3 words), this is mainly because this set contains all user turns in a session, including commands to change the display, which are usually 1-2 words (such as "scroll down" and "back"). About 40% of the utterances are such commands, 32% are click utterances, and 28% are the rest.

Table 2 shows the word error rate (WER) results from these experiments, where we compare a generic large vocabulary 400K word conversational speech recognition language model (LVCSR-LM) with its dynamically adapted version. The table also includes an analysis of impact on performance of click (32%) and non-click (68%) utterance subsets. Overall, with an out-of-vocabulary (OOV) rate of 0.25% and adapting the language models to the visual context improved the WER of the LVCSR-LM from 20.6% to 19.2%. WER for the context-related click utterance subset improved from 28.2% to 23.7% (a relative improvement of 16%), without a degradation on the performance of the rest of the turns. The small improvement (from 17.4% to 17.1%) on the non-click turns can be partially due to domain adaptation as a side effect of adapting to the visual content.

Table 2: *ASR WER with contextual adaptation.*

LM	WER overall	WER Click subset	WER Non-Click subset
LVCSR-LM	20.6	28.2	17.4
LVCSR-LM + adaptation	19.2	23.7	17.1

To study the effects of the gesture intent signal *independent* of how often it is used and the quality of the gesture detector, we complete simulations where all components/measures are real *except* the gesture. Table 3 shows results on a held-out random sample of 75% of the turns in data sets 1 and 2. The table shows the probability of the error types (false accept and miss) using the multimodal score and the intent detector of Equations 3-4. Results are computed for both manual transcription of the speech and automatic speech recognition using the contextual adaptation. The parameters of the detector are varied to show the affects of the size and shape of the gesture focus window (a and b) and the interpolation weight (α) between lexical and gesture-based intent. Since we normalized the scores of the lexical and gesture intent detectors to be [0-1], α can be interpreted as the relative importance of the gesture score in the combination.

For these experiments, we also simulated the human user's gesture intent to control for gesture precision. The simulation places the gesture cursor on an equidistant curve from the intended page element (link). The gesture precision distance, R, is the number of pixels that the cursor is away from the desired object (e.g., web link) and the page. We simulated gestures for two different gesture precisions: $R = 0$ and $R = 20$ pixels. The probability of missing the multimodal intent, P_{miss}, is computed in the operating region where the probability of falsely detecting an intent is low ($P_{fa} = 1\%$). We focus on this operating region due to the sensitivity of users to false positives and the objectionable user experience of the system incorrectly taking actions (clicking).

The best performing multimodal intent detector uses a balanced blend of lexical and gesture ($\alpha = 0.5$) and a broad gesture focus window ($a = 1$,$b = 2$). At these settings, with perfect speech recognition, perfect gesture precision ($R = 0$), and the user gesturing towards the intended page element (link) for 100% of the trials, the P_{miss}(@P_{fa}=1%) $= 8.1\%$. This represents an upper bound on the performance and is a 68.2% error rate reduction (ERR) compared to the single mode lexical intent detector ('No Gesture") with P_{miss}(@P_{fa}=1%) $= 25.5\%$. With the same settings for the gesture focus window, with automatic speech recognition (ASR), and with a gesture precision of $R = 20$, the P_{miss}(@P_{fa}=1%) $= 16.9\%$. This is a 50.3% error rate reduction (ERR) over lexical intent alone.

Using the same development data set used to compute the results in Table 3, we conducted experiments with real human gestures and an automated gesture detection model. For this test, we extracted gesture positions from the data logs that were generated using the model shown in Figure 2. We examined two cases: (1) all logged gestures and (2) only gestures where the user also said "that one". The first case includes all cases where a gesture was detected, which is approximately 30.0% of the test cases (see Table 1). The second case was used to isolate human gesture precision from the errors introduced by the gesture detection model.

