



SpeechWorks

The Art and Science of Spoken Dialog Systems

from Research to Industry

Roberto Pieraccini



SpeechWorks International

17 State Street, New York, NY

SpeechWorks

Talk Overview

Yesterday

- IVR / Touch Tone

Today

- Spoken Dialog Systems
 - The technology
 - The Art

Tomorrow

- Multimodal / Wireless

What we learned

Yesterday

- IVR / Touch Tone

Yesterday

- IVR / Touch Tone

Today

- Spoken Dialog Systems

Spoken Dialog Industry

- Around 1994
- SpeechWorks, Nuance (US), Philips (Europe)
- Goal: replace call centers and IVR with automatic dialog systems
- Business case: cost reduction, customer retention

Communications

■ Auto Attendant

- System: “Please say the name of the person . . .”
- Caller: “Karen Foye”
- System: “Transferring to Karen Foye . . .”

Try it at:

(617) 428 4444

(212) 425 7600



Information Retrieval

- **Product Information**

- System: “What is the name of your Hewlett Packard product?”
- Caller: “Office Jet LX
- System: “Your product is...”

- **Examples**

- UPS, United, HP, American, Amtrak



Information Retrieval

- **Customer Support**

- System: “What type of package are you shipping; to where?”
- Caller: “A letter from 02111 to 90034”

- **Examples**

- **United Airlines, FedEx, American, Guardian 401(k)**



Try it at:

1-800-GO_FEDEX (option 5)

Transactions

- **Speech-enhanced E-Commerce**

- System: “Would you like to place an order?”
- Caller: “Buy 100 shares at limit price”

- **Examples**

- **E*TRADE, Discover Brokerage, TD Waterhouse Australia, Stock exchange of Singapore and more**



Travel

- **United Airlines employee flight reservation**



- **United Airlines flight information system**

Try it at:

1 800 824 6200 (United Airlines)

1 800 896 7317 (Continental)

1-800-THRIFTY, option 1-1

DEMO

Medical Services

- **Mid-Atlantic Medical Services**



Yesterday

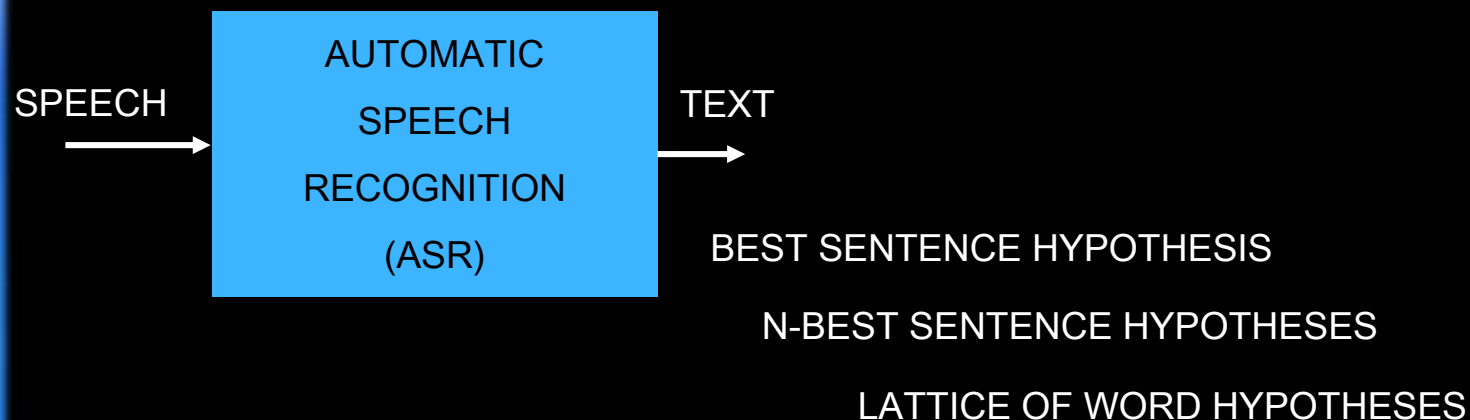
- IVR / Touch Tone

Today

- Spoken Dialog Systems

- **The technology**

The Speech Recognition Component



WORD ACCURACY:
$$\frac{\text{NUMBER OF CORRECT WORDS}}{\text{TOTAL NUMBER OF WORDS}} \times 100$$

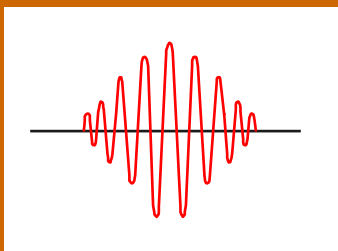
Speech Recognition 101

Caller speaks
the word
"SpeechWorks"

