



WHITE  
PAPER

# From Voice Commands to Natural Conversations The Next Generation of Speech- Driven Entertainment Discovery



# CONVERSATIONAL INTERFACES

**SCOTT:** Computer, last message received and recorded from Captain Kirk.

**ENTERPRISE COMPUTER:** In place.

**SCOTT:** Run it through analyzer. Question: Is it or is it not the Captain's voice?

**ENTERPRISE COMPUTER:** Negative. A close copy.

**SCOTT:** A voice duplicator?

**ENTERPRISE COMPUTER:** 98 percent probability.

**SCOTT:** [to McCoy] Well, they've got them, Doctor, and now they're trying to get us.

– Star Trek: A Taste of Armageddon (episode 1.23, 1967)

While science fiction has long portrayed humans traveling to outer space as the final frontier, it has also depicted speech-driven interfaces as the ultimate human-machine interface. At the same time, the reality of speech-driven interfaces has been anything but the natural, virtually human speech capabilities envisioned. While speech-driven interfaces have been around for decades, practical uses are limited to supporting basic structured queries within hierarchical menus.

Recently, this has started to change. With the wide adoption of smartphones and the increasing use of virtual assistants such as Apple's Siri, speech interfaces that go beyond basic menu navigation and data retrieval have started to catch the fancy of consumers and application developers. Myriad "intelligent virtual assistant" applications have debuted from established companies and unknown start-ups alike. The availability of speech-to-text engines allows the basic speech enablement of almost any application that previously required tactile inputs, such as keypads, keyboards, mice or touchscreens.

Given the relative ease of "bolting on" a speech engine to any application, it is not surprising that the performance of such voice applications and virtual assistants varies greatly. Although there is evidence of serious attempts to try to break through to the futuristic ideals of speech-driven interfaces, most endeavors still rely on relatively structured menus for information retrieval in which spoken keywords simply replace keyed input counterparts.

Existing systems with conversational attributes are inherently task-oriented request-response systems, providing a notion of conversation continuity, though each request-response pair is independent of the other, and the only context maintained is the simple context. For example, a system may only address a user's bank account with a fixed number of accounts within it, and discrete transactions within each account. When it comes to achieving true conversational interfaces, the book is just being written.

In order to create truly conversational interfaces, speech cannot simply be added onto existing application platforms. Speech as an input mechanism with speech-to-text conversion and natural language processing (NLP) serves the basic needs of a speech-driven interface, such as found in interactive voice response (IVR) and call center applications. However, building truly conversational interfaces requires a paradigm-shift approach and a bottom-up foundation that makes speech the central construct of usability design.

Conversational interfaces are features and their implementation should not place an undue burden on existing applications and devices.

## WHAT ARE CONVERSATIONAL INTERFACES?

Conversational interfaces are user interfaces that simulate natural communication qualities on devices and applications, such that users can interact with them in casual language modes – similar to the way humans converse with one another. These interfaces obviate the need for users to conform to hierarchical menus or structured spoken commands, so users get to the end result of their interactions in direct, flexible and intuitive ways. With conversational interfaces, machines appear to adapt to humans, thereby flattening menu-driven functionality.

### Conversational System

Imagine purchasing movie tickets online using a conversational interface.

**User:** I want four tickets for the Tom Cruise movie at Boston Common for this evening.

**Conversational System:** What show timing would you like, 6 p.m. or 7:30 p.m.?

**User:** 6 p.m., please.

**Conversational System:** Please give me your credit card number and expiration date.

**User:** Credit card number is xxx-xxx-1234, expiration 12/15

**Conversational System:** You have purchased four tickets for the movie *Name of Movie* starring Tom Cruise playing at the AMC Loews Boston Common Theater at 6 p.m. this evening. Your credit card ending in 1234 has been charged the amount of \$48. You can use this card to pick up the tickets at the ticket kiosk at the theater. Thank you.

The entire transaction is completed within a couple of minutes.

### Traditional Menu-Driven System

In a traditional menu-driven system (whether on a web page or interactive voice response system), the user would have to:

- Find the movie listing by searching for the movie using Tom Cruise as a search term
- Select the theater location for the movie (likely from a drop-down menu)
- Identify show times and pick the one for which tickets are required (also from a drop-down menu, or check boxes)
- Input the number of tickets required (from a drop-down menu or by entering the number in a field)
- Confirm the inputs, then get directed to checkout
- Input the credit card number to check out

The number of steps required and the time involved are significantly larger, and the process is more tedious.

Imagine what this can do when users are trying to book air travel – juggling dates, flight schedules and ticket prices – or deciding what to watch on TV between hundreds of linear TV channels, thousands of VOD titles and potentially millions of online video titles from multiple services.

While progress is being made toward pure speech-driven interfaces, existing simple request-response style conversational systems suffice only to address specific task-oriented or specific information-retrieval problems in small-sized information repositories. Such as in the above example, even if a simple request-response type system was used, the steps involved would be similar to the traditional menu-driven experience. These systems would fail to perform well on large corpus information repositories.

Conversational interfaces are ideally designed for information retrieval from large, complex information repositories. Examples include television programming, customer support for any large organization, automotive controls and troubleshooting, shopping and e-commerce websites/apps, and travel and tourism.