Table 3: *Summary of multi-modal intent detection with simulated gestures.*

				$R=0$		$R=20$	
	IDW a	IDW b	Gesture α	$P_{miss}@P_{fa}$=1% Manual	$P_{miss}@P_{fa}$=1% ASR	$P_{miss}@P_{fa}$=1% Manual	$P_{miss}@P_{fa}$=1% ASR
No Gesture	-	-	-	-	-	25.5%	34.0%
Gesture	1	0	0.25	10.2%	21.1%	25.5%	34.0%
	1	0	0.50	9.5%	19.4%	25.5%	34.0%
	1	0	0.75	10.0%	19.9%	25.5%	34.0%
	1	1	0.25	10.2%	21.1%	21.5%	31.5%
	1	1	0.50	8.3%	18.3%	16.9%	27.1%
	1	1	0.75	8.6%	18.5%	8.3%	20.1%
	1	2	0.25	10.9%	21.3%	12.3%	23.1%
	1	**2**	**0.50**	**8.1%**	**18.3%**	**7.2%**	**16.9%**
	1	2	0.75	8.3%	18.3%	7.6%	17.1%
	2	0	0.25	10.4%	21.3%	25.5%	34.0%
	2	0	0.50	9.5%	19.4%	25.5%	34.0%
	2	0	0.75	10.0%	19.9%	25.5%	34.0%
	2	1	0.25	10.2%	21.1%	25.5%	34.0%
	2	1	0.50	8.3%	18.3%	25.5%	34.0%
	2	1	0.75	10.0%	19.9%	25.5%	34.0%
	2	2	0.25	10.2%	21.1%	25.0%	33.3%
	2	2	0.50	8.3%	18.3%	23.4%	32.2%
	2	2	0.75	8.3%	18.3%	20.6%	30.8%

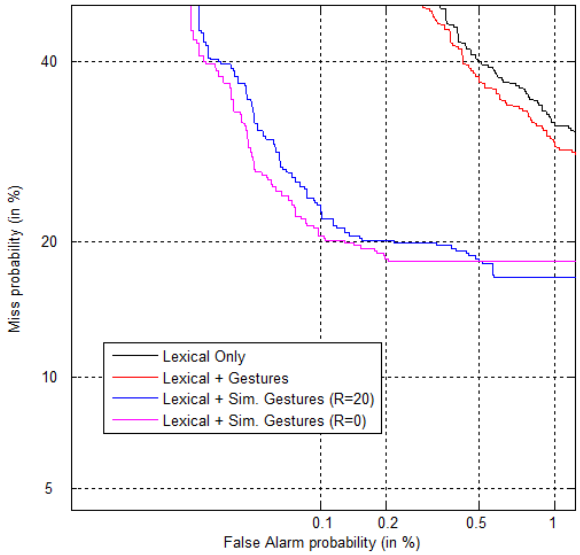

Figure 4: *DET results for multi-modal intent detection*

The results are shown in Table 4. With the introduction of errors due to both human gesture precision and the gesture detection model, the performance over all the trials was $P_{miss}(@P_{fa}$=1%$) = 22.9\%$ (ERR=10.2%) and 31.7% (ERR=6.8%) for manual transcriptions and ASR, respectively. For the case where the user clearly indicated the pointing intention with the phrase "that one" while gesturing, the $P_{miss}(@P_{fa}$=1%$) = 15.4\%$ for both manual and ASR (perfect recognition of the phrase), which is ERR=39.6% and ERR=54.7%, respectively. For this second case, we computed the average gesture precision for humans. Referring to Table 3, the gesture precision (R) for humans was is in the range of 16.4 to 28.6 pixels, depending on the density of the visual content on the screen. In other words, humans are able to precisely gesture towards the intended element. The drop in performance, therefore, was a result of (1) humans only gesturing toward the intended page element 30.0% of the time (see Table 1) and (2) errors in the gesture detection model (see Figure 2).

Figure 4 summarizes the performance of the same experiment for ASR and compares human performance to the upper bound with perfect gesture detection and 100% user participation in gesturing towards the intended page element (link) when speaking. The top two curves show the performance of the real multimodal lexical+gesture detector compared to the baseline (lexical only).

Table 4: *Multi-modal intent detection with real gestures.*

	IDF a	IDF b	Gesture α	$P_{miss}@P_{fa}$=1% Manual	$P_{miss}@P_{fa}$=1% ASR
No Gesture	-	-	-	25.5%	34.0%
All	1	2	0.50	22.9%	31.7%
Gestures + "that one"	1	2	0.50	15.4%	15.4%

6. Conclusions

This paper described the development of a multi-modal dialog system for conversational web search and internet browsing. The work focused on two novel components: dynamic contextual adaptation of speech recognition and spoken language understanding models using multi-modal conversational context, and fusion of users' multi-modal speech and gesture inputs for understanding their intents and associated arguments. The system was evaluated in a living room setup with live test subjects on a real-time implementation of the multimodal dialog system. Results showed a 16% error rate reduction (ERR) for contextual ASR adaptation to clickable web page content, and 7-10% ERR when using gestures with speech. Analysis of the results showed that when users clearly and always indicate pointing intent while simultaneously using voice, the combination of modalities yields a 54.7% ERR over lexical features. While we observed users only point with hand gesture 30% of the time, the result suggests that other, more persistent modalities (e.g., eye gaze) could be used to yield substantial gains over speech alone.