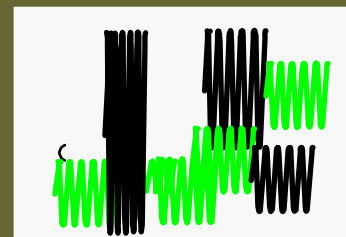


Telephone
Network

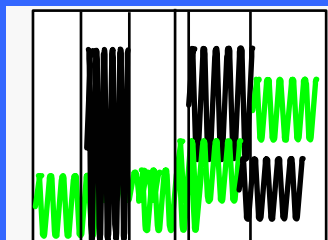
Capture &
Digitization



Spectral
Representation



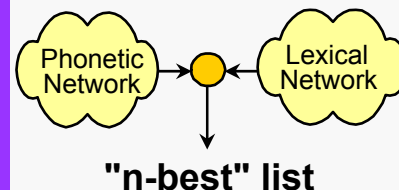
Segmentation



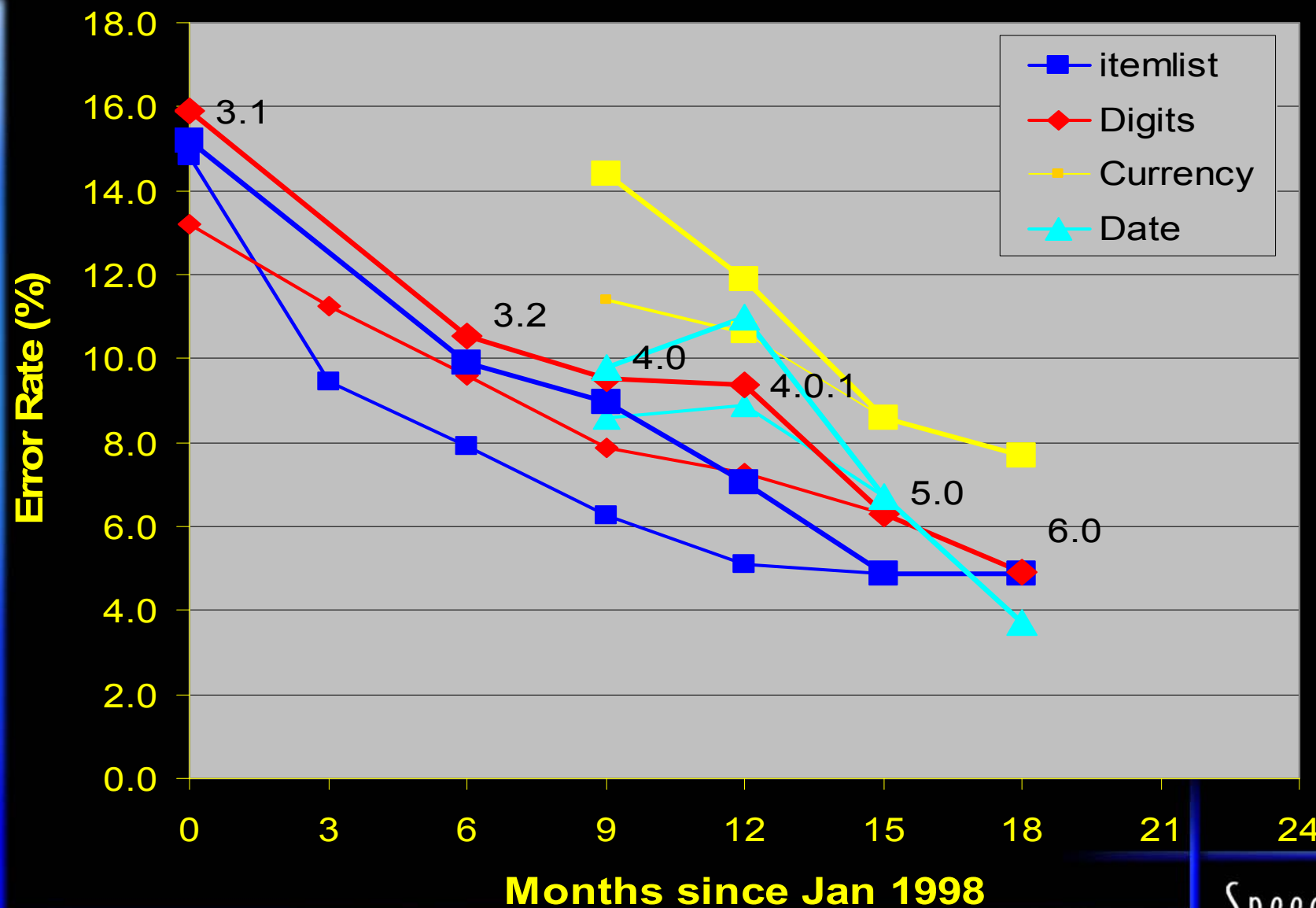
Phonetic
Classification

<u>Phoneme Prob.</u>		
Sound Segment	ao	.92
	b	.22
	ae	.43
	eh	.32
	aw	.51

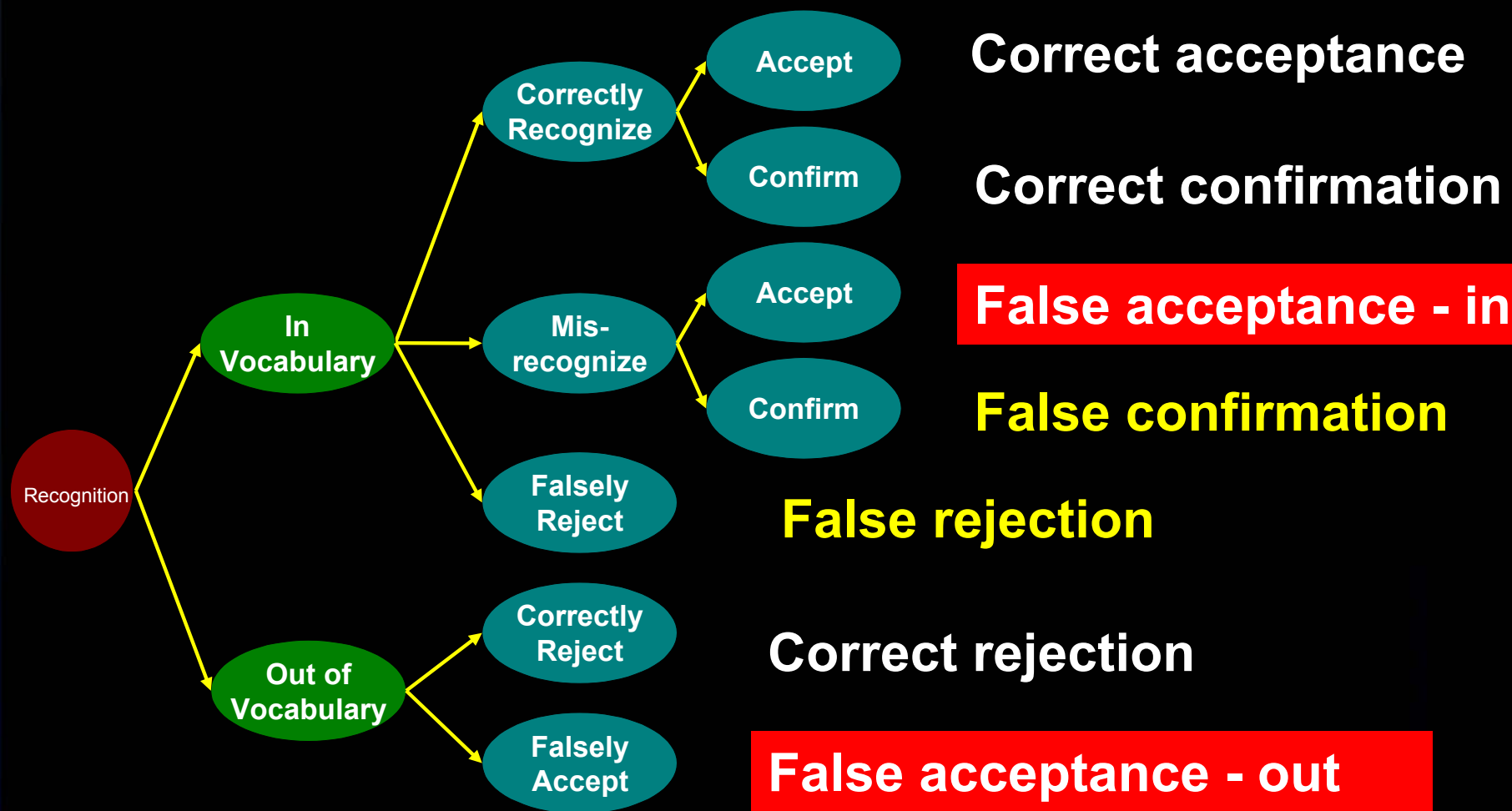
Search & Match



Incremental Improvement



Recognition Accuracy in Dialog Systems



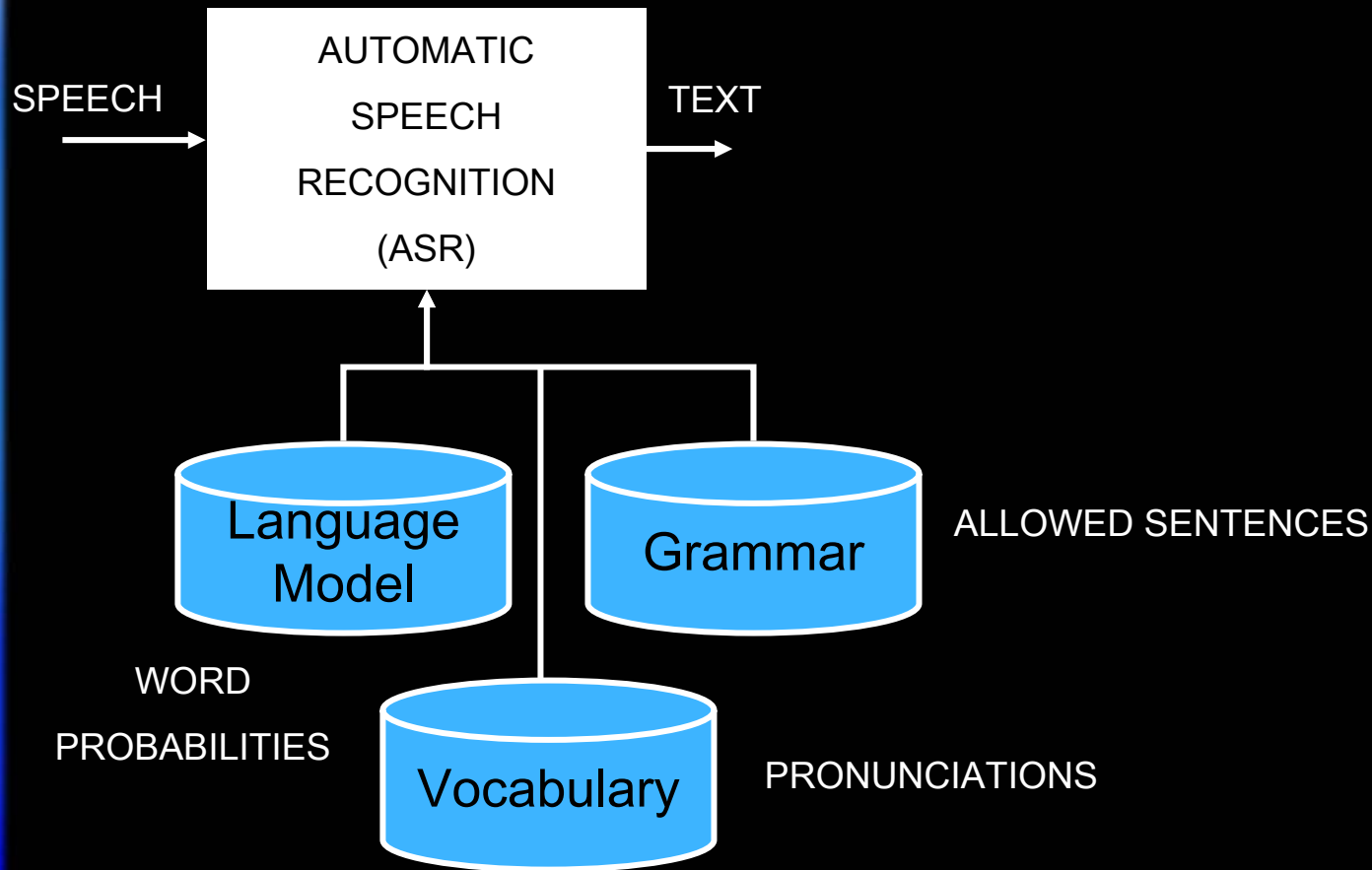
Overall Recognition Accuracy

IN VOCABULARY	
Raw Recognition	95.8
Correct Accept	92
Correct Accept + Correct Confirm	94.2
False Accept	1.8
False Reject	3
False Confirm	1.1
OUT OF VOCABULARY	
Correct Reject	42.5
False Accept	47.8
Out of vocabulary %	16.6
TOTAL NUMBER OF UTT.	8249

Detailed Recognition Accuracy per Dialog State

MODULE	RR_IN	CA_IN	CA+CC	FA_IN	FR_IN	FC_IN	CROUT	FAUT	%OUT	TOTAL
ALTECH_YES_NO_CONF	0.987	0.975	0.975	0.004	0.021	0	0.68	0.32	4.6	546
AMPM_DISAMBIG	1	0.923	1	0	0	0	1	0	23.5	17
ARRIVAL	0.961	0.955	0.961	0.022	0.017	0	0.308	0.462	6.8	191
CABIN_FIRST_COACH	0.964	0.928	0.942	0.014	0.043	0	0.75	0.25	2.8	142
CHANGE	0.898	0.88	0.892	0.036	0.06	0.012	0.583	0.333	6.7	178
CHECK_CONTINUE	0.923	0.923	0.923	0	0	0.077	1	0	7.1	14
CONFIRM_CANCEL	1	1	1	0	0	0	1	0	25	4
CONF_ITIN	0.988	0.972	0.978	0.005	0.014	0.003	0.722	0.083	5.3	682
DATE	0.934	0.881	0.928	0.024	0.019	0.028	0.409	0.364	7.1	617
DEPART	0.904	0.881	0.893	0.034	0.056	0.017	0.278	0.611	9.2	195
DESCRIBE_OPTIONS	0.988	0.953	0.988	0	0.012	0	0.68	0.32	22.5	111
DESCRIBE_OPTIONS_L	0.969	0.917	0.931	0.007	0.055	0.007	0.64	0.314	17	507
DISAMB_AIRPORT	0.964	0.945	0.964	0	0.036	0	0.538	0.385	19.1	68
DTMF_HELP	1	1	1	0	0	0	1	0	18.8	16
EMPLOYEE_ID	0.897	0.813	0.879	0.043	0.026	0.052	0.5	0.242	26.3	472
FAX_NUM	0.933	0.933	0.933	0	0.067	0	0.5	0.5	21.1	19
FLIGHTNO	0.886	0.831	0.871	0.075	0.031	0.024	0.308	0.615	7.9	495
FLIGHTNO_AIRPORT	1	1	1	0	0	0	0.334	0.639	99.3	307
FLIGHTNO_OPTIONS	0.967	0.895	0.909	0	0.087	0.004	0.656	0.297	18.9	339
GET_ACTION_SUMMARY	1	1	1	0	0	0	0.6	0.4	26.3	19
INSTRUCTIONS	0.996	0.969	0.993	0.002	0.002	0.002	0.667	0.056	7.4	484
MENU	0.958	0.925	0.943	0.015	0.033	0.01	0.465	0.416	16.2	623
MORE_SEGMENTS	0.983	0.961	0.966	0	0.028	0.006	0.5	0.25	2.2	183
NA_SPELL_LAST	1	1	1	0	0	0	1	0	60	5
NUM_KNOWN	0.981	0.972	0.972	0.019	0.009	0	0.333	0.667	2.8	109
NUM_TRAV	0.981	0.93	0.956	0.006	0.038	0	0.857	0	4.2	165
PHONE_NUM	0.922	0.872	0.908	0.035	0.043	0.014	0.182	0.636	7.2	152
PIN	0.974	0.957	0.967	0.01	0.013	0.01	0.385	0.462	25.4	409
RELATIONSHIP	0.977	0.938	0.968	0.006	0.023	0.003	0.368	0.474	5.3	360
RETRY	0.958	0.875	0.917	0	0.083	0	1	0	20	30
TIME	0.932	0.848	0.879	0.061	0.061	0	0.364	0.364	7.7	143
TRAVELER	1	1	1	0	0	0	0.16	0.81	97.6	205
WANTS_FAX	0.979	0.916	0.958	0.014	0.028	0	0.75	0.25	2.7	147
WANTS_REPEAT	0.971	0.918	0.941	0	0.035	0.024	0.65	0.1	10.5	190

Vocabularies, Grammars and Language Models



BNF grammars

:BNF

ROOT = \$CallCommand;

\$CallCommand = \$StandardCallCommand;

\$CallCommand = \$PoliteCallCommand;

\$PoliteCallCommand = please \$StandardCallCommand;

\$StandardCallCommand = \$DirectMyCalls \$Where;

\$DirectMyCalls = \$Direct my calls;

\$DirectMyCalls = \$Direct calls;

\$Direct = direct;

\$Direct = send;

\$Where = \$Home;

\$Where = \$Car;

\$Where = \$Work;

\$Home = home;

\$Home = to my home;

\$Car = to my car;

\$Car = to my car phone;

\$Work = to the office;

