Conversational Interfaces for Interactive TV Content Management

Rajesh Balchandran, Mark Epstein, Martin Labsky, Larry Sansone, Sara Basson
T J Watson Research Center

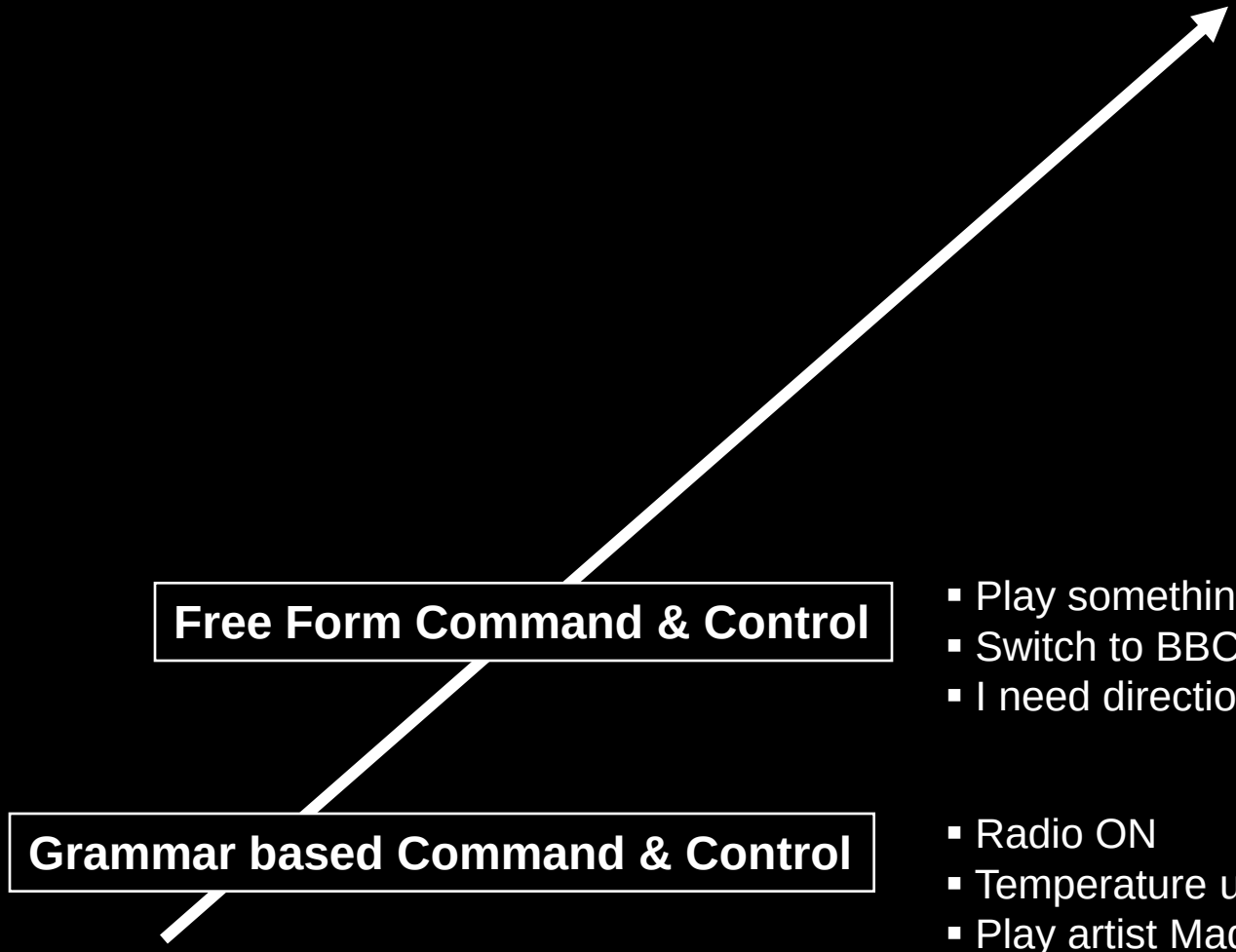
Evolution of devices

- Devices are getting smaller and smaller
- Conventional input & output modalities are becoming unusable

- Explosion in information content and device functionality

- Need direct access to desired information and functions
- Speech based interface can meet these challenges

Evolution to Conversational Interaction



- Play something by Madonna
- Switch to BBC on XM
- I need directions to Cambridge

