The Prospects for Unrestricted Speech Input for TV Content Search

Kent Wittenburg, Tom Lanning, Derek Schwenke, Hal Shubin, Anthony Vetro

TR2006-045 May 2006

Abstract

The need for effective search for television content is growing as the number of choices for TV viewing and/or recording explodes. In this paper we describe a preliminary prototype of a multimodal Speech-In List-Out (SILO) interface in which users' input is unrestricted by vocabulary or grammar. We report on usability testing with a sample of six users. The prototype enables search through video content metadata download from an electronic program guide (EPG) service. Our setup for testing included adding a microphone to a TV remote control and running an application on a PC whose visual interface was displayed on a TV.

ACM Advanced Visual Interfaces, May 2006

The Prospects for Unrestricted Speech Input for TV Content Search

Kent Wittenburg*, Tom Lanning*, Derek Schwenke*, Hal Shubin**, Anthony Vetro*

*Mitsubishi Electric Research Laboratories
201 Broadway
Cambridge, MA 02139 USA
{lanning,schwenke,vetro,wittenburg,woelfel}@merl.com

**Interaction Design, Inc.
http://www.user.com
hal@user.com

ABSTRACT

The need for effective search for television content is growing as the number of choices for TV viewing and/or recording explodes. In this paper we describe a preliminary prototype of a multimodal Speech-In List-Out (SILO) interface in which users' input is unrestricted by vocabulary or grammar. We report on usability testing with a sample of six users. The prototype enables search through video content metadata downloaded from an electronic program guide (EPG) service. Our setup for testing included adding a microphone to a TV remote control and running an application on a PC whose visual interface was displayed on a TV.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: Voice I/O.

General Terms

Algorithms, Design, Human Factors.

Keywords

Television interfaces, multi-modal interfaces, speech interfaces, information retrieval, electronic program guides.

1. INTRODUCTION

Today there is an explosion of video content served to consumers world-wide. Recent industry developments such as the sale of videos through Apple’s iTunes stores for the Video iPod (http://www.apple.com/itunes/videos/) and other Web-based services such as Blinx (http://www.blikx.tv), Google (http://video.google.com), and Yahoo (http://tv.yahoo.com) are evidence the size of this new content base. Internet distribution channels for video that would stream or download directly to TV sets or to Digital Video Recorders are also beginning to appear, e.g., the Netflix/Tivo partership [9]. One proposed solution to the problem of finding video content is personalized recommendation. See, e.g., [8] and other papers in a special issue of *User Modeling and User Adaptation*. A more straightforward solution might be to support search. However, while PC-based Web browsers can provide a reasonable interface for searching for video programming via text entry on keyboards, there is no satisfactory solution currently for standard TV remote controls. Text entry is at best awkward with remote controls that lack both a mouse and a keyboard. Even now, it can be frustrating and time-consuming for users to locate TV shows to record when there might be hundreds of channels and up to two weeks of programming available on an Electronic Program Guide (EPG). Our proposed solution is to add a microphone to remote controls that would enable voice input for searching over ever-growing collections of content available through EPGs. Our approach is to use SpokenQuery technology in a Speech-In List-Out (SILO) interface [3], which allows search terms to be entered that are unrestricted by vocabulary or grammar. The system responds with the best matches it can find even though the speech itself remains ambiguous to the system. At this point little is known about how to design such interfaces for the TV domain and what their prospects for success might be. The research reported on here is a preliminary step towards answering these questions.

In the next section we will discuss related work and how our proposal differs from prior research in speech input for TV interfaces. We follow with a characterization of our prototype and discuss a number of the design decisions that we were forced to make to realize a preliminary system. Then we describe a set of usability experiments, which were conducted over two days with six subjects who were asked to perform tasks associated with video content search. Finally, we conclude with some lessons learned, suggested improvements, and an outlook for the future.

2. RELATED WORK

As with prior work in speech interfaces generally, there have been two basic kinds of proposals for using speech input with TVs. The first is to use speech in order to specify a limited set of commands [5][11]; the second is to develop dialog-based systems that purport to handle errors more gracefully and guide users into using speech that can be understood by the system [6][10]. A problem with the first type of speech interface is that users must learn what they can say in order to be understood and avoid frustration. However, even when speech commands have been more efficient, they have not necessarily led to preference over remote control button interaction [5]. Also, error correction in some form seems to be a requirement [1]. The biggest issue with the second type of interface is the cost and complexity of design and development. Also, it is not