XML grammars – W3C standard

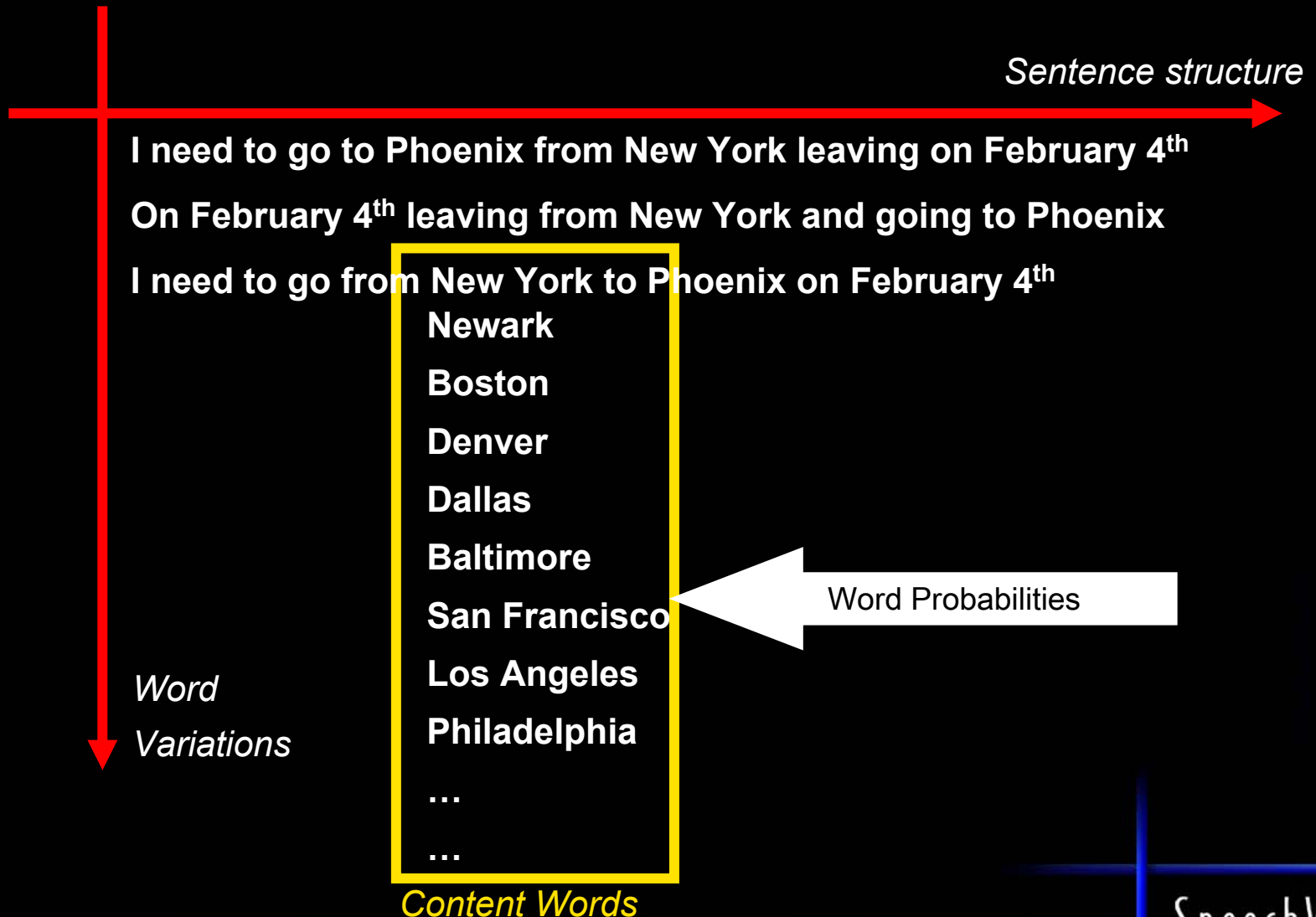
```
<?xml version="1.0"?>

<grammar xml:lang="en-US" version="1.0" root="ROOT">
  <rule id="ROOT" scope="public">
    <item> <ruleref uri="#CallCommand"/> </item>
  </rule>
  <rule id="CallCommand">
    <one-of>
      <item> <ruleref uri="#PoliteCallCommand"/> </item>
      <item> <ruleref uri="#StandardCallCommand"/> </item>
    </one-of>
  </rule>
  <rule id="PoliteCallCommand">
    please
    <ruleref uri="#StandardCallCommand"/>
  </rule>
  <rule id="StandardCallCommand">
    <ruleref uri="#DirectMyCalls"/>
    <ruleref uri="#Where"/>
  </rule>
  <rule id="DirectMyCalls">
    <one-of>
      <item> <ruleref uri="#Direct"/> calls </item>
      <item> <ruleref uri="#Direct"/> my calls </item>
    </one-of>
  </rule>
</grammar>
```

Pronunciations

anesthesiology	ae n ix s t iy z iy aa l ax jh iy
conductor	k ix n d ah k t er
cooking	k uh k ix ng
deposit	d ix p aa z ih t
dynasty	d ay n ix s t iy
ireland	ay r l ax n d
illinois	ih l ax n oy
ohio	ow hh ay ow
psychology	s ay k aa l ax jh iy
public	p ah b l ix k
southwestern	s aw th w eh s t er n
sports	s p ao r t s
sweetheart	s w iy t hh aa r t
texarkana	t eh k s aa r k ae n ax

Language Models



No data

- Estimation of word probabilities without a corpus has to rely on other sources
- Examples:
 - Airport probabilities derived from flight traffic statistics
 - ~ An airport with more traffic or geographically closer to the user is more likely to be spoken in a flight information application
 - Street name probabilities derived from number of buildings on the streets
 - ~ A street with more 'people' is more likely to be spoken in a driving direction application

$$w = (\# \text{ of business}) * 5 + (\# \text{ of high-rises}) * 10 + \# \text{ of residential}$$

Problems with content words

- **Abbreviations, acronyms**

- 14" display w/ anti-glir scrn
- Dr. -> *Drive/Doctor* St -> *Saint/Street*
~ St. John St.

- **Linguistic variations (synonyms, phrase structure, ...)**

A fourteen inches display with anti-glare screen

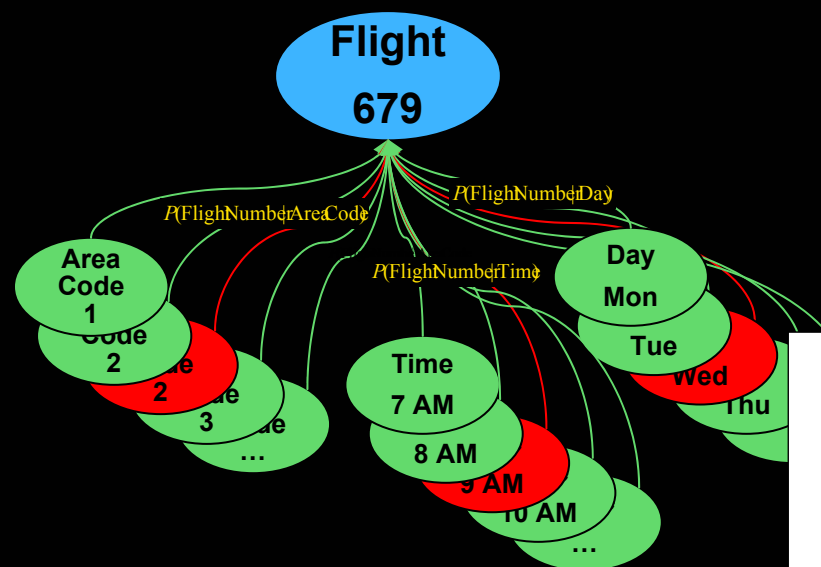
A display of fourteen inches size with an anti-reflection screen

- **Ambiguity**

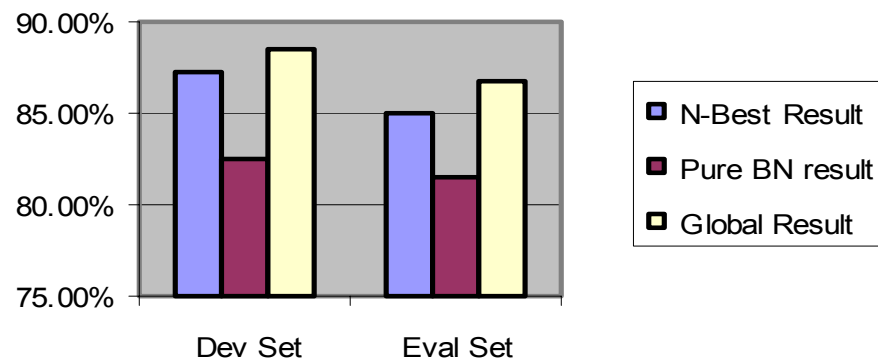
125 twenty seventh street

East or West?