# WHY CONVERSATIONAL INTERFACES ARE IMPERATIVE

The era of touch-based interfaces as the primary mode of interaction will eventually be complemented by conversational interfaces, as users would rather speak their intent and have the system understand and execute. This has been triggered by the significant hardware, software and algorithmic advances making speech-to-text significantly more effective as compared to even a few years ago.

Meanwhile, when one looks at the evolution of connected devices, the form factors and usage trends of devices strongly support the need for conversational interfaces. Among these are:

## Smaller, Input-Constrained Devices

Devices are becoming smaller, portable and mobile. By virtue of their smaller form factors and restricted user interface capabilities, devices are inherently input-constrained relative to their rich functionality, storage and processing capabilities. Conversational interfaces are the ultimate, natural user interfaces that will shape the usability of such devices.

## Processing Power and Speed

At the same time, device size is no longer an indication of the processing power and performance speed it can offer. Handheld and mobile devices pack processing speeds and performance that surpass larger desktop computers from a few years ago. As a result, devices such as smartphones and tablets are being used for multi-functional applications spanning communications, productivity, entertainment and information. The performance of these devices impacts the sales of traditional “workhorse” enterprise and home computing machines, such as desktops and laptops. Given the usage demand of such devices compared with their (small) form factors and mobility, conversational interfaces provide the most suitable usability solutions.

## On-the-Go Usage

Most device usage is done “on-the-go,” where users have limited or no access to tactile and visual controls. For example, cell phone usage while driving is a known hazard, and is even prohibited in many states in the U.S. Thus, place and time of usage supports the utility of conversational interfaces. Increasingly, devices have always-on internet connectivity, whether through mobile and fixed broadband data networks, public Wi-Fi hotspots or other connections, further driving on-the-go anytime, anywhere usage of devices and applications.

## Pent-Up Demand

Decades of sci-fi consumption gave people the awareness that, someday, conversational interfaces would be possible, and technologies such as Siri spurred their appetite for and awareness of them. We expect this pent-up demand to result in a market pull as these technologies and solutions are introduced, thereby creating a cycle of higher supply and rapid innovation.

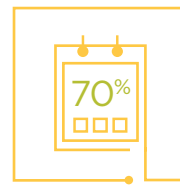
### High Demand for Voice Search Among Pay-TV Subscribers



61 percent would use voice search if offered by their provider.



After viewing a video demonstration of advanced natural-language voice search, 35 percent would be willing to pay an additional fee for the feature.



70 percent would be willing to extend their pay-tv service contract if they were offered better search or recommendations capabilities.

**Source:** Rovi-sponsored survey, September 2014. The research findings were the result of a Web-based survey of 2,000 pay-TV subscribers across the U.S., the U.K., France and Germany conducted in August of 2014.

# PERFORMANCE REQUIREMENTS FOR CONVERSATIONAL INTERFACES

Conversational interfaces require many different technologies and systems that need to work cohesively to deliver desired performance. These technologies include speech processing, speech-to-text, various natural-language processing schemes and others. This paper does not include coverage of all such ingredient technologies and sub-systems, rather specific solutions and advancements that TiVo has implemented to make conversational systems possible. Before we discuss TiVo's conversational interface system components in the next section, it is important to identify three requirements that we consider essential to conversational interfaces and core to our system: disambiguation, statefulness and personalization.

## Disambiguation

Conversational systems need to be capable of precisely disambiguating users' words, phrases and utterances as a primary requirement for conversational interfaces. The underlying requirement behind disambiguation of language is to precisely identify user intent. Relevance, popularity and context are all factors in determining intent and "appearing intelligent" to the user.

Disambiguation is required at multiple levels. For example, phonetic sounds can be attributed to multiple keywords.

Phonetic "Krooz" can be interpreted as "Cruise," as in Tom Cruise the actor, or "Cruz," as in Penelope Cruz, the actress. The system should be able to interpret the phonetic sound and further disambiguate by taking other factors into consideration. For example, if Tom Cruise has a popular film in theaters that month, the system should present "Cruise" as the desired keyword.

Another form of disambiguation required is ascribing the correct meaning to a keyword. For example, the keyword "Eagles" can be ascribed to a sports team (Philadelphia Eagles), a rock band (Eagles) and a bird, among other things. Even within the single context of sports, "Sox" could be ascribed to Boston Red Sox or Chicago White Sox, both of which are Major League Baseball teams.

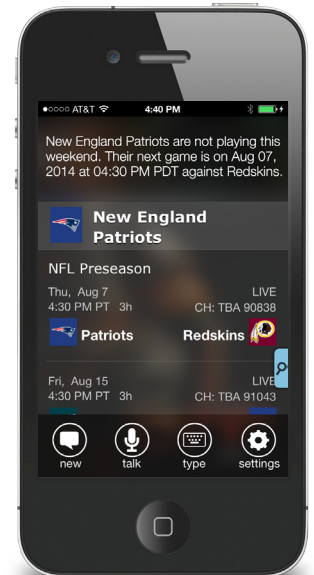
When asked, "Do the Pats have a game this weekend?", the system resolves the ambiguity of "Pats," a commonly used appellation for the New England Patriots American football team, through the implied sports reference of the word "game" as shown below.