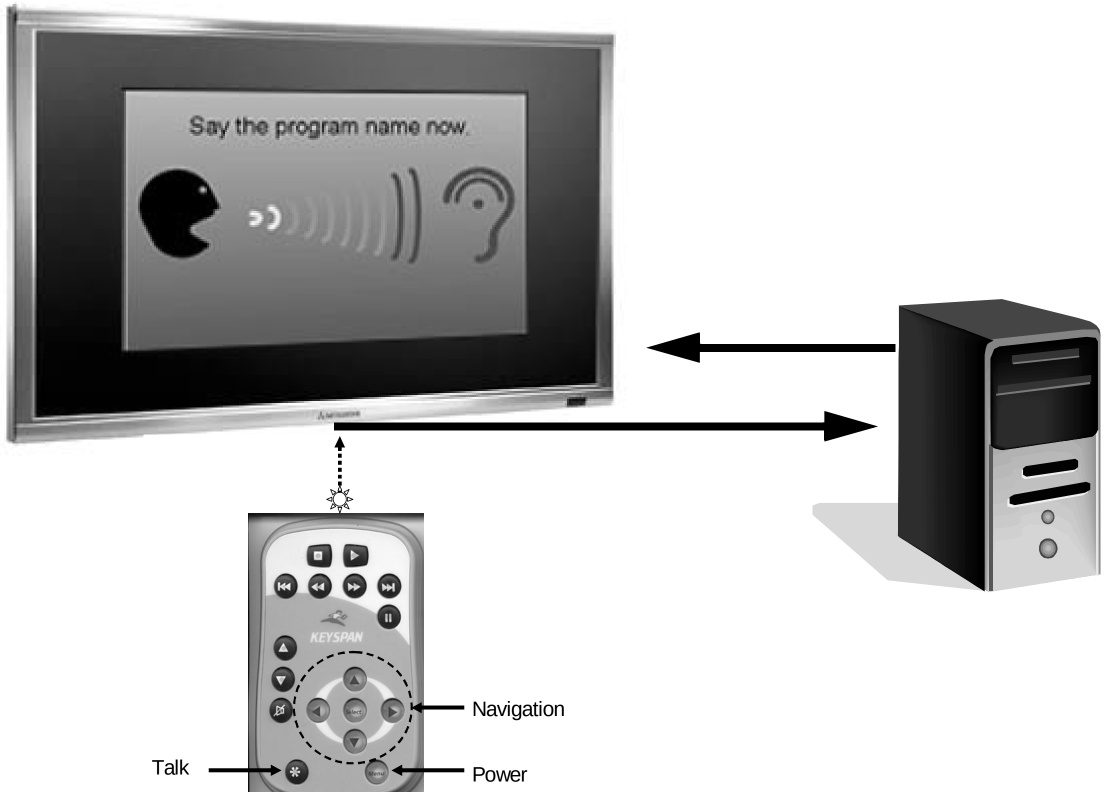


Figure 1: Setup with remote control with push-and-release-to-talk button and visual feedback for audio level on TV screen.

clear that a conversational style interface is suitable for interaction through a remote control with a TV where the current generation expects instant response.