Dynamic Semantic Models

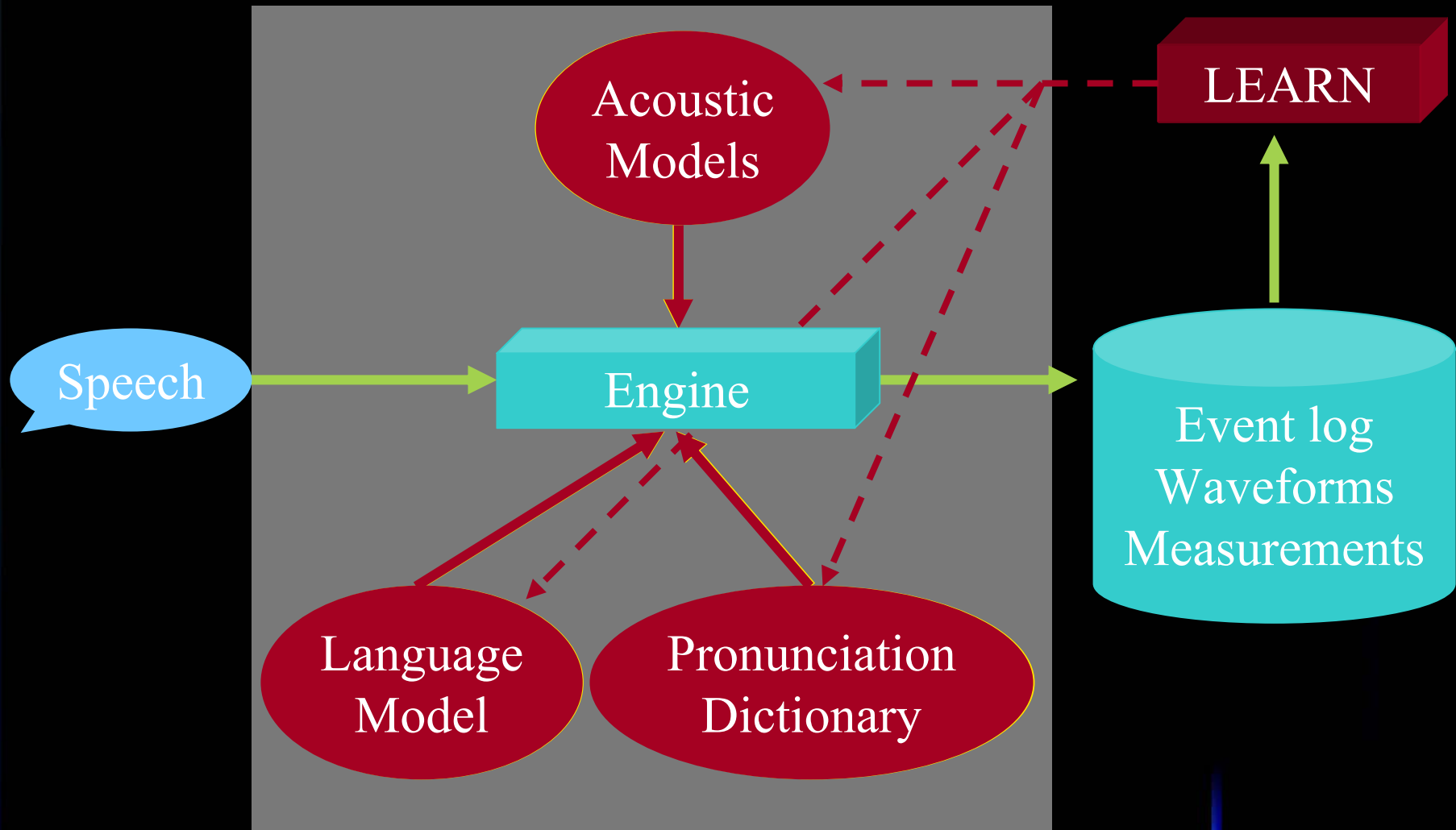


Flight Number Identification

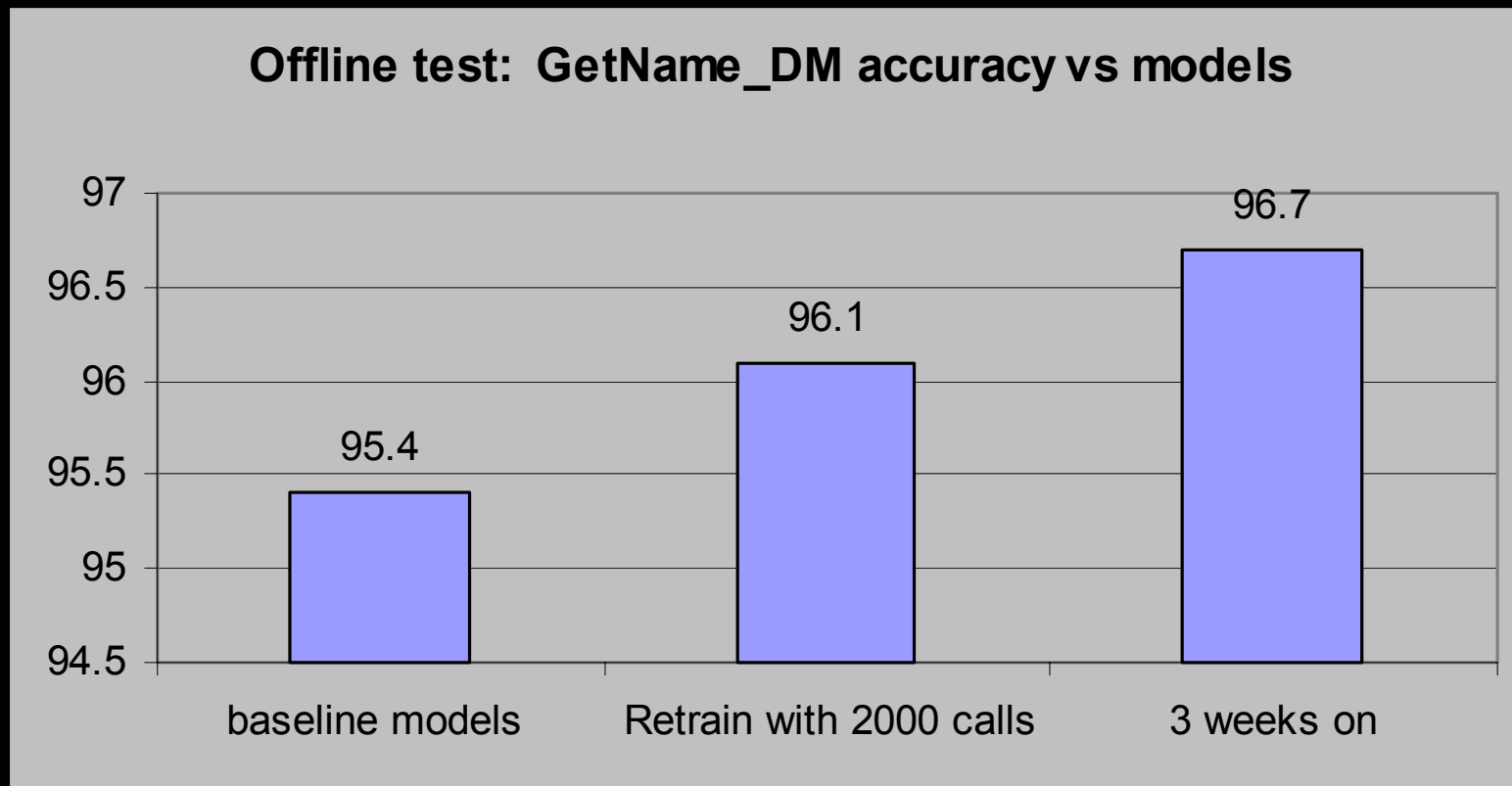


From: Wai, C., Pieraccini, R., Meng, H., "A Dynamic Semantic Model for Rescoring Recognition Hypothesis," Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2001

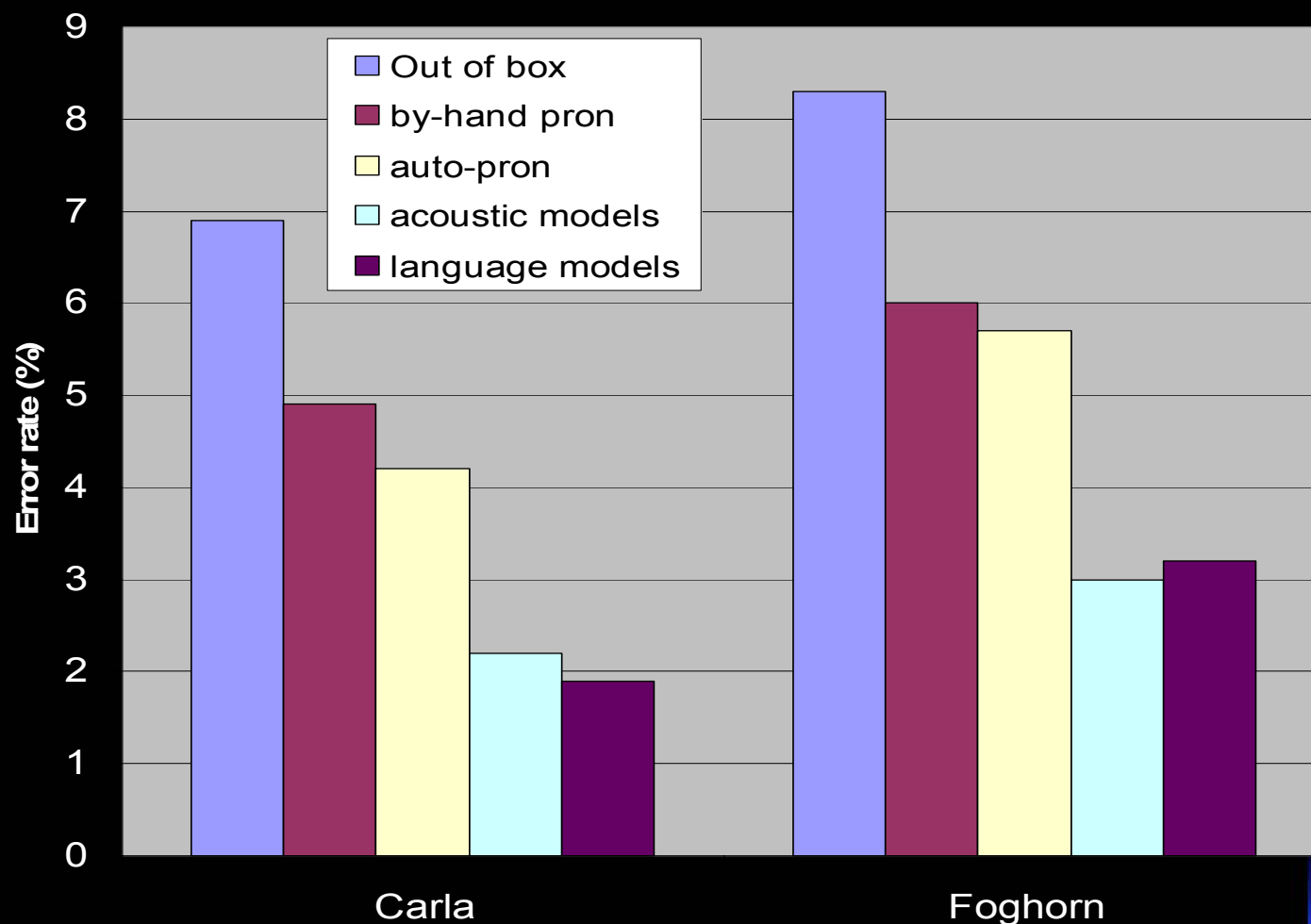
Automatic Tuning (LEARN)



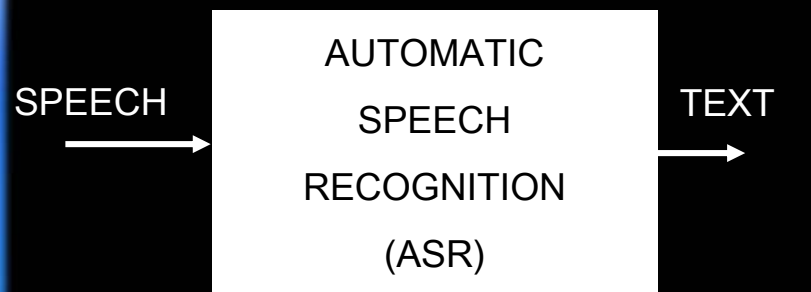
Automatic Tuning (LEARN) performance



Automatic Tuning (LEARN) Performance

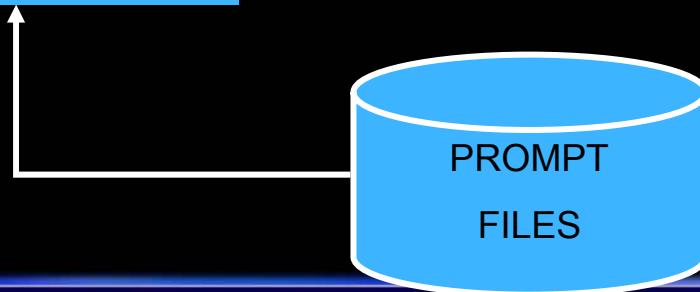
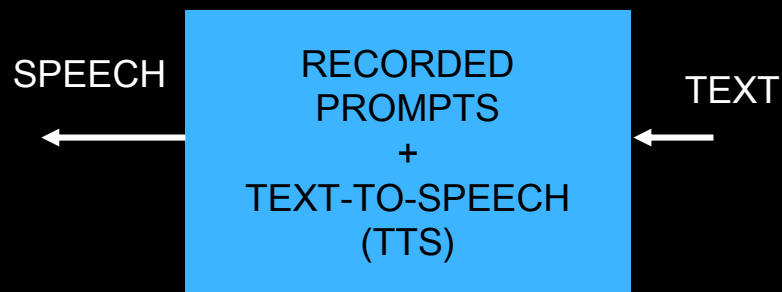


