Summary of Speech Technology and Market Opportunities in the TV and Set-top Box Markets: hands-free remote control systems

DICIT Consortium
(IBM (Praha - Czech Republic, T.J Watson Research Center - USA), ITC-irst (Trento, Italy), University of Erlangen-Nuernberg (Erlangen, Germany), Fracarro Radoindustria (Castelfranco Veneto, Italy), 3Soft (Erlangen, Germany), CitecVoice (Torino, Italy), Alpikom (Trento, Italy))

1. Introduction
This report provides a brief summary of the current state-of-the-art for speech recognition technology in the consumer TV and Set-Top Box (STB) market, examining and comparing several different speech-enabled solutions which are available today. It also provides a brief summary of the market opportunity for speech technology in STB and DVR consumer devices. In particular, the report addresses the use of hands-free remote control devices.

2. Speech Capabilities in the TV and STB Markets
Several devices currently available on the market support speech input for control of various TV functions and services. These devices generally fall into two different categories:
- Stand-alone handheld TV remote controls which allow buttons to be either pressed manually or activated by voice command
- Set-top boxes (either separate stand-alone devices or fully integrated with the TV service) which support voice commands to access certain functions

In general, the speech recognition component runs on the remote control device; however, locating the entire technology in the STB device itself, or adopting a Distributed Speech Recognition (DSR) based solution, seem to be effective alternatives to support a more complex spoken dialogue interface.

The following section provides information on several of the most popular speech-enabled TV devices currently on the market, along with a summary of their speech-based functionalities (as per specifications from the manufacturers).

Accenda
Accenda (see www.accenda.tv) is a stand-alone handheld TV remote control which allows functions on the keypad to be spoken by voice. It is designed to replace the standard infrared remote which is used to control TV’s and other related equipment, and uses speech technology from Innotech Systems. A voice command can activate either a single button or a sequence of buttons on the remote, and several different devices (TV, VCR, DVD, etc.) can be controlled. The device has a built-in microphone and is designed for close-talking operation (30 – 60 cm from the mouth). The underlying speech technology is template-based word matching, and each command must be trained individually for the user’s voice. A maximum of 50 voice commands is supported. Accenda does not have a speech dialog manager or support conversational interaction, although it does provide a pre-recorded acknowledgement to indicate that a voice command or button has been activated.

1 Copyright ©Partners of the DICIT consortium. This paper, or a short extract of it, can be reproduced, republished, or distributed, only if the DICIT Consortium is acknowledged and authors give permission. For further information, please contact the reference person, Maurizio Omologo (ITC-irst, Italy), at the following e-mail address: omologo@itc.it
inVoca

inVoca (see [http://www.remotecodelist.com/remotes/invocamanual.pdf](http://www.remotecodelist.com/remotes/invocamanual.pdf)) is a stand-alone voice-activated TV remote similar to the Accenda remote described above. It uses template-based word matching with a maximum of 50 voice commands which can be trained. Each command can activate either a single button or a sequence of buttons. As with Accenda, there is no capability for complex speech interaction or dialog, although the device does have a small LCD display for feedback, in addition to voice acknowledgements.

PoGo VRC-400

The PoGo VRC-400 (see [http://www.pogoproducts.com/vrc400.html](http://www.pogoproducts.com/vrc400.html)) is a stand-alone voice-activated TV remote similar to the Accenda remote described above. As with Accenda, it uses template-based word matching, although it supports up to 80 voice commands as compared to Accenda’s maximum of 50 words. The device’s 80 commands can also be partitioned across up to four users, with a maximum of 20 commands for each user.

VoiceMe

Human Oriented Technology’s VoiceMe (also marketed in Europe as the Auvisio VA R/C 3000) is a voice-activated table-top infrared remote (see [http://www.hotech.com.tw/products/voiceme/features.htm](http://www.hotech.com.tw/products/voiceme/features.htm)) which is designed to replace or supplement a standard handheld remote. Its speech recognition technology uses speaker-dependent template-based word matching, and supports a maximum of 30 voice commands. Unlike other similar voice-activated remote controls, the VoiceMe remote is not intended for handheld use and does not have a full keypad for accessing functions manually. This 15cm diameter device is instead designed to be placed near a standard set-top box at some distance from the user (up to 5 meters away as per the device’s instruction manual), and is activated only by voice commands. VoiceMe also uses an “always listening” mode of interaction, whereby a special trigger word is spoken to get the unit’s attention before speaking a command. Each voice command can activate up to 3 functions of a standard remote, and several different devices (TV, VCR, DVD, DVR, etc.) can be controlled. Since VoiceMe uses speaker-dependent speech technology, it can only be trained to recognize a single user. VoiceMe does not have a speech dialog manager or support natural conversational interaction.