Where such disambiguation is not apparent, the system resolves ambiguities through intelligent dialogue with the user until such disambiguation is completed and intent is determined.

For example, in the "Sox" example, the system asks a disambiguating question as shown below.

If the system has the intelligence to understand what's going on in the world — what's timely and trending — this disambiguating question may not be necessary at all. The system would know which team is competing in the playoffs, for example, and present the more relevant result to the user.

A key corollary to disambiguating speech is the requirement that ambiguity resolution exploits domain-specific structural knowledge. In other words, conversational interfaces built for specific vertical market applications (e.g., travel, television, automotive) can achieve faster, more precise disambiguation than general-purpose applications such as web search. Disambiguation through dialogue also requires statefulness.





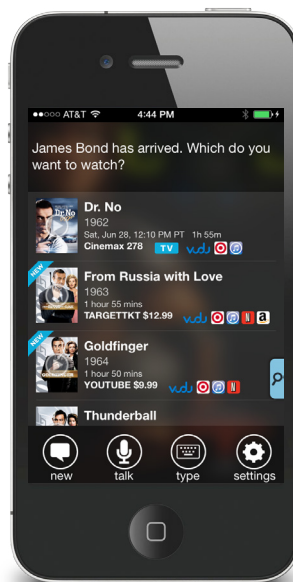
## Statefulness

In the course of disambiguation, a conversational system must be able to maintain the context or state of the dialogue. This is a prerequisite to resolving ambiguities and advancing the conversation in a logical way. It is not uncommon for users to switch the context arbitrarily. For example, when looking for a movie to watch on TV, a user may arbitrarily decide to first schedule a recording of an upcoming game before he forgets. The conversational system must therefore not only be capable of maintaining the context of a dialogue in order to complete the disambiguation, but it must also recognize when that context has changed within the same dialogue session and adapt accordingly. Statefulness also implies domain-specific structural knowledge.

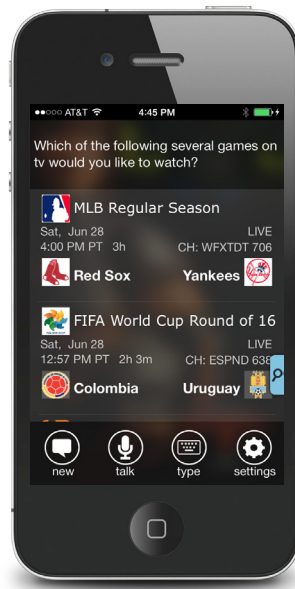
In the example below, the term “thrillers” is interpreted as a movie genre. Thereafter, the term “Bond” is interpreted as the well-known category of James Bond movies that are thrillers.



Subsequently, when the user asks for “old ones,” the system recognizes that the context is still James Bond movies.



Conversely, when the user switches the context from James Bond movies (and movies in general) to sports, the system is able to infer the new context or change of state.



A more advanced example of context-matching along with disambiguation would be where a user browsing Al Pacino movies may want to use the term “Tony Montana” to identify the movie in which Al Pacino has played a character with that name. The system should be able to recognize that Tony Montana in that context refers to the movie *Scarface*, not another actor whose name is also Tony Montana.

## Personalization

Using a conversational system without personalization capabilities is like talking to a stranger every time. While this can work as a natural conversational system, over time users will deem it inadequate. For example, if a user asks the television, “What time is the game tonight?” and the conversational system needs to clarify that question without remembering the user lives in Boston and watches Red Sox games regularly, the system will fail to satisfy the user. Consequently, conversational systems need to be able to personalize responses and results to learned user preferences. In addition, conversational systems need to personalize responses spatiotemporally (i.e., with the awareness of location, time of day, day of week, time of year and other such variables).

Continuing the above Sox example, when a user may ask, “When is the Sox game tonight?” the system should be able to recognize that the user is referring to the Boston Red Sox and not the Chicago White Sox, given the user’s location (Boston) and/or prior viewing history.