The Output Component



TEXT TO SPEECH

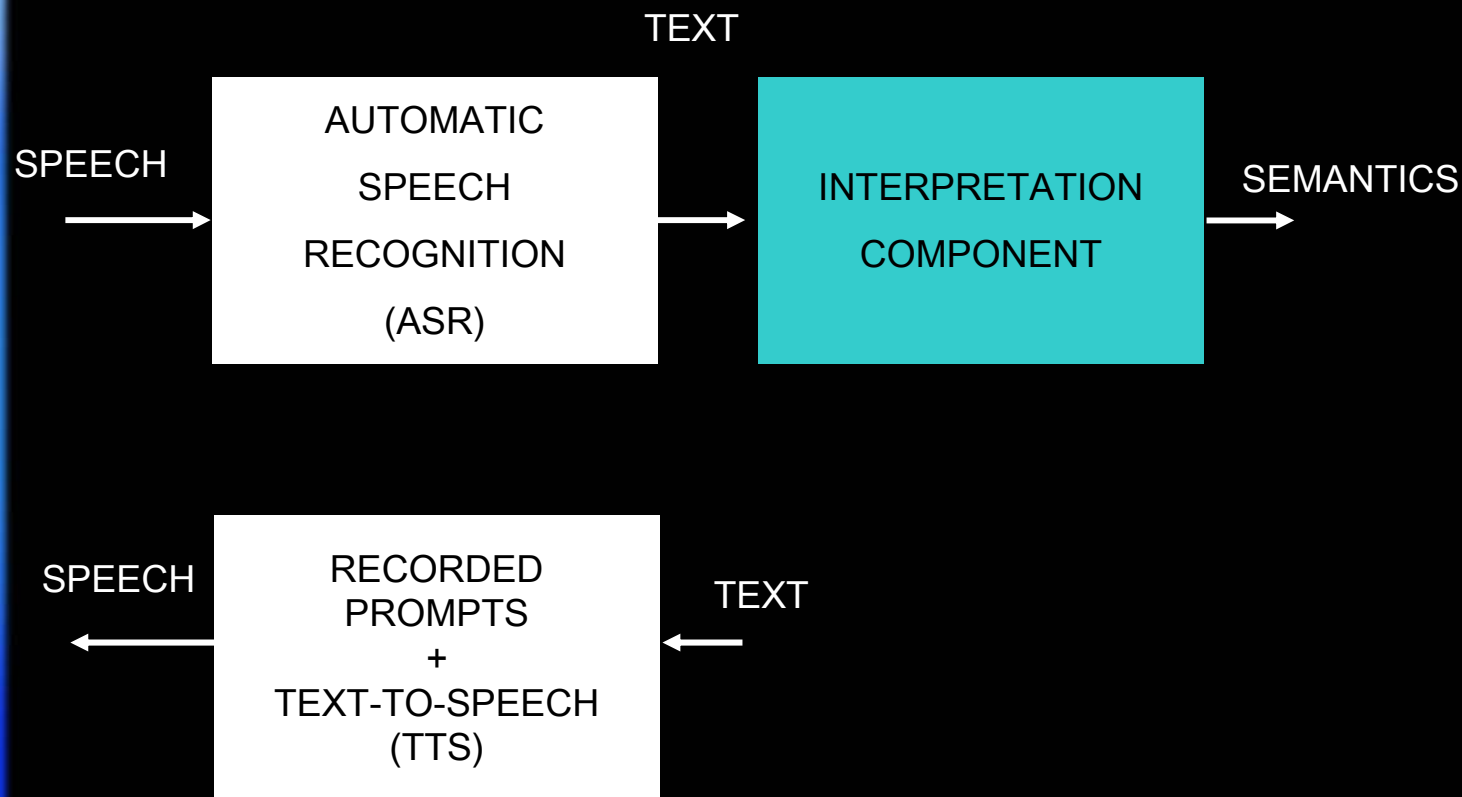
RECORDED PROMPT



Mixing TTS and Prompts

Welcome to the SpeechWorks flight information line. I can give you up to the minute arrival and departure information for all flights. Please tell me which flight you need information about. If you don't know the number for the flight you're interested in, say "I don't know it", so we can look it up some other way. Otherwise, please enter or say the flight number. For example, you could say "1-6-0-9".

From Words To Concepts



Functions of the interpretation component

- Normalization

Kennedy Airport

J F K

Kennedy International Airport

J F K Airport

New York J F K



JFK


Functions of the interpretation component

- Normalization
- Improve accuracy, confidence and efficiency of the recognizer

N	RECOGNITION RESULT	INTERPRETED RESULT
1	direct my calls to my car phone	direct calls to car
2	direct calls to my car	
3	please direct my calls to my car	
4	send calls home	direct calls to home
5	direct calls to my home	direct calls to office
6	send my calls to the office	

Functions of the interpretation component

- Normalization
- Improve accuracy, confidence and efficiency of the recognizer
- Extract semantics

Flying from san francisco to new york on may 12th  ORIGIN: SFO
DESTINATION: NYC
DATE: 05/12

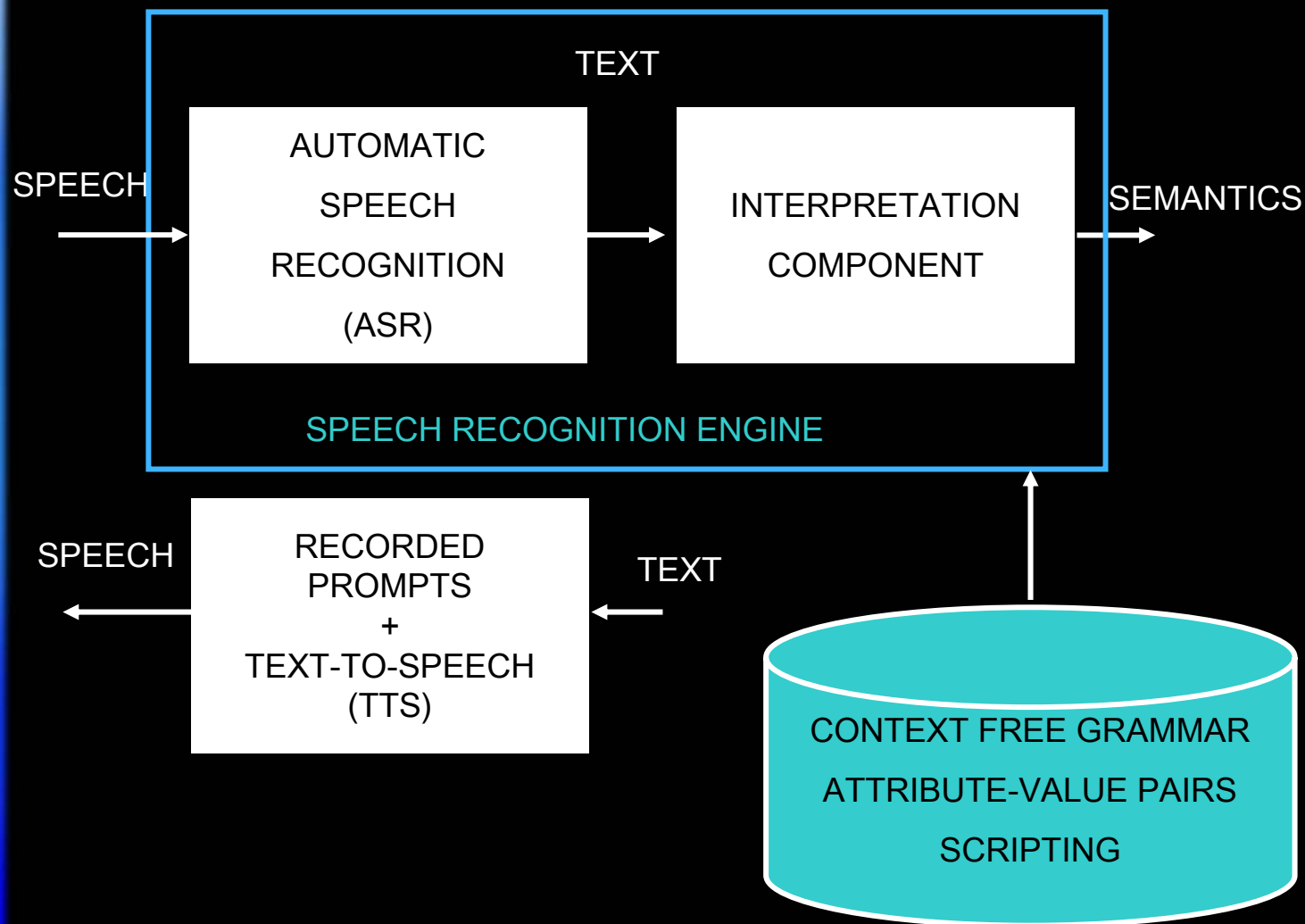
ATTRIBUTE-VALUE PAIRS

Functions of the interpretation component

- Normalization
- Improve accuracy, confidence and efficiency of the recognizer
- Extract semantics
- Integrate domain knowledge



From Words To Concepts



Augmented Grammars

\$ITINERARY = \$FROM \$TO;

\$FROM = from \$AIRPORT;

\$TO = to \$AIRPORT;

\$AIRPORT = [new york] (J F K)(kennedy) [airport];

\$AIRPORT = (boston | logan) [airport];

Augmented Grammars

\$ITINERARY = \$FROM \$TO;

<script>

origin = \$FROM.VALUE;

destination = \$FROM.VALUE;

INVALIDATE = \$origin == \$destination;

</script>

\$FROM = from \$AIRPORT;

<script>

VALUE = \$AIRPORT.VALUE;

</script>

\$TO = to \$AIRPORT;

<script>

VALUE = \$AIRPORT.VALUE;

</script>

\$AIRPORT = [new york] (J F K)(kennedy) [airport];

<script>

VALUE = JFK;

</script>

\$AIRPORT = (boston | logan) [airport];

<script>

VALUE = BOS;

</script>

Do we need natural language?

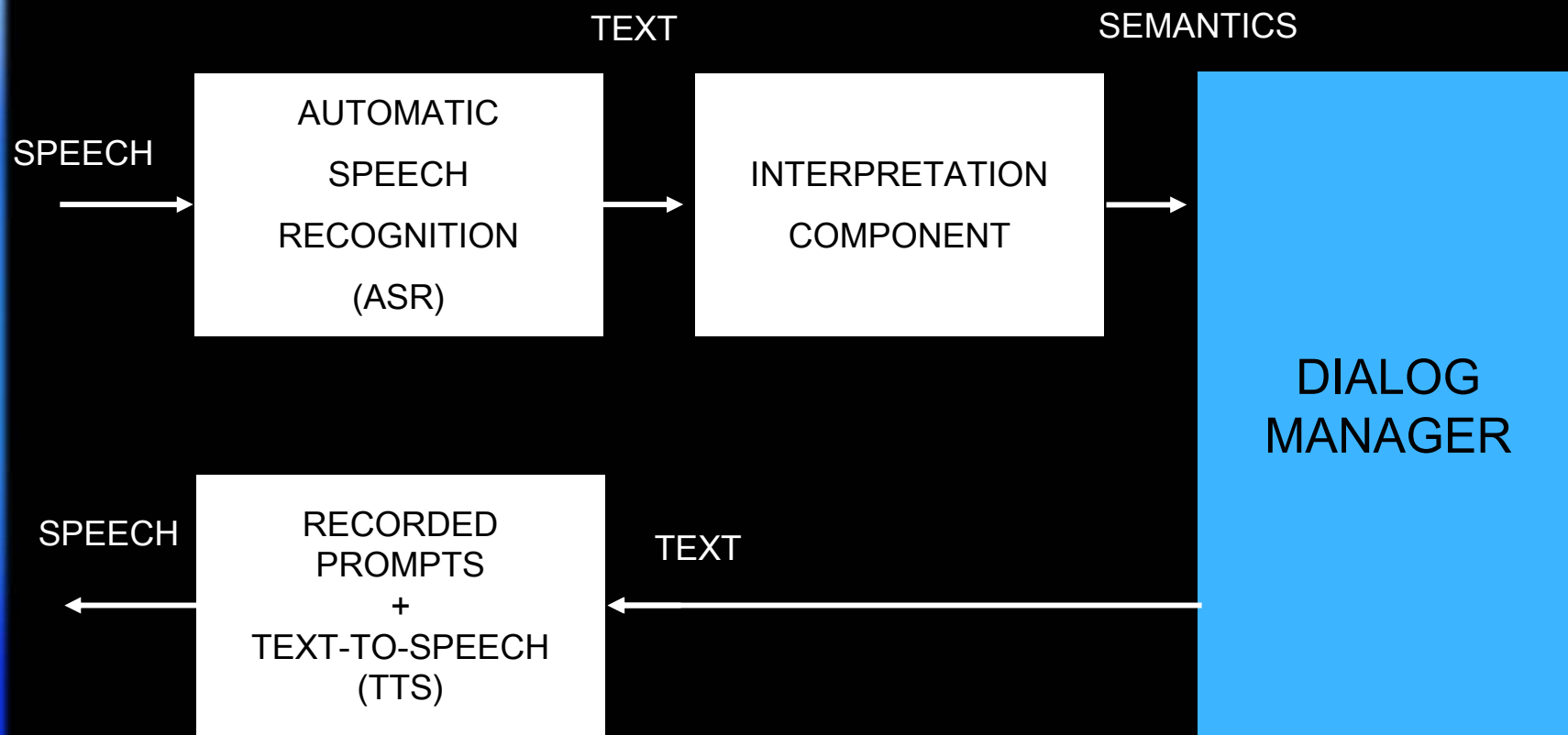
I would like a large thin crust pizza with pepperoni and mushrooms and two bottles of soda.