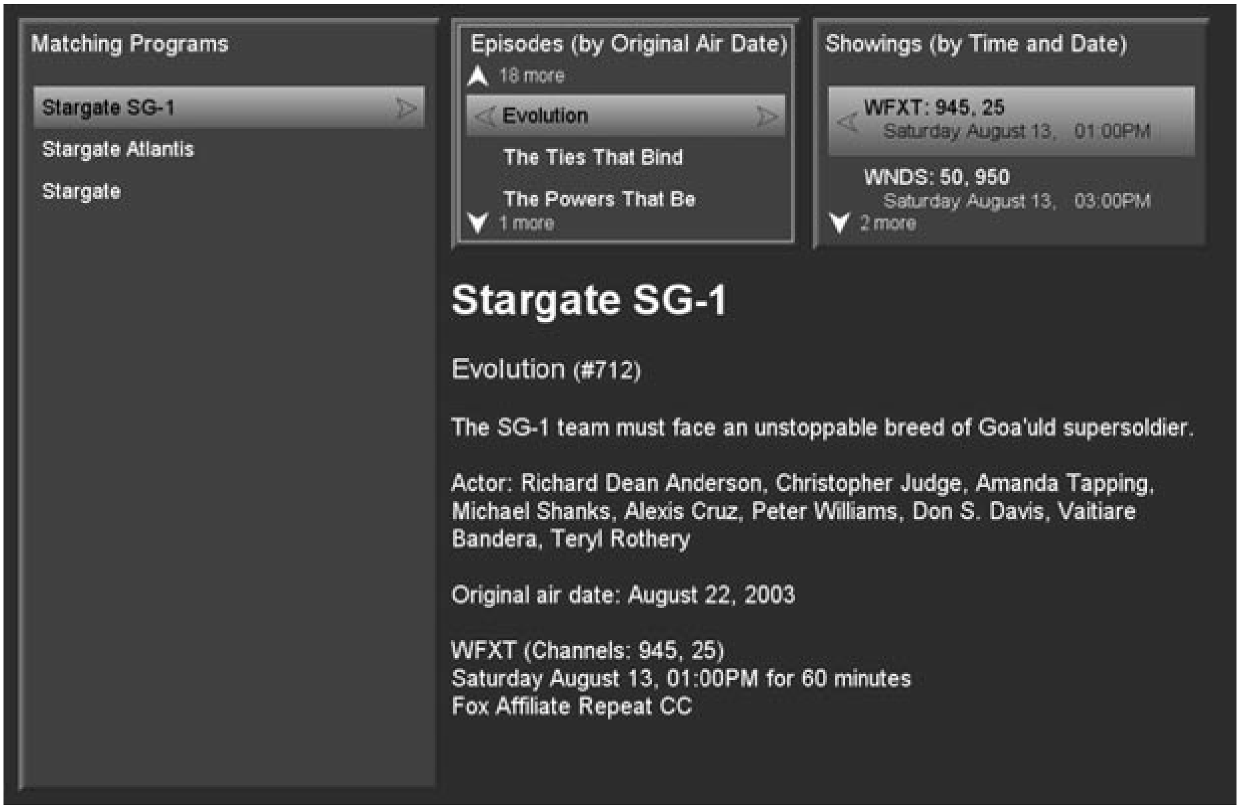
A new model for application of speech to interfaces is the Speech-In List-Out paradigm [3] based on SpokenQuery technology described in [12][13]. The basic concept is to utilize the output of a speech recognition engine not as a full specification of a text query, but rather as a set of words and/or bigrams with probabilities that can be used to match against the indexed target set. Conceptually, this style of interface would appear to a user as a sort of spoken version of Google. However, a significant difference is that the system cannot easily display back to the user what it has "understood" (as a text input box does) except in the form of a list of ranked matches against the target set. Instead of taking the "best guess" of the speech engine as the query, it computes a vector of all possible words and/or bigrams that the speech engine determines might have been said and uses that structure as the query. Our hypothesis is that such a system can avoid the problems of speech error recovery by eliminating the need for fully disambiguating the spoken input. However, its success ultimately relies on the accuracy of its retrieval performance.

A number of prototypes have been built to exhibit SpokenQuery technology including document retrieval with cell phones, point-of-interest search for navigation systems, and music search using car audio systems. An experiment in [4] showed that subjects peformed better on a simulated driving task while searching for music with the SILO system than with the standard GUI button-based interface. Retrieval performance for music collections and for these other application domains has been promising. Some test results were reported in [3] and [13] as well as anecdotally in [4]. This present paper is the first report on usability studies that we are aware of, and it also is the first to consider application of SpokenQuery in a SILO model in the TV content domain whose information structure is more complicated than, say, music collections.

3. PROTOTYPE

The prototype we built for the purposes of this study was a limited emulation of an interface for a television with an embedded personal video recorder (PVR). It allowed the viewer to find and schedule the recording of a program using a SILO design. The prototype software ran on a PC that was connected to a high-definition television (720p) and controlled by a remote with an embedded microphone. The prototype used actual EPG data for a two week period extending into the future from the date of the testing. Although we did not allow the subjects to restrict the program information to their own channels or programs they had at home, the information was actually what they would receive in the Boston area should they have one of the higher-end service plans for cable or satellite (hundreds of channels).

Our software suite made use of the public domain Sphinx 3 speech engine [2] as well as MERL's SpokenQuery modules. The application and user interface was written in Java and ran on a Windows XP personal computer.



Matching Programs

Stargate SG-1

Stargate Atlantis

Stargate

Episodes (by Original Air Date)

18 more

Evolution

The Ties That Bind

The Powers That Be

1 more

Showings (by Time and Date)

WFXT: 945, 25
Saturday August 13, 01:00PM

WNDS: 50, 950
Saturday August 13, 03:00PM

2 more

Stargate SG-1

Evolution (#712)

The SG-1 team must face an unstoppable breed of Goa'uld supersoldier.

Actor: Richard Dean Anderson, Christopher Judge, Amanda Tapping, Michael Shanks, Alexis Cruz, Peter Williams, Don S. Davis, Vaitiare Bandera, Teryl Rothery

Original air date: August 22, 2003

WFXT (Channels: 945, 25)
Saturday August 13, 01:00PM for 60 minutes
Fox Affiliate Repeat CC

Figure 2: View of TV screen after search results are returned. Left column is matching program/series titles, the first of which is selected. Middle column is list of episodes. Third colum is list of showings (time/channel).