- Radio ON
- Temperature up
- Play artist Madonna

Evolution to Conversational Interaction

- I'd like to listen to Sonata Number Eleven
  - I found Sonata Number Eleven in 2 albums:
    - The Best of Beethoven &
    - The Music of Mozart

    Which one would you like?
- The second one

Conversational Interaction

- I'd like to listen to Sonata Number Eleven by Beethoven

Key Technologies
- Natural Language Recognition and Understanding
- Dialog Management

Free Form Command & Control

- Play something by Madonna
- Switch to BBC on XM
- I need directions to Cambridge

Grammar based Command & Control

- Radio ON
- Temperature up
- Play artist Madonna

Next Generation of Voice Interaction

- User Interface characteristics
  - User input options and characteristics
    - Fixed vs. free form with multiple token extraction
    - Single mode vs. multi-modal
  - Nature of Interaction
    - Menu Driven vs. conversational
- Technology aspects
  - Speech recognition & interpretation
  - Dialog management
    - Single mode vs. asynchronous and multi-modal
    - Single device vs. multiple asynchronous devices
    - Static vs. dynamic content


= IBM Research Focus Areas

Demonstrations

Music Selection by Voice

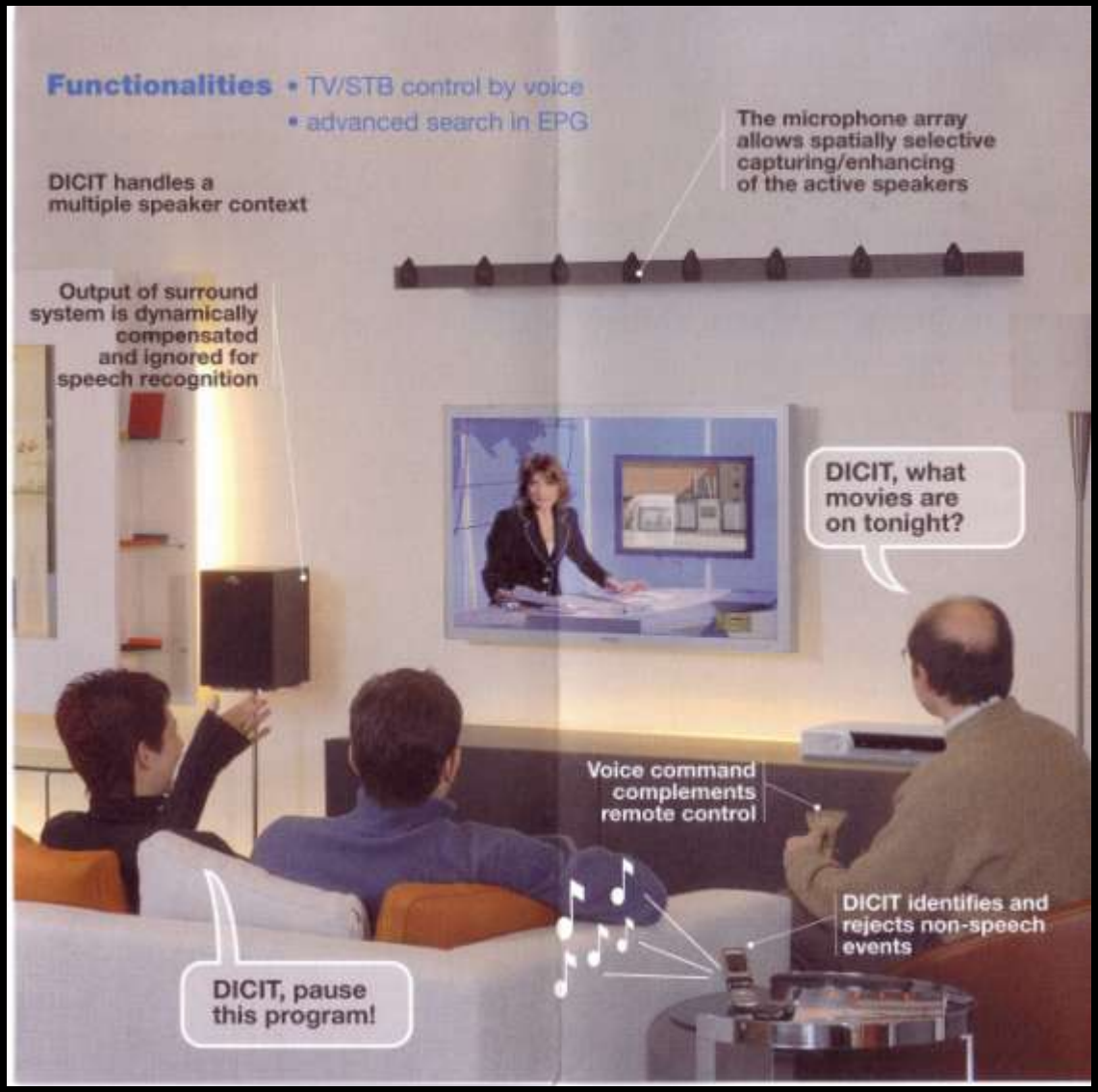
- IPod thumbwheel & similar interfaces are ineffective after a few thousand songs
- IPod shuffle has no display
- Combinations of artist, song etc. not possible