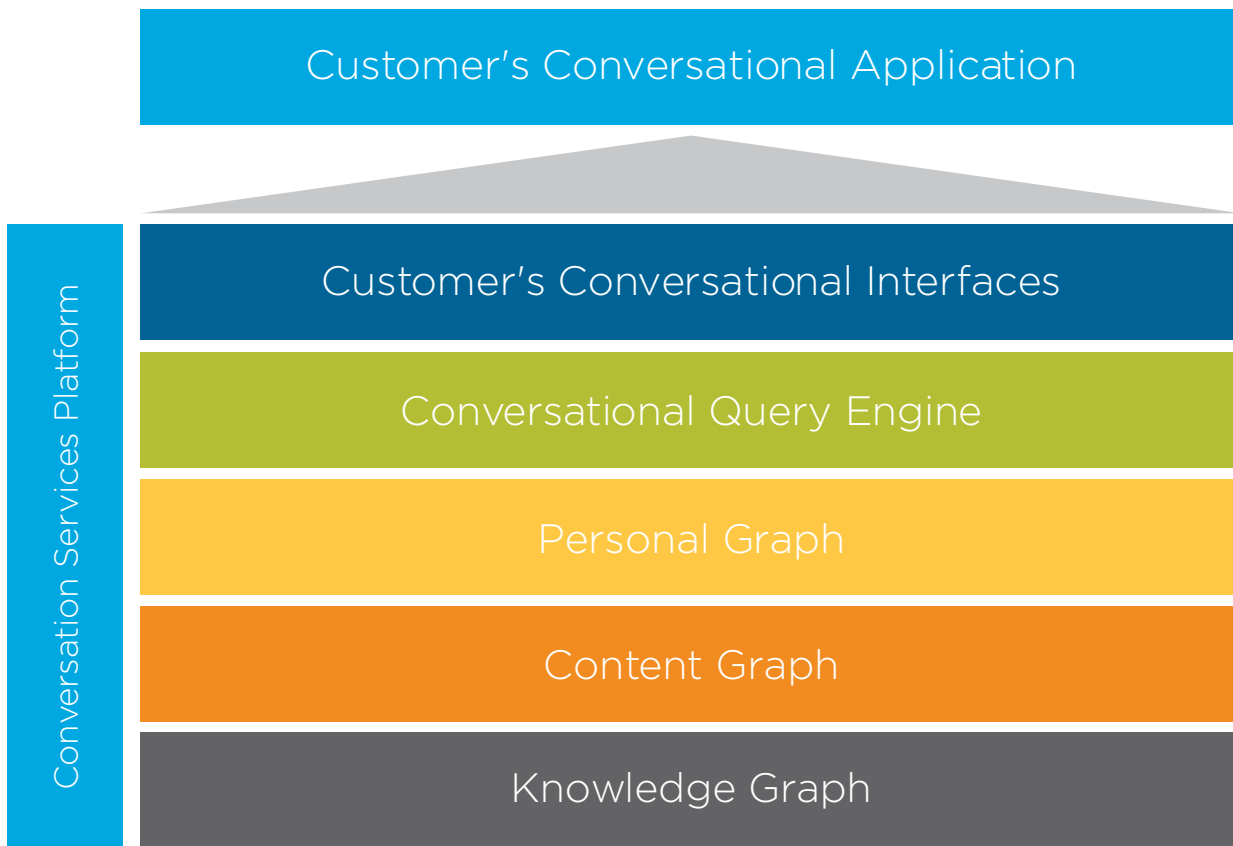
The natural resolution of ambiguities combined with the ability to maintain state and personalization across exchanges enables users to continue to seamlessly explore the information domain with the ease inherent in human interactions, where a user may indirectly refer to a person, concept or topic uttered in an exchange much later in a conversation. TiVo’s conversational interfaces also implicitly determine convergence/divergence/termination of a topic of discussion from each exchange and tailor responses accordingly. Therefore, conversational interfaces, unlike current simple request-response systems, make the dialogue very close to human interaction where ambiguity resolution, context-dependent user preferences weighting, and topic convergence/divergence/termination all happen naturally in a conversation.

By addressing these different characteristics that are central to a natural conversation flow in human interactions, TiVo’s conversational interfaces aspire to inch closer to the ideal of simulated natural human conversations with devices and applications.

**APPENDIX A** of this white paper demonstrates a number of use cases to further illustrate how a good conversational system performs under various real-world conditions requiring disambiguation, statefulness and personalization.

## CONVERSATION SERVICES

While exploring the specific mechanisms underlying TiVo's conversational interfaces is beyond the scope of this paper, it is important to highlight that our conversation services make conversational interfaces possible.



It should be evident from the performance requirements of conversational interfaces that building solutions with truly conversational speech interfaces requires a comprehensive, end-to-end approach, which TiVo's conversation services provide. The platform is backed by more than 65 patent applications and 32 issued patents. The key elements of conversation services are:

### Knowledge Graph

Conversational system challenges are complex search challenges. Semantic disambiguation to yield accurate search results requires a knowledge graph approach that makes it possible to map search results to intent, not simply to search terms and keywords.

At its core, a knowledge graph comprises:

- Dynamic reference of named entities (smart tags) that exist in the world across media, entertainment, geography, celebrities/ personalities, sports and specialized named entities that exist in enterprise verticals (e.g., e-commerce, travel, health, etc.)
- Algorithms to extract, de-duplicate and disambiguate the entities across sources
- Relationships across the disambiguated entities encoded as edges of the graph



## Content Graph

The content graph comprises the actual content assets that are consumable by users, whether these are videos, contact names and addresses, database entries, or others. The content graph maps these content assets within a library to the knowledge graph. As such, a knowledge graph can support multiple content graphs. At the same time, since content graphs are vertical-specific, such as movies and television, travel and leisure, healthcare and so on, knowledge graphs must be augmented with vertical-specific data space characteristics. Correspondingly, conversational systems must be implemented for specific vertical market applications, as opposed to generically for any or all market applications, in order to meet stated performance requirements.

## Personal Graph

Personalization is a key attribute of true conversational systems. The personal graph implicitly tunes the conversational system to individual users in order to provide the capabilities that simulate natural human conversations.