3.1 Interaction

As shown in Figure 1, the remote control had a button to trigger listening, four buttons (up, down, left, and right) for cursor movement, and a button for selection in the middle. Other than the power button, these were all we used for the experiment. A receiver mounted under the television converted button presses on the remote to simulated key presses on the PC's keyboard. The remote contained an embedded microphone that was connected to the audio input on the PC.

The television screen started with a blank screen that was tuned to the programming available on channel 2 when the viewer pressed the power button. The viewer could start a search by pressing the listen button at any time, except when the television was already listening. When the viewer pressed the listen button, the prototype displayed a real-time audio level meter and played a short audio prompt tone as shown in Figure 1. The viewer would then speak terms for the search. Once the system had determined the viewer had finished talking, the prototype emitted an end-beep and displayed the results, an example of which is shown in Figure 2.

As indicated, the results display was composed of three lists and one detailed view. The lists, from left to right, were (1) a list of series (labeled “Matching Programs”) that matched the spoken query, (2) a list of episodes for the series selected in the first list, and (3) a list of showings for the episode selected in the second list. The detail view contained the combined information from the selected series, selected episode, and selected showing. The prototype results screen always had a selected series, episode, and showing unless there were no matches.

The list of matching series, labeled "Matching Programs", was rank-ordered by best match to the query. The list of episodes was ordered by the original air date, and the list of showings was ordered by the time and date of the showing.

If the viewer pressed the select button, the PVR would simulate the recording of the program, and then the television returned to the programming on channel 2. The viewer could also use the cursor buttons to move up and down or left and right through the lists to select the correct showing.

3.2 Information Model

The primary entities in our EPG database were series, episodes, showings, stations, and channels. An example of a series would be "Stargate SG-1". An episode in that series, such as "Evolution", might have 4 showings on 2 stations. One of those stations might have an analog channel, "25" and also a digital channel, "25-1". This model is close to existing open standards and the EPG data we received from a commercial service (Tribune Media Services). We defined a hierarchical structure where series contained episodes and episodes contained showings. Stations were associated with channels.

A key question for us was "What entity should viewers search for?" We knew ultimately that a viewer had to select a particular showing to complete the task of scheduling the personal video recorder. However, as we will explain, this end goal did not necessarily imply that viewers needed to search on showings directly. Our SILO multimodal interface paradigm allowed for designs in which the viewer could speak a phrase and then browse intermediate results listings using the buttons on the remote control.

Our design process thus began with identifying the entities that the user would be able to search for and then creating a written and spoken index for the words associated with those entities. A quick analysis of the data for a two week period showed there were ~123,000 different showings of ~24,000 episodes. After creating pseudo series for each program that was not part of any series, the analysis revealed ~7,500 unique series. This implied our two-week dataset would have approximately 155,000 items.

A small study and several interviews indicated to us that indexing on showings was a problem. In the first place, retrieval performance was in part determined by the size of the database. A reduction in the database size would tend to increase accuracy of matching spoken queries to results. As it turned out, showings were not easily distinguishable by query terms that a user might use. The ranking of showings, particularly if they were from the same episode, was also a problem. But the most important factor was our conclusion that viewers would not actually want to search on showings. There are just too many of them. It made more sense to consider searching on episodes.

Thus our next choice was to consider indexing and searching on episodes. Aggregating showings into their common parent episodes reduced the number of items from ~125,000 to ~25,000. However, there were still issues. At first we simply concatenated the text of each attribute for the episode and its parent (pseudo)series to create the index. However, adding all the words contained in the episode’s actors, description, directors, genre, name, and ratings attributes did not produce the results we desired. Searches for the series entitled "Lost" were literally lost within the hundreds of other episodes that contained the word "lost" in their description. We incorporated entity and attribute weighting to tune the rankings but that only slightly improved the results. In fact, by spending considerable time with the data, we realized that more often than not, viewers would not know the words in the episode names in any event. Series names such as “Lost” or “NFL Football” seemed to us a better bet.

Therefore our next move was to narrow the choice set even more by limiting searches to only (psuedo)series, which we labeled as “Matching Programs” in the interface. The indexed database now had approximately 7,500 items. We specified the language representation of a series to be the name of the series and then added in all stations or networks that carried the series. The viewer would now need to use the remote control to manipulate an on-screen browser to select among the matching series and then select the desired episode and showing from there. We described the feature as “Program Title Search” for the purposes of this experiment. We decided to proceed with usability testing in order to determine whether program title search might be a feature that appealed to viewers.