TV Program Selection by Voice

- Remote Control is limited and only provides menu based hierarchical interaction
- Search is very difficult

➢ Voice based interaction helps users navigate through thousands of songs and programs easily and efficiently

DICIT – Functional Overview



- Front End System
  - Acoustic Scene Analysis, Speaker Localization
  - Acoustic Echo Cancellation
  - Speech Detection
- ASR & Dialog System
  - Audio input from front end
  - Natural Language Speech recognition & understanding
  - Dialog Management
  - Managing Electronic Program Guide data, grammar generation
  - Controlling Set Top Box
  - Handling inputs from Remote Control

First Prototype Evaluation

- Close talking and far field ASR tests carried out at the 3 sites – FBK, Amuser & Elektrobit
  - 150 English sentences were spoken by 2 users at each site
  - Close talking and far-field audio was collected in parallel
- Evaluation with naïve users
  - Carried out by Amuser and Elektrobit

ASR Far Field Tests at 3 Sites

			e16i1479 - Adapted Acostic Model						
			Close-Talking Mic				Far Field Mic		
	Grammar	Counts	Amuser	EB	FBK	YKT	Amuser	EB	FBK
4	Initial-Ok-Grammar	68 / 64	97.8	100.0	100.0	100.0	100.0	100.0	100.0
5	On-Grammar	116.0	97.5	98.1	89.7	100.0	88.9	88.9	87.9
6	TV-Screen-Grammar	296.0	95.1	95.6	82.4	100.0	87.3	90.4	73.6
7	EPG-Grammar	524 / 51	86.0	87.3	66.7	96.0	79.1	81.0	62.4
8	Selection-Grammar	480.0	85.4	90.8	66.3	95.6	77.5	81.1	64.6
9	Search-By-Grammar	436 / 42	87.6	91.2	74.3	97.1	79.4	84.3	65.9
10	Ambiguous-Channel-Grammar	252 / 24	92.8	94.2	79.5	96.7	86.7	92.5	77.9
11	Settings-Grammar	216.0	96.0	98.0	85.2	98.0	90.0	91.0	83.3
12	User-Id-Grammar	108.0	97.3	100.0	85.2	100.0	93.3	98.0	88.9
13	Not-On-The-Air-Grammar	120.0	98.9	98.3	86.7	100.0	96.7	95.0	88.3
14	Help-Grammar	104.0	98.7	100.0	88.5	96.2	96.2	100.0	84.6

			e1600012 - Base Acoustic Model						
			Close-Talking Mic				Far Field Mic		
	Grammar	Counts	Amuser	EB	FBK	YKT	Amuser	EB	FBK
4	Initial-Ok-Grammar	68 / 64	100.0	100.0	100.0	100.0	100.0	100.0	100.0
5	On-Grammar	116.0	98.8	100.0	94.8	96.3	91.4	83.3	87.9
6	TV-Screen-Grammar	296.0	95.1	97.1	81.1	97.1	88.7	89.7	74.3
7	EPG-Grammar	524 / 51	86.5	92.1	64.3	93.7	77.5	78.6	56.2
8	Selection-Grammar	480.0	86.5	90.8	69.6	94.7	81.6	80.3	60.0
9	Search-By-Grammar	436 / 42	89.9	92.2	72.4	94.1	82.0	81.9	60.7
10	Ambiguous-Channel-Grammar	252 / 24	93.9	96.7	73.8	95.0	86.1	90.0	71.3
11	Settings-Grammar	216.0	98.0	98.0	86.1	98.0	90.0	85.0	80.6
12	User-Id-Grammar	108.0	96.0	100.0	88.9	100.0	93.3	98.0	83.3
13	Not-On-The-Air-Grammar	120.0	98.9	96.7	86.7	100.0	93.3	90.0	81.7
14	Help-Grammar	104.0	98.7	100.0	84.6	96.2	94.9	90.4	92.3