At its core, the personal graph comprises the following functions and capabilities:

- It is a concise mathematical model based on statistical machine learning
- It is a hyper-personalized, contextual learning engine that implicitly learns individual behavioral patterns and interests
- Its spatiotemporal model learns time of day, day of week with current location, device, etc., intersected with short-term and long-term memory
- User activities and interests are captured in terms of named entities (smart tags) which are nodes in the knowledge graph
- Smart tags and weights are the fuel for the personal graph

## Conversational Query Engine

The conversational query engine is the front end of the conversational system on which the actual conversational interfaces are implemented. It is the engine that binds all of the other parts of the system together, including delivering the following feature sets:

- Conversational interfaces that simulate natural language dialogues, based on entity extraction and subsequent intent determination through fast graph computation algorithms
- Linguistic features (pronouns, verbs etc.) are mapped to certain node types of the knowledge graph and adjectives typically handled as filters
- Machine learning of linguistic features per vertical data space
- Stateful intent determination enabled by personal graph
- Disambiguation of user intent even where the natural language query lacks discernible entities (as discussed in prior sections)
- Automatic inference of conversation state (if it is a continuing dialogue vs. new session)
- Instant results for “search-as-you-type” functionality
- Robust content discovery features including home screen recommendations (“more like this”) and collaborative filtering based on characterizing individual, and aggregate user behavior through the personal and knowledge graph
- APIs for multiple devices (smartphones, tablets, TVs, etc.) that share the cloud-based infrastructure for transparent hyper-personalization

## Implementing TiVo's Conversational Interfaces

While requiring an end-to-end approach, conversational interfaces are ultimately an essential feature of many existing and emerging products and services. This implies that practical implementation of conversational interfaces should allow for integrating them with existing product and services platforms. Intelligence based on our knowledge graph and conversational query engine resides in the network and cloud. Based on individual customer requirements, this can live within a customer's or enterprise network. In other implementations, TiVo hosts this in the cloud for different services. Calls to end terminal devices can be implemented through API calls without any burden being placed on the end device or application requirements to support storage or processing for conversational capabilities.

The personal graph can also be implemented entirely in the cloud or over a network. In some cases, the personal graph may reside on the end terminal device in order to conduct personalization in off-network mode (e.g., airplane mode).

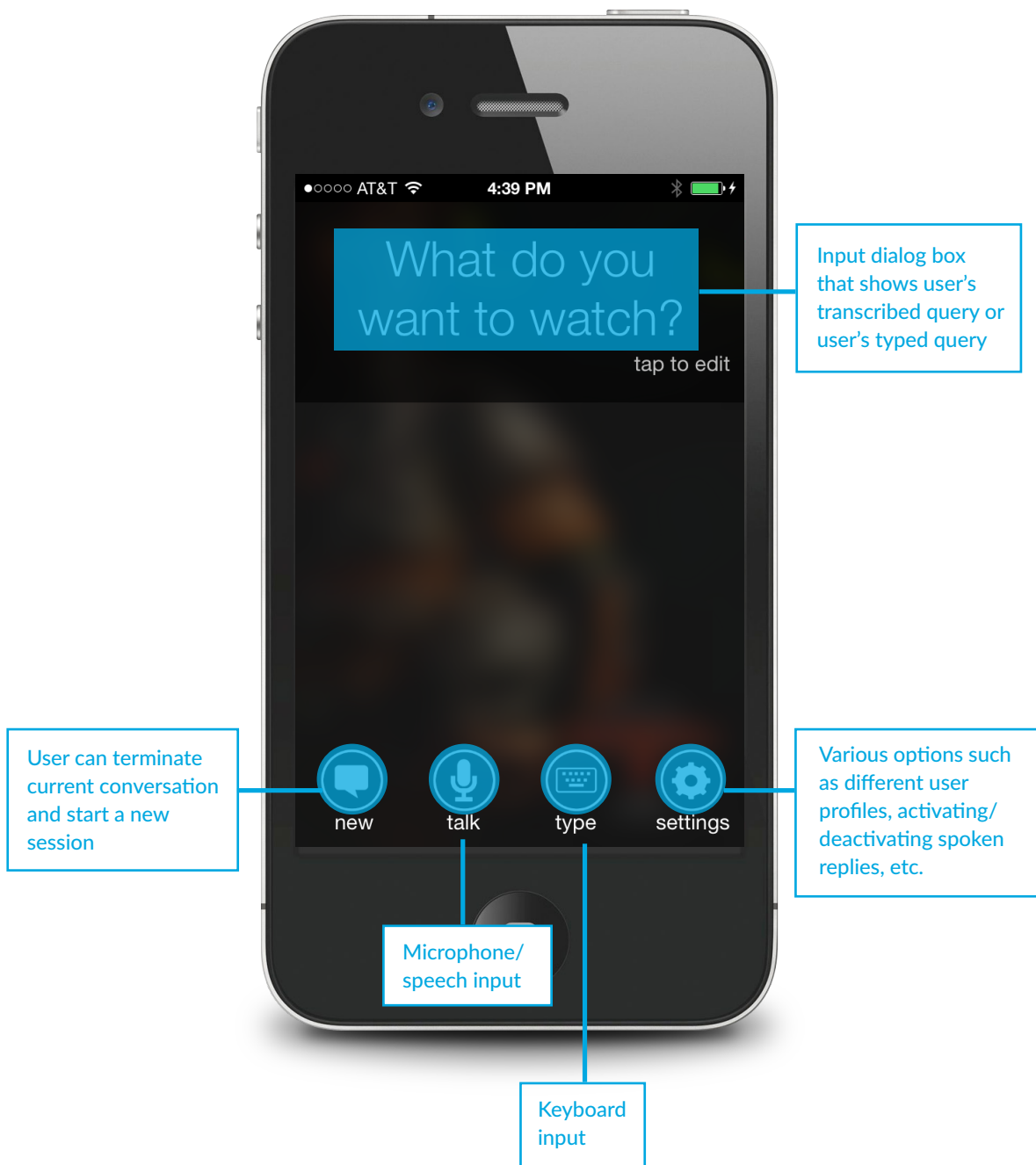
Consequently, while TiVo's conversation services are built from ground up as an end-to-end conversational interface platform, implementation is intended to be an extension of existing systems and application platforms. For different verticals and new markets, TiVo will augment the knowledge graph and integrate the customer's content graph to operate with the conversational platform. In this way, the support for new conversational interfaces is truly incremental, quick, efficient and affordable using our platform.