4. USABILITY TESTING

User studies were conducted at MERL following the principals of usability engineering [7]. A consultant was engaged to help design and carry out the study. During the user sessions, the consultant asked participants to find TV programs to watch or record. The purpose was to identify strengths and weaknesses of the system and to offer recommendations for improving it.

4.1 Study setup

Participants sat on a couch in a lab at MERL, watching a large-screen TV. They used either a wired microphone glued to a small remote or a separate clip-on wired mic with the same remote pictured in Figure 1.

The research team made changes during the study. Removing less likely items from the data made the results more relevant. Introducing a delay into audio capture made recognition better (although it was still a problem). Removing network names from the data set may have made recognition better but interfered with how some participants interacted with the system. These changes made the system easier to use, but did not affect the outcome.

The consultant sat with participants individually to facilitate the study. He administered an initial questionnaire on background information and then gave each person a series of tasks. He asked follow-up questions after each session. Members of the MERL research team observed remotely through a video hookup.

The consultant employed a think-aloud protocol, where he interacted to introduce new tasks and follow up on interesting points, but he did not necessarily answer all of the participants' questions. He also tried to keep them comfortable when recognition wasn't working well.

4.2 Participants

We used an outside firm to recruit participants. They recruited eight people based on our screening questionnaire, but we had two no-shows. Our goal was to find participants that were likely to be in the target market for high-end televisions that might include a spoken query feature such as we were testing. We required a minimum income, and we wanted participants who spoke English clearly with no noticeable accent or speech impediments. (One participant had a noticeable local accent and one had a very strong one.) The backgrounds of the six participants in our study are partially summarized in Table 1. Some quotations from participants are included in what follows, referenced according to Table 1. (P1 and P5 were no-shows.)

4.3 Tasks

Some tasks were communicated only verbally while others were given in written form. Some tasks that came from printed listings were very specific. For example, the consultant showed participants a listing from the newspaper that had a program circled and asked them to find the program:

Other tasks simulated viewing suggestions from friends, from memory or based on interest. For example:

- "Someone told you about a program about container ships on The Discovery Channel sometime this week. Can you find it so you can either watch or record it (depending on the schedule)?" In this case, "container ships" was in the title of the program, but not in the description.

- "See if there are any programs this week about cooking turkeys for Thanksgiving." This task was very open ended, although there were a number of relevant programs.

Table 1: Information about participants. Recruiting requirements included a minimum income and some experience with state-of-the-art TV services.

Ref	Sex	Age	Income	TV & accessories	TV service	Three favorite shows	Activities
P2	F	48	50K+	Smaller TV	Cable/Comcast	Sitcoms, Old Movies, musicals	Set up device, browsed
P3	M	38	50K+	Smaller TV,Tivo	Cable/ RCN	News, TBS, Discovery & History	Set up device, browsed
P4	F	33	50K+	Smaller TV,Tivo	Cable/Comcast	Lost, The Apprentice, Desperate Housewives	Set up device, browsed
P6	M	32	50K+	Large Screen	Cable/Comcast	Sports, comedy, History Channel	Set up device
P7	M	38	50K+	Large screen	Cable/Comcast	Lost, The Tonight Show, Animal Kingdom	Set up device, browsed
P8	F	36	50K+	Larg screen, Tivo, smaller TV	Cable/Comcast	Fox, reality shows, drama series	Set up device, browsed

- "Remember the episode of M*A*S*H where everyone's afraid that Captain Pierce was killed at the front? See if it's on this week." This task was presented either orally or on paper. It required participants to find the program M*A*S*H and then find an episode that was similar to the description, but used different words; this simulated a friend's recommendation or a dim recollection.

- "The FX program, Cops, features police officers in different cities. You used to live in Seattle--do they ever show that city's cops?" The word "Seattle" was in each of two episode names in the result set, but only in the description of one of them.

The consultant also allowed participants to search for things that they were interested in.

4.4 Significant findings

When it worked, test participants enjoyed using the system. Successful interactions were very quick. If the right program was first on the list (and therefore highlighted for action), participants frequently hit the Select button without looking at the other columns.

P4: "[It's] easy to get around"

P6: "This is a pretty neat device here… sophisticated."

In general, participants were comfortable with using the microphone and the remote control. Many moved the remote to their mouths to talk. Others kept it steady and dipped their head towards the control when the consultant pointed out that moving the mic caused noise. It appeared that viewers would be able to learn appropriate usage with appropriate audio level feedback. However, P3 was not sure if the Talk button was a press-and-hold or press-and-release operation.

However, most of the participants picked up the remote control, pressed the Talk button and then paused, thinking about what to say. This may be partly due to the testing situation, but it may be a natural response as well. It would be best to automatically detect the onset of speech as well as the end of speech.

