A Multi-modal Spoken Dialog System for Interactive TV

R. Balchandran, M. Epstein, G. Potamianos, and L. Seredi
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
rajeshb, meps, gpotam@us.ibm.com

ABSTRACT

In this demonstration we present a novel prototype system that implements a multi-modal interface for control of the television. This system combines the standard TV remote control with a dialog management based natural language speech interface to allow users to efficiently interact with the TV, and to seamlessly alternate between the two modalities. One of the main objectives of this system is to make the unwieldy Electronic Program Guide information more navigable by the use of voice to filter and locate programs of interest.

Categories and Subject Descriptors: H.5.2 [User Interfaces]: User interface management systems; Voice I/O; Natural language

General Terms: Human Factors.

Keywords: Natural language speech interface for TV.

1. INTRODUCTION

Over the years there has been an explosive growth in the number of television channels and programs available to consumers. Most households have access to hundreds of channels and listings of thousands of programs each week. However, the primary interface to navigate through these programs has remained the same for the last twenty-thirty years, namely, the Remote Control (RC). With the advent of the Electronic Program Guide (EPG) the remote control has been expanded with arrow and selection buttons for navigation. However, even as the remote control is very effective for basic tasks such as switching channels and controlling the television set such as volume, it is very inefficient in dealing with tasks such as selecting a channel by name instead of number, searching for programs in the program guide, etc. The main reason for this is that the remote control has to perform these tasks using a restricted keypad and a set of hierarchical on-screen menus – making the task of selecting an item by name very cumbersome.

A better interface is clearly needed with speech based interaction being a natural choice. Indeed, during the last few years a number of products such as, *Accenda, inVoca, PoGo, VoiceMe, Promptu* etc. have been introduced in the market, which integrate remote control and speech recognition to control TV. However, the use of these devices is not easy as the set of allowable commands is quite restrictive and static and they do not cater to the dynamic nature of the EPG. Our goal is to overcome these limitations.

In this demonstration we will show a prototype system that aims to complement the standard TV remote control with a speech interface that is much more effective at direct specification of "names" and also can enable direct access to most functions. The goal of this system is not to replace the remote control by voice – rather it aims to make the two modalities work interchangeably so that the most effective mode can be used at any point in time. The system gets updated periodically to be in sync with the latest content of the EPG.

One of the compelling features of this system is the support for *Natural Language Understanding*(NLU) enabling users to phrase their requests in a flexible manner without being restricted to the use of keywords. Additionally, users can provide multiple pieces of information in a single request, thereby enabling power users to efficiently use the system. At the same time, the system is also designed to guide naive users through step-by-step prompts. The overall user interface – both voice and haptic – and the interaction with various peripheral components is managed by a dedicated programmable Dialog Manager.

This demonstration system is part of the of the overall prototype being developed within the EU DICIT [3] project, which includes far-field microphone technologies to enable users to speak to the television from a distance of a few meters [1], [2]. This demonstration primarily focuses on the multi-modal and dialog aspects of this system.

2. SYSTEM OVERVIEW

This system uses IBM's Embedded Via Voice (EVV) for natural language speech recognition and interpretation. It also uses IBM's CIMA (Conversational Interaction Management Architecture) for multi-modal Dialog Management and component integration. EPG data is obtained from *XMLTV* (www.xmltv.org).

3. FUNCTIONAL OVERVIEW

In the prototype system, the Voice User Interface (VUI) has been designed to work hand-in-hand with the remote control, so users can use the most efficient modality at any

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.
ACM 978-1-60558-198-9/08/10.

point in the interaction. In addition the VUI caters to both naive and experienced users. The following is a summary of supported functions:

- Basic functions such as channel selection by name and number, volume control using relative and absolute values etc.
- Use of pop-up windows for disambiguation lists, and other informative messages.
- Support for expert and guided modes of interaction (selectable by user).
- Support for user profiles for favorite channels etc.
- Multiple ways for scrolling and selection of list items. For example, users can select by saying the name of the desired item, saying "select the fifth one", "number five", or "select that" to choose a highlighted item. Users can also employ the numeric buttons on the remote or scroll using the arrow buttons and select with the OK button or use a combination of voice and remote control.
- Direct access to most functions from the top level menu in addition to traditional directed dialog.
- Includes support for extensive ambiguity resolution – three types of disambiguation situations are considered (in combination) - ambiguity in token value, ambiguity in combination of two or more tokens and N-best results (including free-form results) when ASR confidence is low. The collective ambiguity is presented to the user for resolution.
- Users can restate requests to recover from errors without always having to start over.

Browsing for programs

The EPG typically includes thousands of programs and finding programs of interest is a challenge to users. With the limited search capability available in current interfaces, most users loose patience after scrolling through a few screens of programs - even for a single channel.

One of the main objectives of the prototype system is to make the unwieldy EPG more usable by providing a variety of mechanisms to filter and locate programs of interest. Users can filter using several categories such as, channel, genre, artist name, program name, day and time etc. The multi-slot recognition capability of the NLU system allows experienced users to specify one or more of these filter criteria in a single request. Alternatively, users can also add/remove criteria over several turns incrementally narrowing down their search. The following example illustrates these concepts. Consider that the user wants to locate programs of a certain genre at a certain point in time. The user could achieve this in one of the following ways:

Naive user:

User: Show me the guide

System: (EPG is displayed)

User: Search by genre

System: Choose from business, drama...

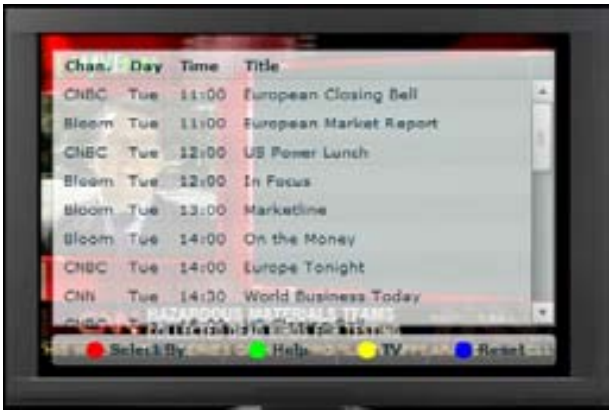

Figure 1: Filtered EPG list as seen in prototype

System: (Genre values are listed on screen)

User: business OR the first one OR (user selects by remote)

System: Listing programs for genre drama

System: (Filtered EPG list is displayed)

User: Show programs for tomorrow

System: Listing programs in genre business for Tuesday

System: (EPG list is updated)

Experienced User:

User: What's available in business for Tuesday afternoon

System: Listing programs in genre business for Tuesday

System: (Filtered EPG list is displayed)

The system lists matching programs as shown in Figure 1.

4. SUMMARY AND FUTURE WORK

In this demonstration we have presented an advanced prototype that implements a multi-modal interface for control of the television. The functions of the dialog application along with mechanisms for browsing for programs in the EPG were also described.

5. ACKNOWLEDGMENTS

This work was partially supported by the IST EU FP6 IST-034624 DICIT project [3].

6. REFERENCES

[1] R. Balchandran, M. Epstein, L. Seredi, M. Omologo, M. Matassoni, and R. Manione. The DICIT project: an example of distant-talking based spoken dialog interactive system. In *Workshop on Spoken Language Understanding and dialog Systems, IIS 2008*

[2] J. Huang, M. Epstein, and M. Matassoni. Effective Acoustic Adaptation for A Distant-talking Interactive TV System. In *Interspeech 2008*

[3] DICIT (Distant-talking Interfaces for Control of Interactive TV). [online] *http://dicit.fbk.eu*