- Users are acquainted with business logic
- Higher performance of structured systems
- New users appreciate system guidance

I don't know, but my TV seems pretty much dead and black smoke is coming out of it, I also hear a buzzing kind of sound.

- Users do not know the business logic
- A structured system is impractical

Dialog Management



Discourse Management

- **Timeouts**

- ~ *I didn't hear you, please try again.*

- **Low confidence**

- ~ *I didn't understand.*

- **Confirmation**

- ~ *Did you say IBM?*

- **Preventing repeated mistakes**

- ~ *Did you say Boston? ... Did you say Boston?*

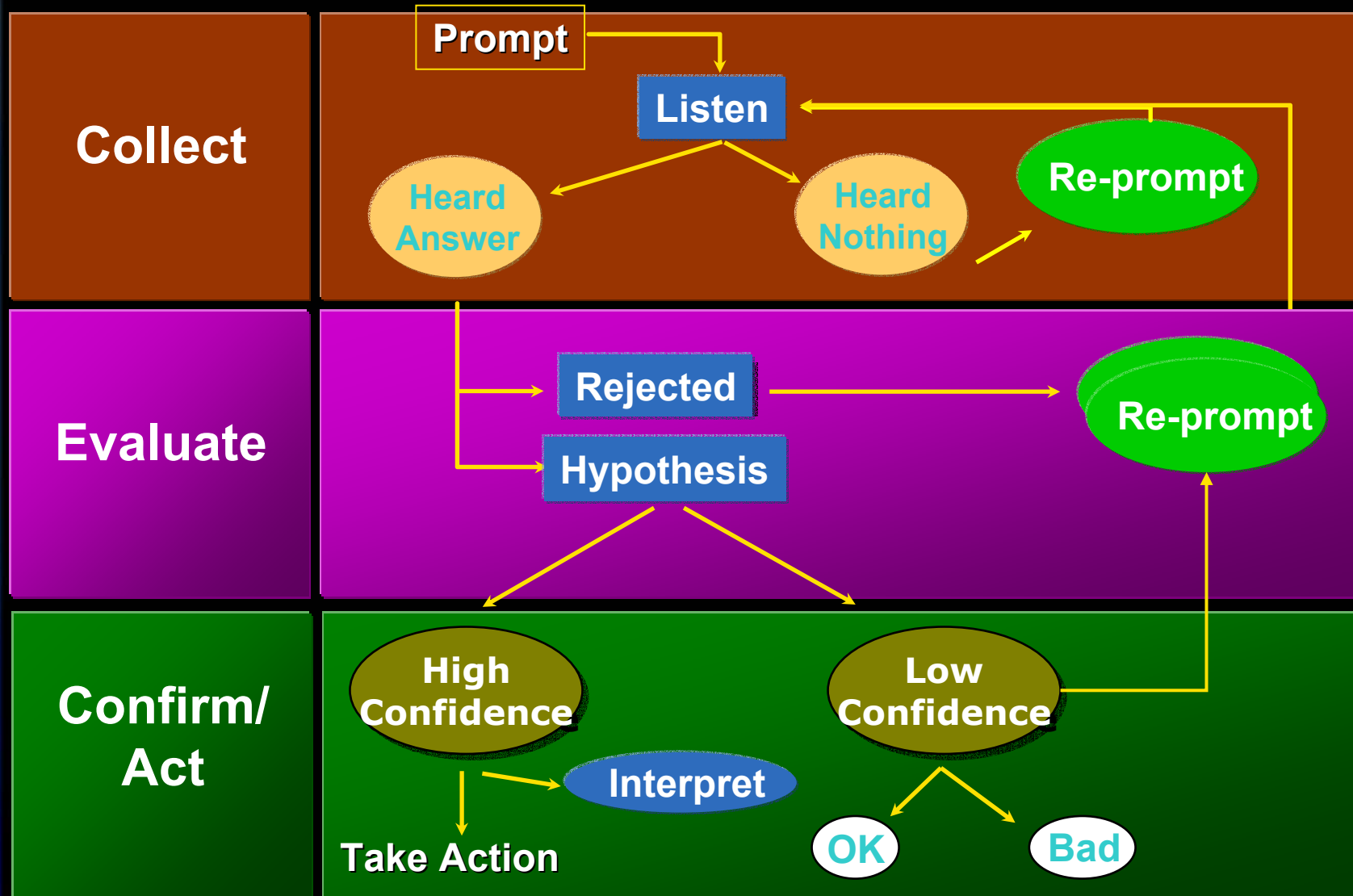
- **Fallback**

- ~ *I still can't understand. Please spell that.*

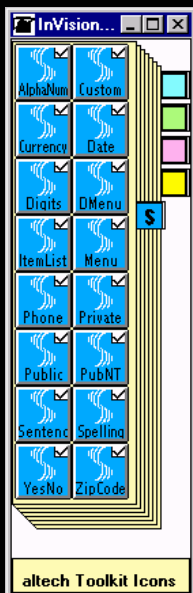
- **Disambiguation**

- ~ *Which Steve did you mean?*

Dialog Modules



Dialog Module Types



Address

Speaker Verification

Voice Enrollment

"Plug & Play"