Participants expected voice recognition/retrieval to simply work. Their actions and comments indicated that they wanted to say something and see the right program listed first. Despite the fact that recognition/retrieval will never be at 100%, these participants nevertheless expected it.

P4: "I think if I said 'Friends' it would be right at the top [of the program list]."

P6: "If you give it a voice command it should give the result."

P6: “This is too much work for using a voice command. You don't want to go through this." [after repeated attempts]

When recognition/retrieval failed, recovery strategies often did not improve the result. Some of the strategies we noted follow. See also Table 2.

- Repeating the same utterance. Slight changes in inflection sometimes returned very different results.

- Using more or less of the title. Example: "Johnny Cash", then "I walk the line", then "Johnny Cash I walk the line".

- Changing inflection to a question, as if one were asking the recognizer rather than telling it what to do. P2 and P3 especially did this.

- Other changes to pronunciation. Example: When a search for "Friends" didn't work, P3 said "Friendzz".

- Going off into other dimensions. Example: Trying "Cops", then "Cops on FX", then "FX programming" -or- "cooking turkeys", then "cooking shows", then "The Cooking Channel", then "Thanksgiving", then "Martha Stewart".

- Adding detail. Example: When "Friends" didn't work, one participant tried "Friends baby shower", adding detail from the episode title or description.

- Saying. Individual. Words. Instead. Of. Continuous. Speech. To. Make. The. System. Understand. Better. Like. Talking. To. A. Small. Child. P3 and P4 especially did this.

- Some people spoke more slowly to try to help the system understand what they were saying. This may also model speaking to a young child.

Table 2: Examples of sequences of utterances by participants in the face of failed recognition/retrieval.

	Task: Find the Johnny Cash special	**Task 4: Cooking turkey for Thanksgiving.**	**Task 9: A specific Barney & Friends episode**	**Task 7: 7pm news on Channel 7**	**Task 10: A show about container ships on Discovery Channel.**	**Task 11: Specific episode of Cops**
P2			"Barney" "Children's programming" "PBS television?" "WGBH"	"WHDH" "7 o'clock news" "Local news programming" "NBC programming"	"Discovery channel"	"Seattle police "Cops" "Cops on FX" "Seattle cops" "Cops on FX" "FX programming
P3		"Cooking tips" "Thanksgiving cooking"		"Channel 7" "Channel 7" "Channel 7" "Channel 7" "Channel 7"	"Discovery Channel" "Discovery Channel container ships"	"Cops"
P4	"Johnny Cash" "I walk the line" "Johnny Cash I walk the line"	"Cooking turkeys" "Cooking show" "The Cooking Channel" "Thanksgiving" "Martha Stewart"			"Discovery Channel content for ships" [unclear]	"Cops"
P6	"I walk the line"		"Barney in concert" [a VHS title!] "Barney riding bikes"		"Discovery Channel cargo ships"	
P7	"I walk the line"	"Cooking channel" "Thanksgiving food" "Food channel" "Food channel" "Food"				"Cops cops cops" "Cops Seattle
P8	"Johnny Cash I walk the line"	"Cooking channel"				

- Some people moved the microphone closer to their mouths, or moved their mouths closer to the mic (despite instructions not to move the mic too much).

Participants did not simply talk louder, as the stereotype of talking to a non-native speaker might suggest.

Despite clear instructions and apparent understanding, participants did not confine their searches to program titles only. This deviation from the model was often apparent in recovery mode, but not only in this case. Examples follow.

- "New England Patriots." Language from episodes. The language in episodes is important particularly for sports programs, where titles are very general (NFL Football or MLB Baseball) and the important information is in the episode name (Dallas Cowboys at Philadelphia Eagles or World Series).

- A content-based search, matching a program description. Example: A recommendation from a friend to find "A show about brewing beer on the Discovery Channel." Or "The M*A*S*H episode where…"

- An ill-defined content-based search. Example: P4 wanted to find a program about investing with a host whose first name she remembered as "Jimmy". She tried "The Jimmy Show", "Jimmy investing".

- A category-based search, such "children's programming".

- "Food Channel." Restricting searches to channels or networks. One participant said that she only watches a few of the channels she has and would like to restrict searching to those channels.

As observers noted, "subjects see this as a general search tool. They expect to say date or channel or time or program name or whatever. Rankings need to be refined to make this work."