ASR Far Field Tests at 3 Sites – Adaptation Effect

			e16i1479 - Adapted Acostic Model						
			Close-Talking Mic				Far Field Mic		
	Grammar	Counts	Amuser	EB	FBK	YKT	Amuser	EB	FBK
4	Initial-Ok-Grammar	68 / 64	97.8	100.0	100.0	100.0	100.0	100.0	100.0
5	On-Grammar	116.0	97.5	98.1	89.7	100.0	88.9	88.9	87.9
6	TV-Screen-Grammar	296.0	95.1	95.6	82.4	100.0	87.3	90.4	73.6
7	EPG-Grammar	524 / 51	86.0	87.3	66.7	96.0	79.1	81.0	62.4
8	Selection-Grammar	480.0	85.4	90.8	66.3	95.6	77.5	81.1	64.6
9	Search-By-Grammar	436 / 42	87.6	91.2	74.3	97.1	79.4	84.3	65.9
10	Ambiguous-Channel-Grammar	252 / 24	92.8	94.2	79.5	96.7	86.7	92.5	77.9
11	Settings-Grammar	216.0	96.0	98.0	85.2	98.0	90.0	91.0	83.3
12	User-Id-Grammar	108.0	97.3	100.0	85.2	100.0	93.3	98.0	88.9
13	Not-On-The-Air-Grammar	120.0	98.9	98.3	86.7	100.0	96.7	95.0	88.3
14	Help-Grammar	104.0	98.7	100.0	88.5	96.2	96.2	100.0	84.6

			e1600012 - Base Acoustic Model						
			Close-Talking Mic				Far Field Mic		
	Grammar	Counts	Amuser	EB	FBK	YKT	Amuser	EB	FBK
4	Initial-Ok-Grammar	68 / 64	100.0	100.0	100.0	100.0	100.0	100.0	100.0
5	On-Grammar	116.0	98.8	100.0	94.8	96.3	91.4	83.3	87.9
6	TV-Screen-Grammar	296.0	95.1	97.1	81.1	97.1	88.7	89.7	74.3
7	EPG-Grammar	524 / 51	86.5	92.1	64.3	93.7	77.5	78.6	56.2
8	Selection-Grammar	480.0	86.5	90.8	69.6	94.7	81.6	80.3	60.0
9	Search-By-Grammar	436 / 42	89.9	92.2	72.4	94.1	82.0	81.9	60.7
10	Ambiguous-Channel-Grammar	252 / 24	93.9	96.7	73.8	95.0	86.1	90.0	71.3
11	Settings-Grammar	216.0	98.0	98.0	86.1	98.0	90.0	85.0	80.6
12	User-Id-Grammar	108.0	96.0	100.0	88.9	100.0	93.3	98.0	83.3
13	Not-On-The-Air-Grammar	120.0	98.9	96.7	86.7	100.0	93.3	90.0	81.7
14	Help-Grammar	104.0	98.7	100.0	84.6	96.2	94.9	90.4	92.3

Road to the Final Prototype

- Optimize echo cancellation and speech detection
  - Better rejection of false triggers by TTS prompts
- Carry out automated, consistent and reproducible tests of the full system across all labs
- Enhancements to dialog design, ASR, and NLU
  - Improved rejection and error recovery

Thank you!

Backup

User Interface Characteristics of Voice Interaction

Richness and flexibility of requests

- Grammar based command and control & parameter input
- Free form for command and control

- Free form with multiple token extraction from voice input
  - "I'd like to hear the Fifth Symphony by Beethoven"
  - Take me to the city center of *Boston*"

Richness and flexibility of the interaction

- Menu driven (Directed Dialog)
  - Pre-scripted call flow, tedious hierarchical menus
  - Specific commands must be used at specific times
  - Correcting errors typically requires restarting

- Conversational Interaction
  - No need to remember specific commands, direct access to most functions
  - User can correct, even after several turns, without having to always start over
  - Extensive disambiguation
  - System can remember context and past history for more intuitive & personal interaction