Yes/No

Digits

Alphanum

Voice Menu

Item List

Spelling

Phone

Zip Code

Soc. Security #

Credit Card #

Basic w/ NL Phrasing

Natural Number

Currency

Date

CC Expir. Date

Time

Name

Natural Language

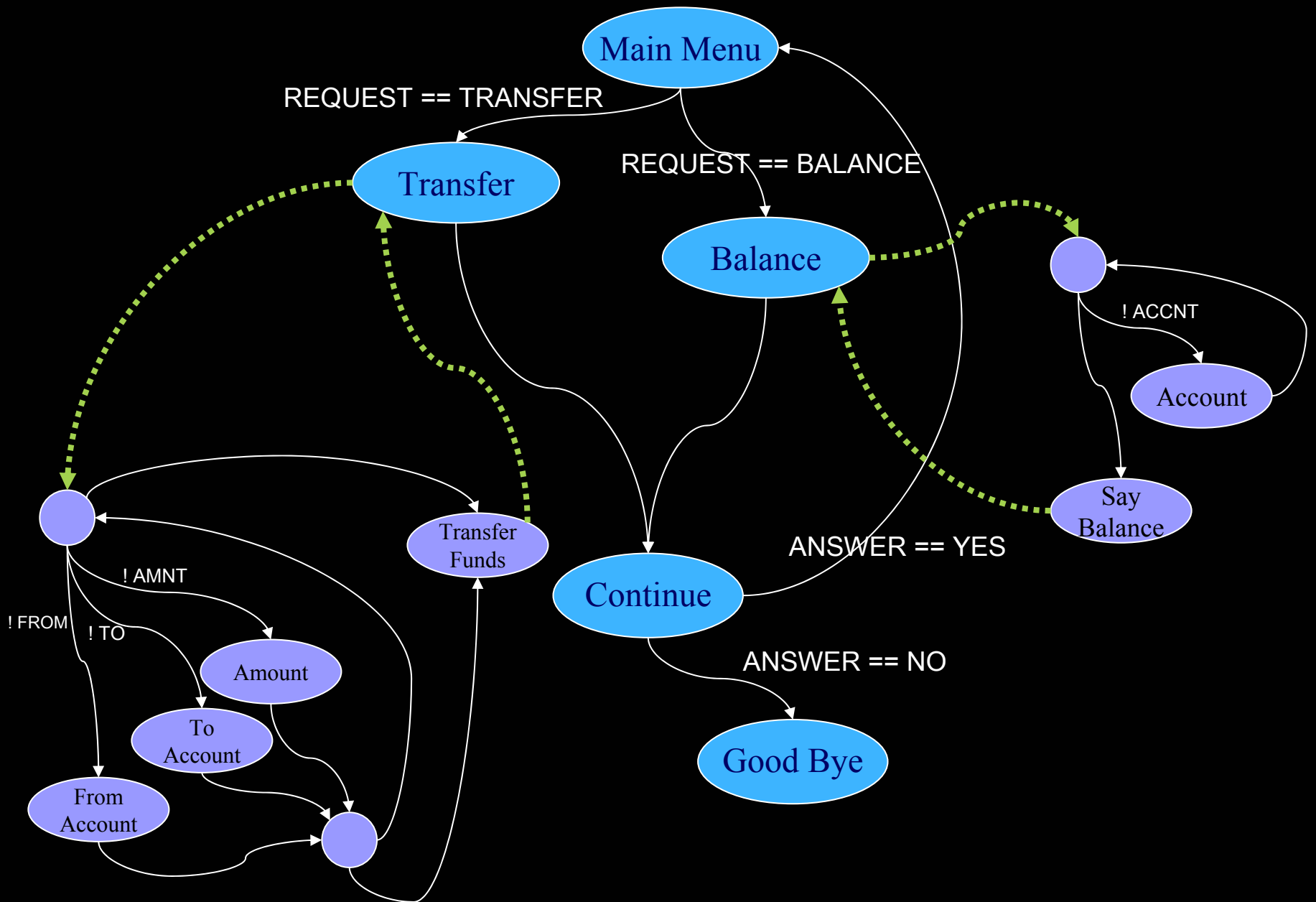
Custom Context

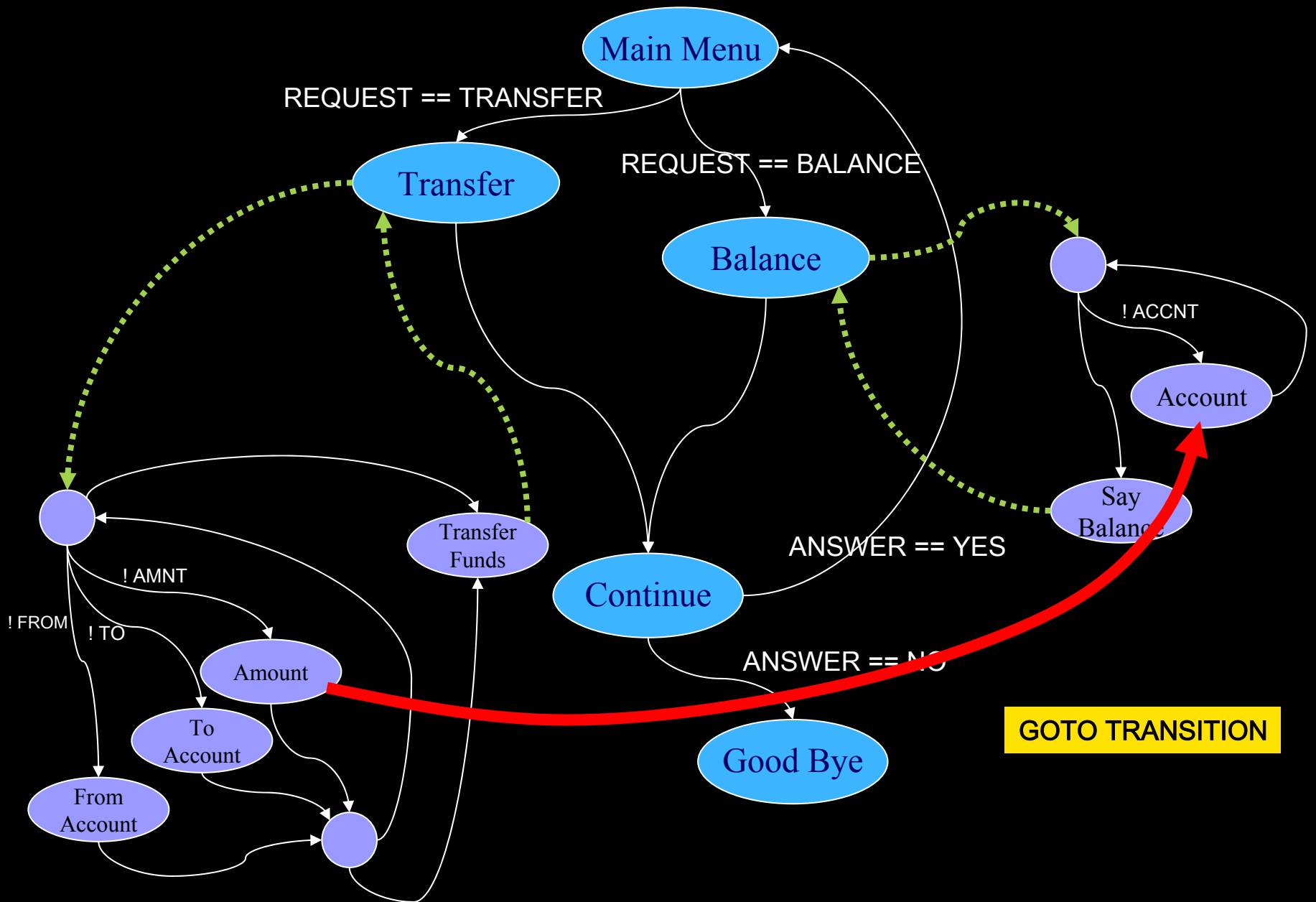
**User-Defined
Natural Language**

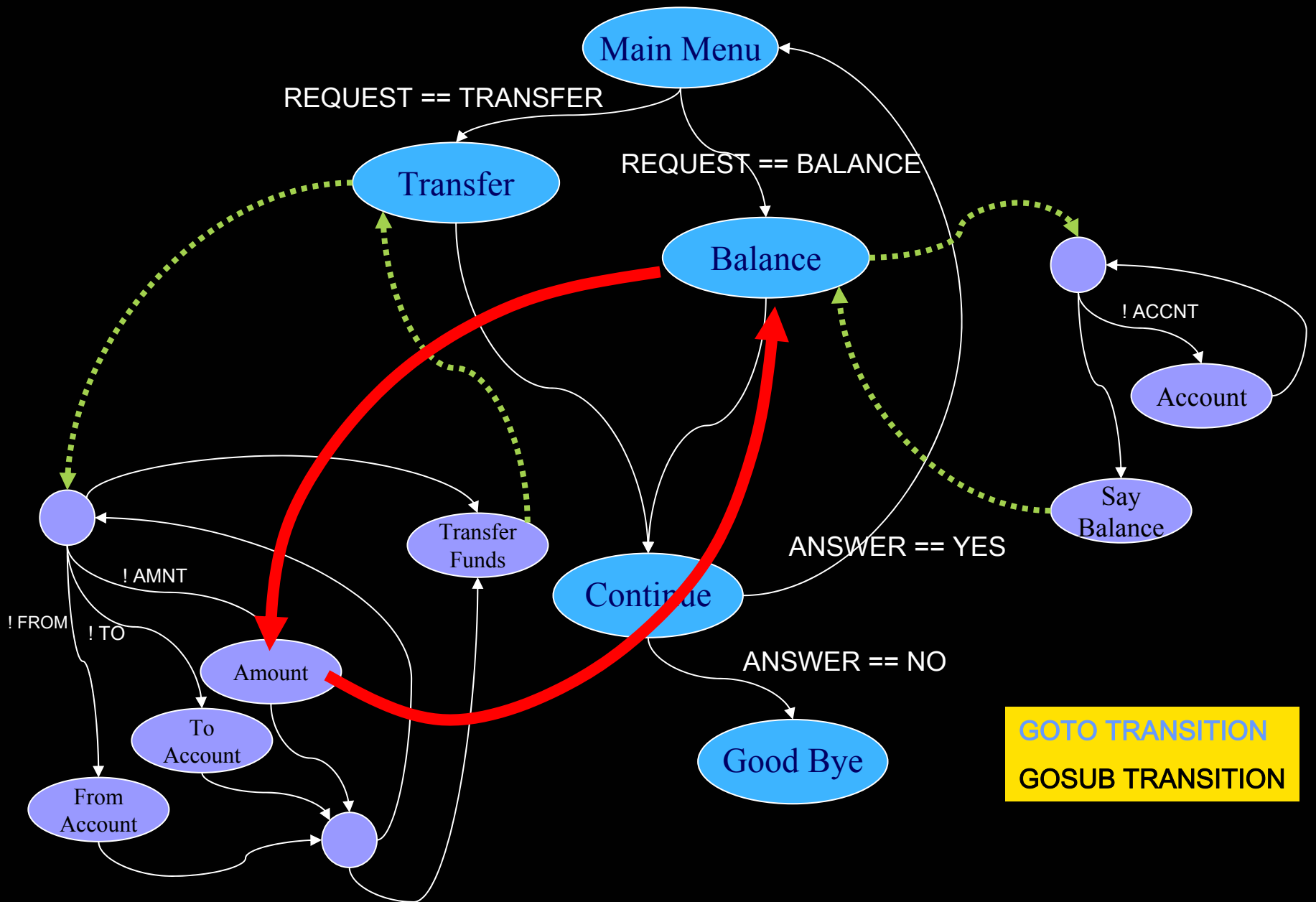
High Level Dialog Representation (call flow)

- **Recursive Transition Network Representation**
 - Modularization, Sub-dialogs
- **Nodes and Actions**
- **Transitions and Conditions**
- **Frame representation of Dialog State**

TYPE_OF_TRANSACTION: "TRANSFER"
ORIGINATING: "CHECKING"
DESTINATION: "SAVINGS"
AMOUNT: 500.25
DATE: "10-02-2001"
USER_DATA:
 PERSONAL:
 ID: "124578"
 SSN: "540 00 1234"
STARTING_DATE: "02-02-1998"
CURRENT: true
ACCOUNT_NUMBER: 3
CHECKING_BALANCE: 1245.55