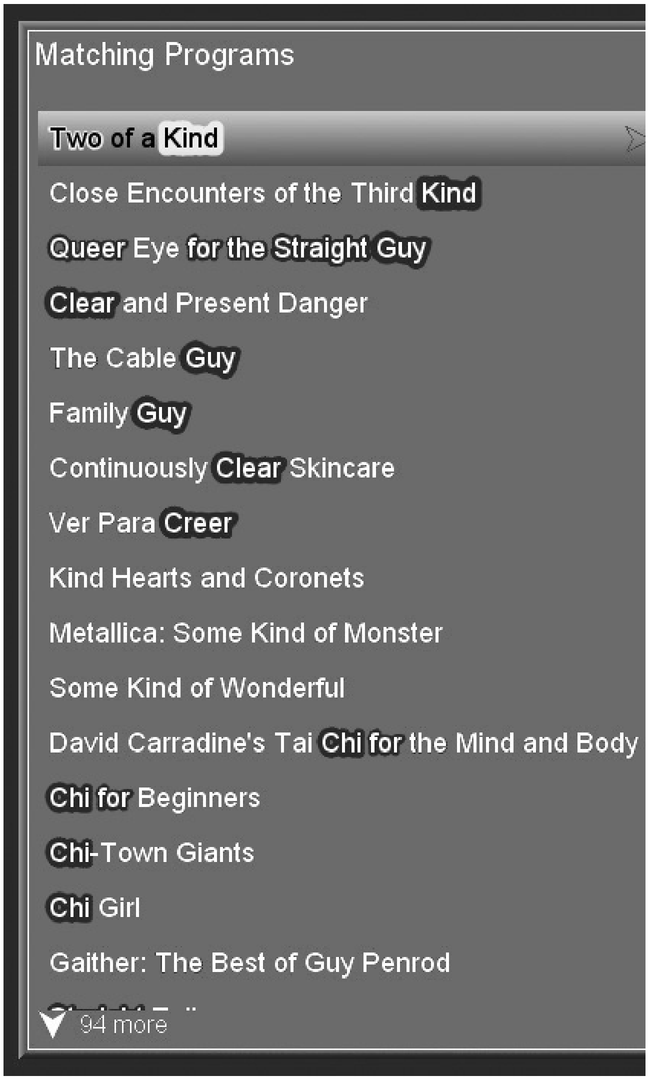
Users were not interested in searching through a list of program names but were happier to look through episodes. As with searching the Web, the first page of results is all that matters. (P4 scrolled through the entire list of 100 in her first task, but that seemed to be because of the testing situation.) In fact, we observed participants who didn't seem to accept a result if it wasn't the first item in the program list.

On the other hand, scrolling through episodes was fine, perhaps entertaining. That might have been because the list was typically shorter. It may also have been because it was more rewarding to look at episodes for the requested program. P7 appeared to enjoy reminiscing as he looked through episodes of a favorite program.

P2: "I don't want to scroll through 100 results."

P4: "I have to scroll through all 100 to see it?"

P6: "Having to search [browse the list] like this defeats the purpose [of the speech interface]."



Matching Programs

Two of a Kind
Close Encounters of the Third Kind
Queer Eye for the Straight Guy
Clear and Present Danger
The Cable Guy
Family Guy
Continuously Clear Skincare
Ver Para Creer
Kind Hearts and Coronets
Metallica: Some Kind of Monster
Some Kind of Wonderful
David Carradine's Tai Chi for the Mind and Body
Chi for Beginners
Chi-Town Giants
Chi Girl
Gaither: The Best of Guy Penrod

94 more

Figure 3: A variable highlighting feature could indicate what the system thought it might have heard and how that contributed to the results.

5. DISCUSSION & FUTURE WORK

Indications from the usability testing are that SILO interfaces for TV search could be accepted in the marketplace but only if the high expectations of retrieval/recognition performance can be met. However, it is obvious that such a speech-based search feature will never work at 100%, so if it is to be successful, users must learn to expect something less. How good is good enough? That is a question we are not able to answer with this study.

We see two possible directions for improving recognition/retrieval performance for interfaces of this type. One is to build more constraints into the search before the user utters the query. For example, one could imagine a GUI in which a user chooses, say “Program Title Search” or “Actor Search” or a category such as “Sports” before stating the query. However, besides adding more complexity to the interface, this sort of design may risk failing to constrain users’ behavior anyway. Perhaps a better course of action would be to work on statistical ranking methods outside the sound domain in order to improve the retrieval performance much as current Internet search engines such as Google do. A variety of data could be used to improve such rankings such as program popularity in general or in particular households. In fact, the methods used by personalized recommendation services [8] might contribute to ranking results. We believe also that our sound- and language-based statistics can be further improved.