7. Acknowledgements

The authors would like to thank Malcolm Slaney for helpful discussions related to this work.

8. References

[1] G. Tur and R. De Mori, Eds., *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, John Wiley and Sons, New York, NY, 2011.

[2] R. A. Bolt, "Put-that-there: Voice and gesture at the graphics interface," *Computer Graphics*, vol. 14, no. 3, pp. 262, 1980.

[3] G. Taylor, R. Frederiksen, J. Crossman, J. Voigt, and K. Aron, "A smart interaction device for multi-modal human-robot dialogue," *Ann Arbor*, pp. 190–191, 2012.

[4] R. Balchandran, M.E. Epstein, G. Potamianos, and L. Seredi, "A multi-modal spoken dialog system for interactive tv," in *Proceedings of the 10th international conference on Multimodal interfaces*, 2008, pp. 191–192.

[5] A. Ibrahim and P. Johansson, "Multimodal dialogue systems for interactive tv applications," in *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces*, 2002, pp. 117–122.

[6] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor, "MATCH: an architecture for multimodal dialogue systems," in *Proceedings of the ACL*, Philadelphia, PA, July 2002.

[7] S. Oviatt, P. Cohen, L. Wu, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, and D. Ferro, "Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions," *Human-computer interaction*, vol. 15, no. 4, pp. 263–322, 2000.

[8] R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel, "Natural human-robot interaction using speech, head pose and gestures," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004, pp. 2422–2427.

[9] L. Heck, "The conversational web," in *IEEE Spoken Language Technology Workshop(SLT), Keynote*, 2012.

[10] Wikipedia The Free Encyclopedia, "Conversation," http://en.wikipedia.org/wiki/Conversation.

[11] D. Hakkani-Tür, G. Tur, L. Heck, A. Celikyilmaz, A. Fidler, D. Hillard, R. Iyer, and S. Parthasarathy, "Employing web search query click logs for multi-domain spoken language understanding," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 419–424.

[12] D. Hakkani-Tür, G. Tur, L. Heck, and E. Shriberg, "Bootstrapping domain detection using query click logs for new domains," in *Interspeech*, 2011.

[13] G. Tur, D. Hakkani-Tür, D. Hillard, and A. Celikyilmaz, "Towards unsupervised spoken language understanding: Exploiting query click logs for slot filling," in *Interspeech*, 2011.

[14] D. Hakkani-Tür, L. Heck, and G. Tur, "Exploiting query click logs for utterance domain detection in spoken language understanding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5636–5639.

[15] D. Hakkani-Tür, G. Tur, R. Iyer, and L. Heck, "Translating natural language utterances to search queries for slu domain detection using query click logs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4953–4956.

[16] G. Tur, D. Hakkani-Tür, L. Heck, and S. Parthasarathy, "Sentence simplification for spoken language understanding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[17] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and L. Heck, "Learning when to listen: Detecting system-addressed speech in human-human-computer dialog," in *Interspeech*, 2012.

[18] G. Tur, Minwoo Jeong, Ye-Yi Wang, D. Hakkani-Tür, and L. Heck, "Exploiting semantic web for unsupervised statistical natural language semantic parsing," in *Proceedings of Interspeech*, 2012.

[19] L. Heck and D. Hakkani-Tür, "Exploiting the semantic web for unsupervised spoken language understanding," in *Proceedings of the IEEE SLT Workshop*, Miami, FL, 2012.

[20] D. Hakkani-Tür, L. Heck, and G. Tur, "Using a knowledge graph and query click logs for unsupervised learning of relation detection," in *Proceedings of the ICASSP*, 2013.

[21] L. Heck, D. Hakkani-Tür, and G. Tur, "Leveraging knowledge graphs for web-scale unsupervised semantic parsing," in *Interspeech*, 2013.

[22] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[23] C.D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.