## Conclusion

Conversational interfaces are the inevitable next phase of interfaces required for the emerging era of smart, connected devices. Technology and market forces are driving toward conversational interfaces at a rapid pace. Conversational interfaces, however, are not simply speech enablement on existing solutions. Rather, they require a comprehensive approach with specific technical and usability requirements to perform as natural and intelligent interfaces. Disambiguation, statefulness and personalization are essential requirements for true conversational interfaces. At the same time, given the requirements of conversational interfaces, they are ideally implemented for specific vertical data spaces. A semantic platform based on a knowledge graph implemented in TiVo's conversation services delivers to the performance requirements of conversational interfaces. Our platform also addresses implementation challenges by allowing conversational interfaces to be integrated with existing products and services through a comprehensive yet flexible and affordable cloud-based implementation.

## APPENDIX A

### Opening Screen:

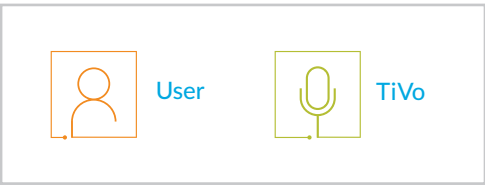


### Screen Capture Demonstration of TiVo's Conversation Services:


Please see the following pages for a walk-through demonstration with screen captures.

**Note:** These these are actual queries conducted on an iPhone 4S over a cellular data network.

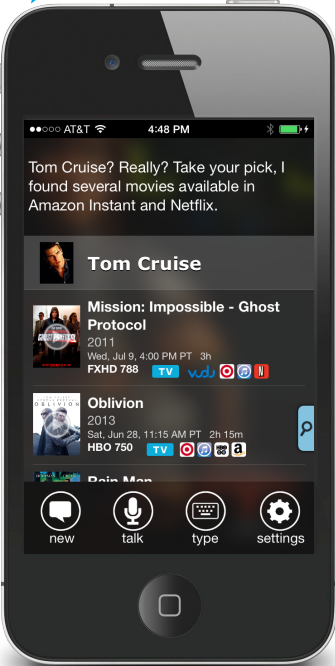
# EXAMPLE 1: SWITCHING PROGRAM SOURCE



1



Spoken query  
"...Tom Cruise  
movies"




Tom Cruise? Really? Take your pick, I found several movies available in Amazon Instant and Netflix.

**Tom Cruise**

**Mission: Impossible - Ghost Protocol**  
2011  
Wed, Jul 9, 4:00 PM PT 3h  
FXHD 788

**Oblivion**  
2013  
Sat, Jun 28, 11:15 AM PT 2h 15m  
HBO 750


new talk type settings



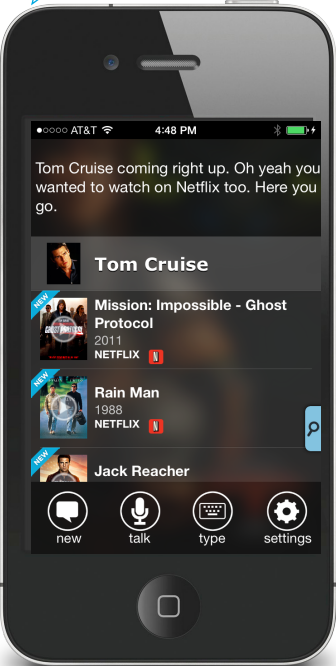
Spoken reply  
with transcribed  
output

**Note:** Results are shown as sliding cards for available movies (can be rendered as list, thumbnails, others as well).

2



Naturally spoken  
response "yes" –  
to TiVo



Tom Cruise coming right up. Oh yeah you wanted to watch on Netflix too. Here you go.