An issue with SILO interfaces that was evident in our studies is that when the participants’ queries failed, there was little information available to the user to understand why it failed. Subsequent to the studies reported on here, we did come up with a proposal that may partially address this issue. We call it variable highlighting. An example is shown in Figure 3. In this example, the user uttered “Queer Eye for the Straight Guy,” which was returned third in the list of results with five of the six words getting some highlighting. However, the recognition engine concluded that it might have heard other words as well. In fact, the word “kind” received the highest score, as indicated by the most intense highlighting. It is easy to surmise that this (mistaken) scoring was the biggest contributing factor to the result that “Two of a Kind” ranked highest on the result list. With this sort of a visual feedback, we hypothesize that users may at least come to some understanding of how and why SpokenQuery returns the results that it does when those results do not match the users’ expectations. Our hope is that it would lead to more successful recovery strategies should users employ them and that it may also lead users to be more forgiving of the system. We will leave the testing of this hypothesis to future work.

6. CONCLUSION

Our main findings from this study are, first, that a SILO design based on SpokenQueries could address a need in TV content search if retrieval performance can meet user expectations. Second, that recognition/retrieval performance and overall UI design in this domain needs further development to manage such expectations realistically. Besides recognition/retrieval performance improvements, we believe that other forms of feedback are needed in the case of inevitable failures. We suggested one idea in the form of variable highlighting of terms in results. Clearly, more research is needed.

7. ACKNOWLEDGMENTS

We thank Bhiksha Raj and Peter Wolf and other MERL contributors for their original conception of SpokenQuery technology used here as well as Bent Schmidt-Nielsen for his insightful comments and contributions to the research. We are grateful to Tribune Media Services for making their EPG data available to us for this research through their TMSDataDirect service.

8. REFERENCES

[1] Berglund, A., and Qvarfordt, P. Error Resolution Strategies for Interactive Television Speech Interfaces. In *Proceedings of International Conference on Human-Computer Interaction (INTERACT ‘03*) (Zurich, Switzerland, September 1-5, 2003). IFIP, Amsterdam, 2003, 105-112.

[2] CMUSphinx: The Carnegie Mellon Sphinx Project. http://cmusphinx.sourceforge.net/html/cmusphinx.php.

[3] Divi, V., Forlines, C., van Gemert, J.V., Raj, B., Schmidt-Nielsen, B., Wittenburg, K., Woelfel, J., Wolf, P.; and Zhang, F. A Speech-In List-Out Approach to Spoken User Interfaces. In *Proceedings o f Human Language Technology Conference (HLT 2004)* (Boston, Massachusetts May 2-7, 2004). Association for Computational Linguistics, 2004, 113-116.

[4] Forlines, C., Schmidt-Nielsen, B., Raj, B., Wittenburg, K., and Wolf, P. A Comparison between Spoken Queries and Menu-based Interfaces for In-Car Digital Music Selection. In *Proceedings of International Conference on Human-Computer Interaction (INTERACT ‘05)* (Rome, Italy, September 12-16, 2005). IFIP, Amsterdam, 2005, 536-549.

[5] Ibrahim A., Lundberg J. and Johansson J. Speech Enhanced Remote Control for Media Terminal. In *Proceedings of Eurospeech '01* (Aalborg, Denmark, September 2001). International Speech Communcation Association, Bonn, Germany, 2001, Volume 4, 2685-2688.

[6] Johansson, P. MADFILM--A Multimodal Approach to Handle Search and Organization in a Movie Recommendation System. In *Proceedings of the 1st Nordic Symposium on Multimodal Communication* (Helsingör, Denmark, September 25-26, 2003). Nordic Network For Multimodal Interfaces, 3003, 53-65.

[7] Nielsen, J. *Usability Engineering*. Morgan Kaufmann, 1st edition, 1994.

[8] O’ Sullivan, D., Smyth, B., and Winson, D. Improving the Quality of the Personalized Electronic Program Guide. *User Modeling and User-Adapted Interaction*, 14, 1 (2004), 4-36.

[9] Stone, B. I Want a Movie! Now. *Newsweek Magazine*, Sept. 13, 2005.

[10] Wahlster, W. SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. In Krahl, R., Guenther, D. (eds), *Proceedings of the Human Computer Interaction Status Conference 2003* (Berlin, Germain, June 2003). DLR, 2003, 47-62.

[11] Welcome to Promptu. *http://www.promptu.com*.

[12] Wolf, P., and Raj, B. The MERL SpokenQuery Information Retrieval System: A System for Retrieving Pertinent Documents from a Spoken Query. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*(Lusanne, Switzerland, August 26-29, 2002). IEEE, 2002, Vol. 2, 317-320.

[13] Wolf, P., Woelfel, J., van Gemert, J., Raj, B., and Wong, D. SpokenQuery: An Alternate Approach to Choosing Items with Speech. In *Proceedings of International Conference on Speech and Language Processing (ICSLP)* (Jeju Island, South Korea, October 4-8, 2004). ISCA, 2004, 221-224.