AgileTV Promptu

AgileTV's Promptu system (see [http://www.promptu.com/](http://www.promptu.com/)) is a fully-integrated voice-activated set-top box which uses a handheld remote for voice input. Unlike stand-alone voice-activated handheld remotes which do speech recognition processing inside the remote itself, Promptu uses DSR technology to encode voice input and transmit it over the service provider’s cable connection for remote processing at the cable provider’s central office. This allows significantly more speech processing power to be available (and therefore more sophisticated voice functions) compared to systems which do all speech processing inside the handheld device itself. Promptu’s handheld remote contains a microphone and a push-to-talk button which is held down when speaking voice commands. The remote transmits the voice signal over infrared connection to the set-top box, which then encodes and sends it over the cable connection. Promptu uses speaker-independent phonetic-based speech-recognition technology, which means the system does not need to be trained for each user and the vocabulary can be flexibly defined according to the particular context. Voice commands with Promptu follow a fixed grammar format, depending on the category of the command. To tune the TV to a particular channel, a user may say “Channel 7” or “CNN”, for example. Since Promptu is integrated with the service provider’s cable network, it also has access to electronic program guide (EPG) information, unlike simple stand-alone devices. This allows commands to be spoken for scanning or searching the EPG information. For example, a user may say “Scan Sports” to scan through all sports channels, or “Find Spider-Man” to search the EPG for any channels and broadcast times at which “Spider-Man” can be watched. However, Promptu does not allow naturally-spoken commands which do not follow it’s pre-defined grammars, and does not have any speech dialog manager for conversational interaction with users.
2.1. System Comparisons
Most of the existing systems listed above are simple speech-enabled replacements for handheld infrared remote controls, which offer a limited set of voice commands for activating keys (or a sequence of keys) on the remote. This limited speech capability is due both to the limited processing potential of most handheld battery-powered devices, as well as the lack of integration and access to EPG and other STB information. Only the Promptu system allows access to a richer set of commands for searching and selecting program guide information by voice. On the other hand, simple voice-activated remote controls can be easily installed and set up to work with almost any existing STB or TV device, whereas the Promptu system requires a complete centralized speech server infrastructure to be set up and maintained by the cable service provider (and cannot work with satellite-based TV services which don’t have a cable connection for DSR transmission).

In terms of audio capabilities, only the VoiceMe system is designed to be operated hands-free at a far distance from the user, making use of both a far-talking microphone as well as a special “trigger word” which activates the device to listen for commands. All of the other systems use a close-talking microphone which is contained in the handheld remote, along with a push-to-talk button which must be pressed when speaking a voice command.

3. Market Opportunity for Speech in the STB and DVR Markets
Market research firm InStat expects the worldwide digital set-top box market to grow to 91 million units in 2005 and 130 million units in 2008. This rapid growth is driven by high consumer demand for several varieties of digital TV service (satellite, cable, IP-DSL, terrestrial digital HDTV, etc.), as well as new STB capabilities such as TV time shifting, as exemplified by TiVo and other similar systems.

As the sophistication and features of these STB’s and services grow, naturally the complexity of the user-interface required to access and control these services will also increase. This will place demand on new ways for users to easily access these services, such as through speech commands or multimodal interfaces which combine voice and visual interaction.

Shipments of advanced-feature set-top boxes are rapidly growing relative to basic-feature STB’s, according to InStat. Figure 1 below shows expected unit shipments of cable-based digital STB’s through 2008, for both basic and advanced market segments.
Digital Video Recorders in particular are one of the fastest growing segments of the consumer TV equipment market during the past couple years. Some of the popular leaders in this segment include TiVo and ReplayTV, in addition to a growing number of other satellite and cable STB’s which are now integrating DVR capabilities. As shown in Figure 2 below, unit shipments of hard-disk based DVR devices rose from 4.6 million in 2003 to 11.4 million in 2004, and are expected to grow by 58% to 18 million units in 2005 according to InStat. Satellite, cable, and DVD+DVR combination devices comprise the largest share of the DVR market segment.
Worldwide revenues in the DVR market are also expected to grow steadily between 2005 and 2009. As shown in Figure 3 below, DVR product revenues are forecast to grow to $6.7 billion in 2005 and $8.3 billion in 2006.

![Figure 3: Worldwide DVR Product Revenue (US $ in Millions) (Source: InStat MDR 5/05)](image)

Because of the numerous functions which are supported by DVR-enabled STB’s for accessing, recording and replaying content, speech and multimodal interfaces seem especially relevant for this segment of the market. Navigating through hundreds of channels of programs and lengthy electronic program guides, as well as searching and scheduling content to be recorded on DVR, can become challenging tasks when only a manually-operated handheld remote is available. Making these tasks available through a friendly and easy-to-use “conversational” speech interface can greatly improve the ability of consumers to successfully use these new devices and services.

One of the key challenges to enabling sophisticated speech interfaces is the memory and CPU processing limitations on the device where speech input is being recognized. This is clearly evident from many of the current speech-enabled TV remote controls mentioned above, which have a very restricted set of available voice commands due to the limited resources on a battery-powered handheld device. Performing the speech recognition on a platform with greater processing power can greatly extend the vocabulary and capabilities of the speech interface, as evidenced by the Promptu system which uses server PC’s to handle the speech processing. However, networked server-based speech systems such as Promptu do provide an obstacle to widespread deployment of speech-enabled STB’s, since a large speech server infrastructure must first be set up by the cable service providers. The best balance between advanced speech capabilities and ease of deployment may therefore be to locate speech processing on the STB device itself, which can provide significantly more capabilities than handheld devices while eliminating the need for a large server infrastructure to be deployed beforehand.