**Tom Cruise**

**Mission: Impossible - Ghost Protocol**  
2011  
NETFLIX

**Rain Man**  
1988  
NETFLIX

**Jack Reacher**


new talk type settings




Returns results  
for Netflix

**Note:** Retains context of Tom Cruise movies without user needing to.

3



Changes source from Netflix to  
iTunes in normal, casual spoken  
language: "what about iTunes"



Back

**Mission: Impossible - Ghost Protocol**  
2011 PG-13 IMDb 7 94%  
Directed By: Brad Bird  
Runtime: 180 minutes

Details Crew Related Programs


Wed, Jul 9 4:00 PM PDT 3 hours FXHD

WATCH LATER

Available on:

Netflix	iTunes	\$3.99
Vudu	Orbit	\$17.99
Google play		\$3.99

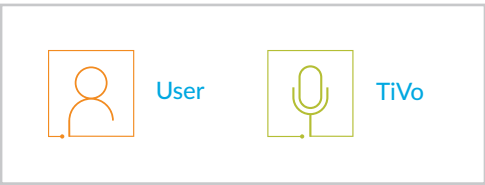
new talk type settings



Acknowledges change  
of source to iTunes  
and modifies results

**Note:** User can continue to modify, refine and advance query in this way while maintains context.

# EXAMPLE 2: REFINING SEARCH CRITERIA



1

User looking for "Robert Duvall's movies"

Quite a few movies to choose from with Robert Duvall.

**Robert Duvall**

**The Godfather**  
1972  
Sun, Jun 29, 1:00 PM PT 3h  
HBO 203

**The Godfather: Part II**  
1974  
Sun, Jul 6, 8:40 AM PT 3h 25m  
HBO 812

**Apocalypse Now**  
1979  
2 hour 27 mins

Given the large list of movies with Robert Duvall, TiVo assists user with prompt for "recent ones"

2

Normal spoken English response "yes..." - without reference to "Robert Duvall" or "movies"

I got Jack Reacher for you.

**Robert Duvall**

**Jack Reacher**  
2012  
2 hour 10 mins  
YOUTUBE \$5.99

**Jayne Mansfield's Car**  
2012  
Mon, Jun 30, 12:20 PM PT 2h 10m  
STARZ 245

**Seven Days in Utopia**  
2011  
1 hour 38 mins  
ITUNES \$4.99

Understands user's normally spoken response and refines the search to "recent" results (goes from 1972 *Godfather* to 2012 *Jayne Mansfield's Car*)

Note: There are multiple results as evident from the "dots" under the card that allow the user to scroll across the recent results.

3

Inserts additional criteria of "Nicole Kidman" without referencing any previously stated search criteria

Nicole Kidman and Robert Duvall last appeared in Days of Thunder.

**Nicole Kidman**

**Robert Duvall**

**Days of Thunder**  
1990  
1 hour 47 mins  
ITUNES \$3.99

Retains the context of "recent" movies of "Robert Duvall" but also adds the requirement of "Nicole Kidman," thereby showing movies that are recent Robert Duvall movies that also feature Nicole Kidman

4

Further refines query to see if the movie is available on Netflix

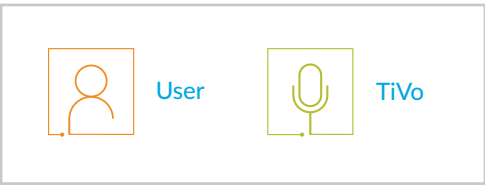
Playing Days of Thunder on Netflix.

**Days of Thunder**  
1990  
1 hour 47 mins  
NETFLIX

Retains the reference of the prior selection and applies additional user-stated requirement to the result

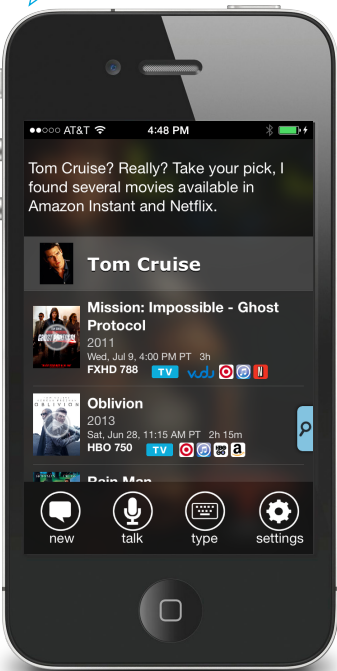
Note: The user does not reference the movie.

# EXAMPLE 3: SWITCHING SEARCH CRITERIA



1

Searches for actor and director combination

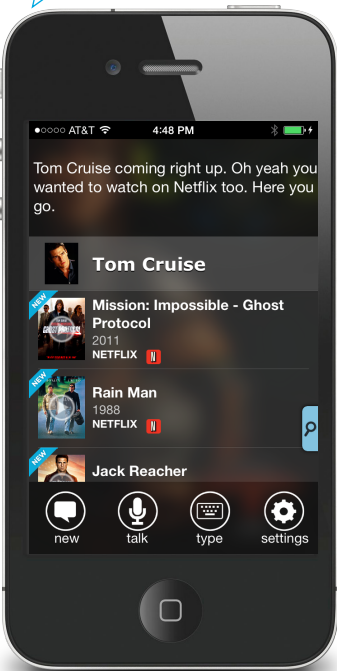


Gives results of movies by combined actor and director results

**Note:** This could have been a different query combination such as "Harrison Ford movies from the 70s."

2

Switches context (director) in normal spoken English



Recognizes that user wants movies of same actor as before (Harrison Ford), but with different director

**Note:** "Harrison Ford" was not mentioned again.

3

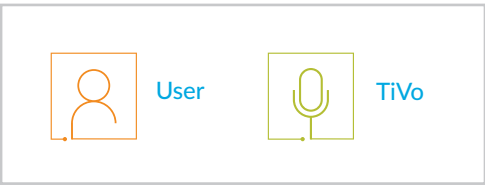
Asks for movie source without naming the actor, director or movie names again



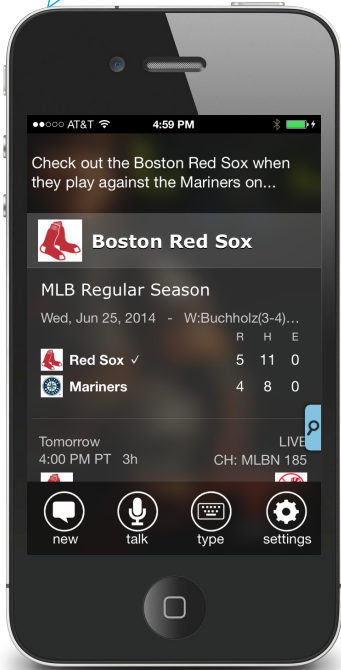
Provides results based on ongoing refinements and context of user's request - "George Lucas movies with Harrison Ford available on Netflix"



# EXAMPLE 4: EXAMPLE OF LINEAR (DATE/TIME) TV LISTINGS



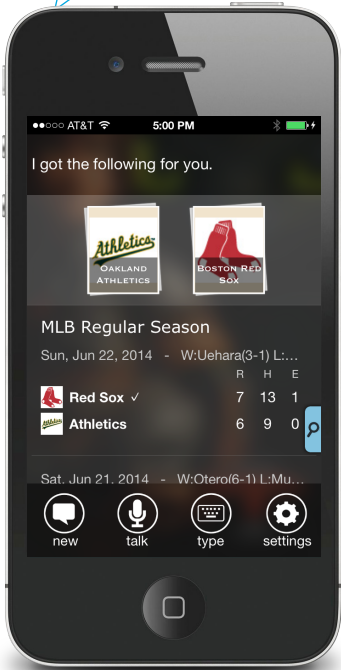
1 User searches for TV listing



Finds the game "tonight" per user request

**Note:** User does not specify date or time.

2 Switches context without naming the team (Red Sox) again

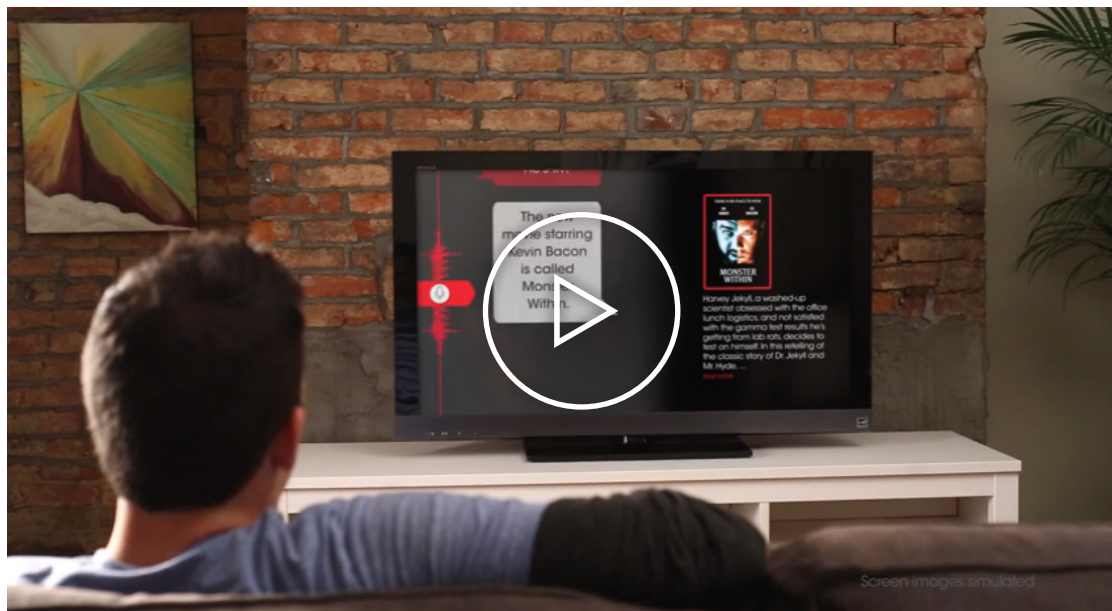


Maintains context of Red Sox as the basis of query and looks up listings

**Note:** The query would have been answered even if the user said "Oakland" only as normal spoken language.

## APPENDIX B

### Introduction to TiVo's Conversation Services:



## ABOUT TIVO

TiVo is leading the way in the discovery and personalization of digital entertainment. We help power top brands around the world with market-leading guides, metadata, recommendations, audience analytics and advanced advertising solutions. With products deployed through an innovative cloud-based platform, TiVo is enabling customers worldwide to increase their reach, drive consumer satisfaction and create a better entertainment experience.

TiVo's pioneering technologies bridge the semantic gap in usability for connected devices and applications based on our proprietary knowledge graph semantic database. Built on the knowledge graph engine, TiVo's conversation services are used by Tier 1 service providers in North America and major smartphone manufacturers globally for intelligent search, personalization and recommendations solutions.