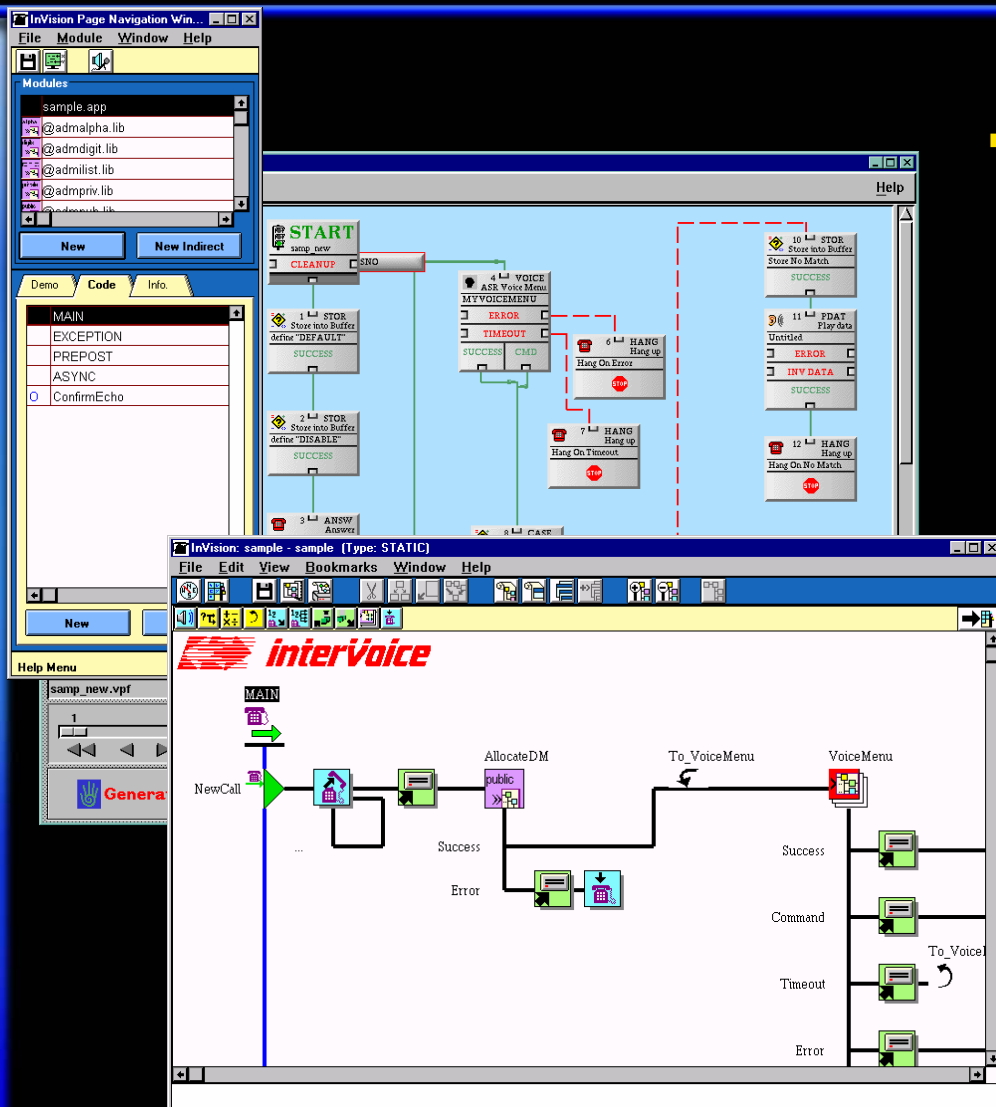


GOTO TRANSITION
GOSUB TRANSITION

Logic included in the dialog manager

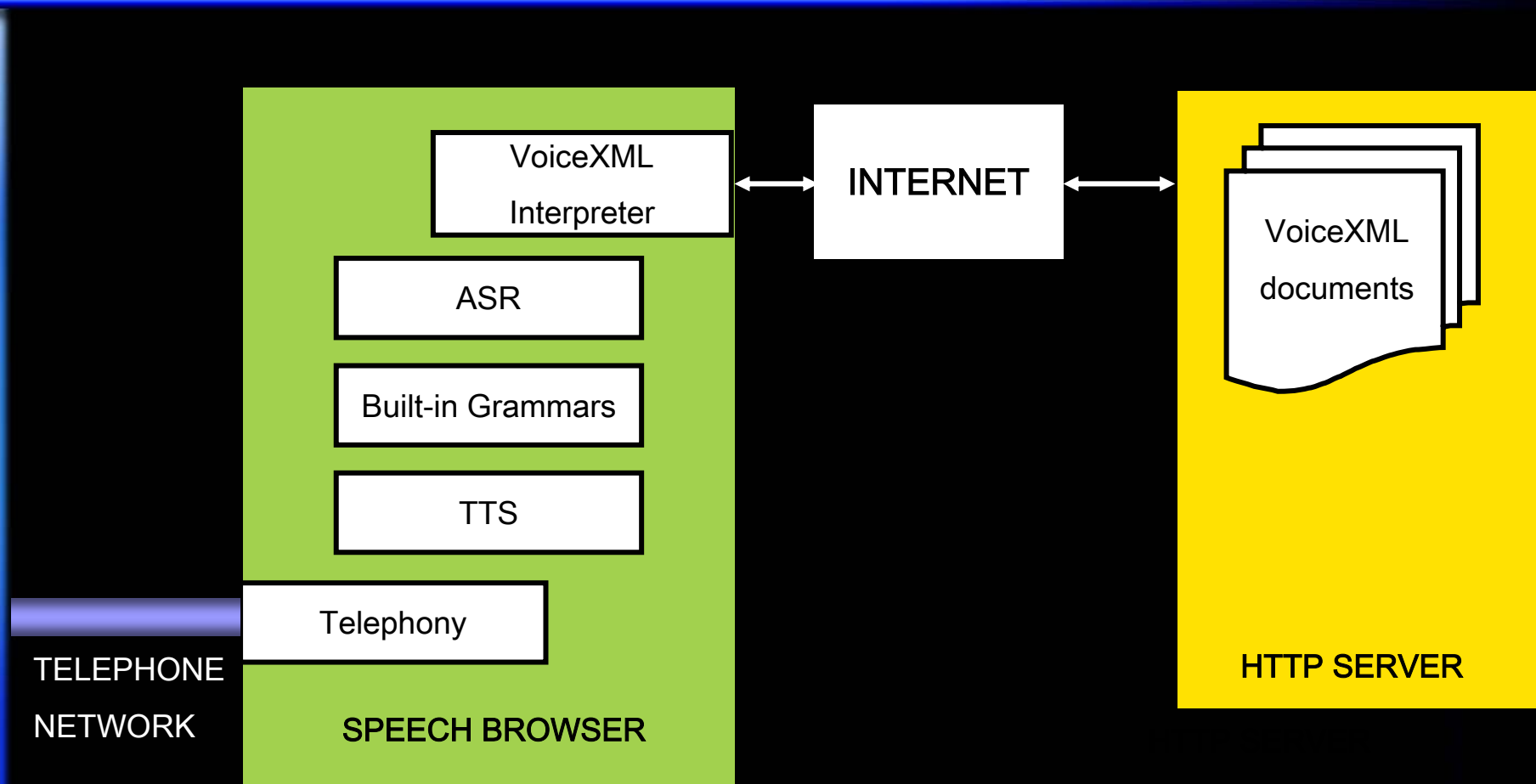
- **UI Universals**
 - Backup, Cancel, Repeat, Help, Main Menu, ...
- **Personalization**
 - Dialog Usage
 - Word Usage
- **High Level Patterns**
 - Form filling
 - List navigation
 - One step correction

Application Development Environments



- Dialog Modules fully integrated in several existing IVR development environment
 - Native C/C++
 - Invision/Intervoice
 - Aspect
 - ArtiSoft (Visual Basic + Activex DMs)

VoiceXML dialog systems



VoiceXML dialog systems

TELEPHONE
NETWORK

Telephony

```
<prompt>
  Beverages cost
  <sayas class=money> 1.50 </sayas>
  per serving.
  <audio src="/Chaching.wav"/>
</prompt>

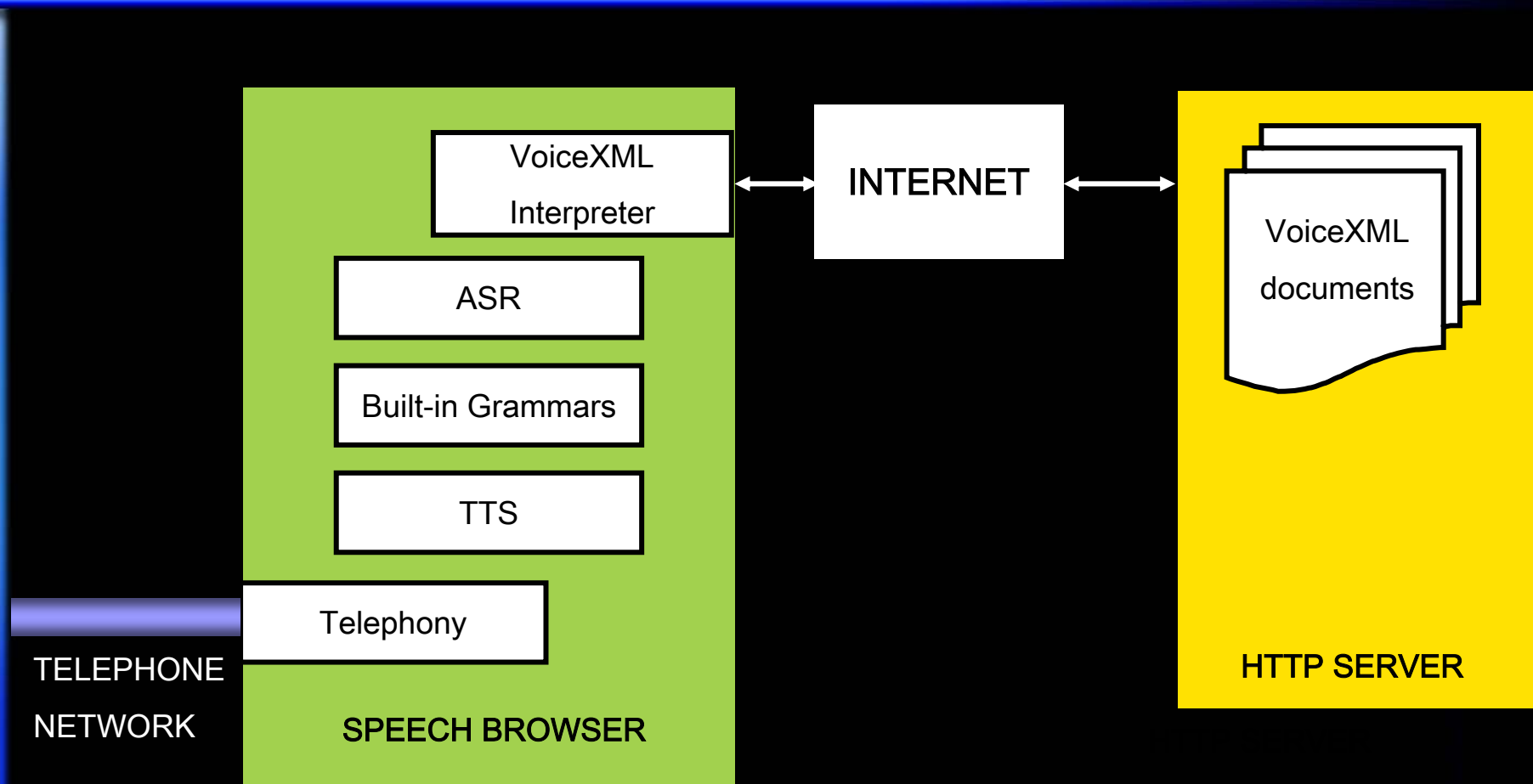
<form>
  <field name="drink">
    <prompt>
      Would you like coffee or juice?
    </prompt>
    <grammar src="drink.gram"/>
  </field>
</form>

<if cond="drink == 'coffee'">
  <goto "Coffee.asp" />
<else> <menu>
  <prompt> Would you like </prompt>
  <choice dtmf="1" next="OJ.asp"/>
  Orange Juice </choice> or
  <choice dtmf="2" next="Ap.asp"/>
  Apple Juice </choice> </menu>
</if>
```

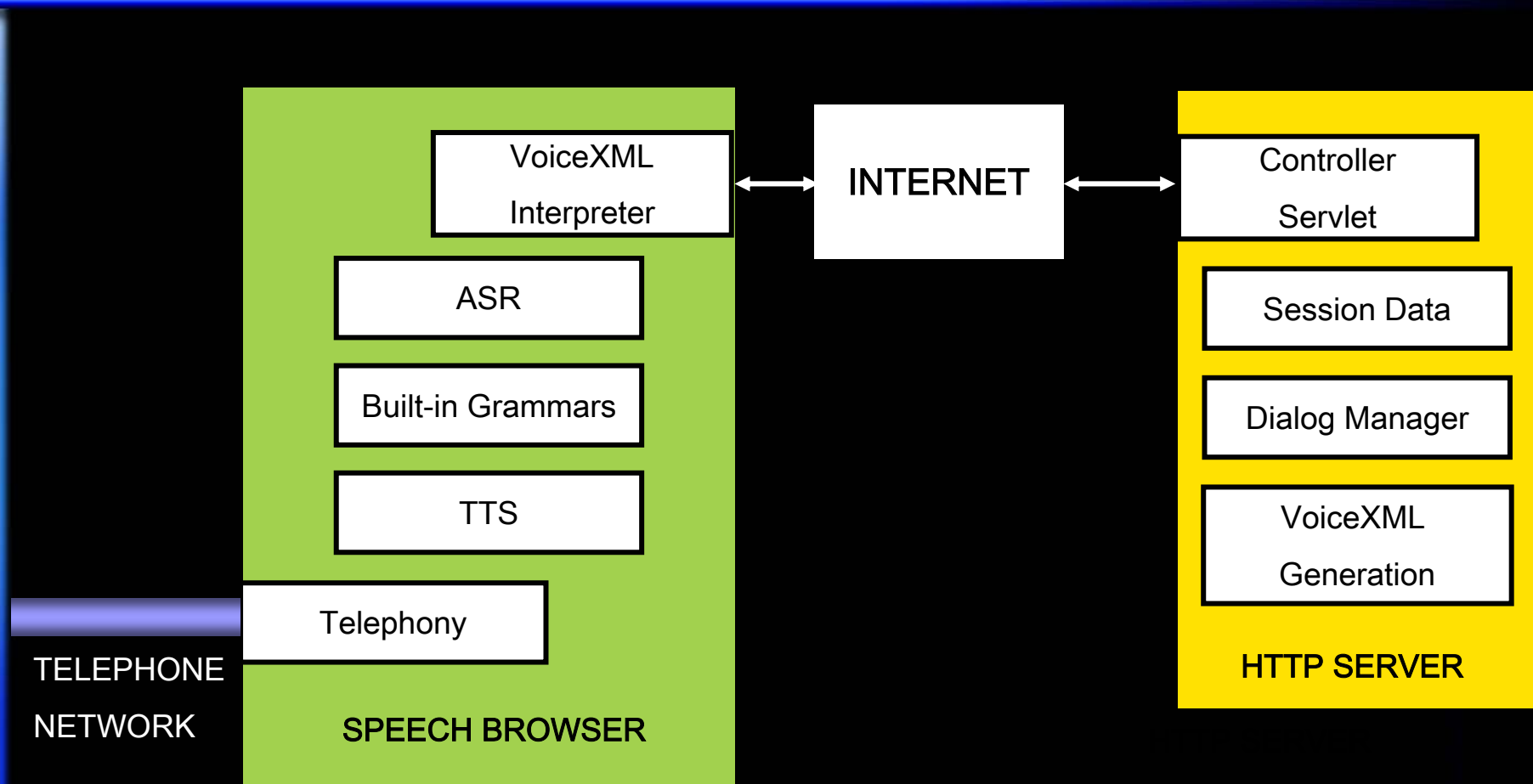
VoiceXML
documents

HTTP SERVER

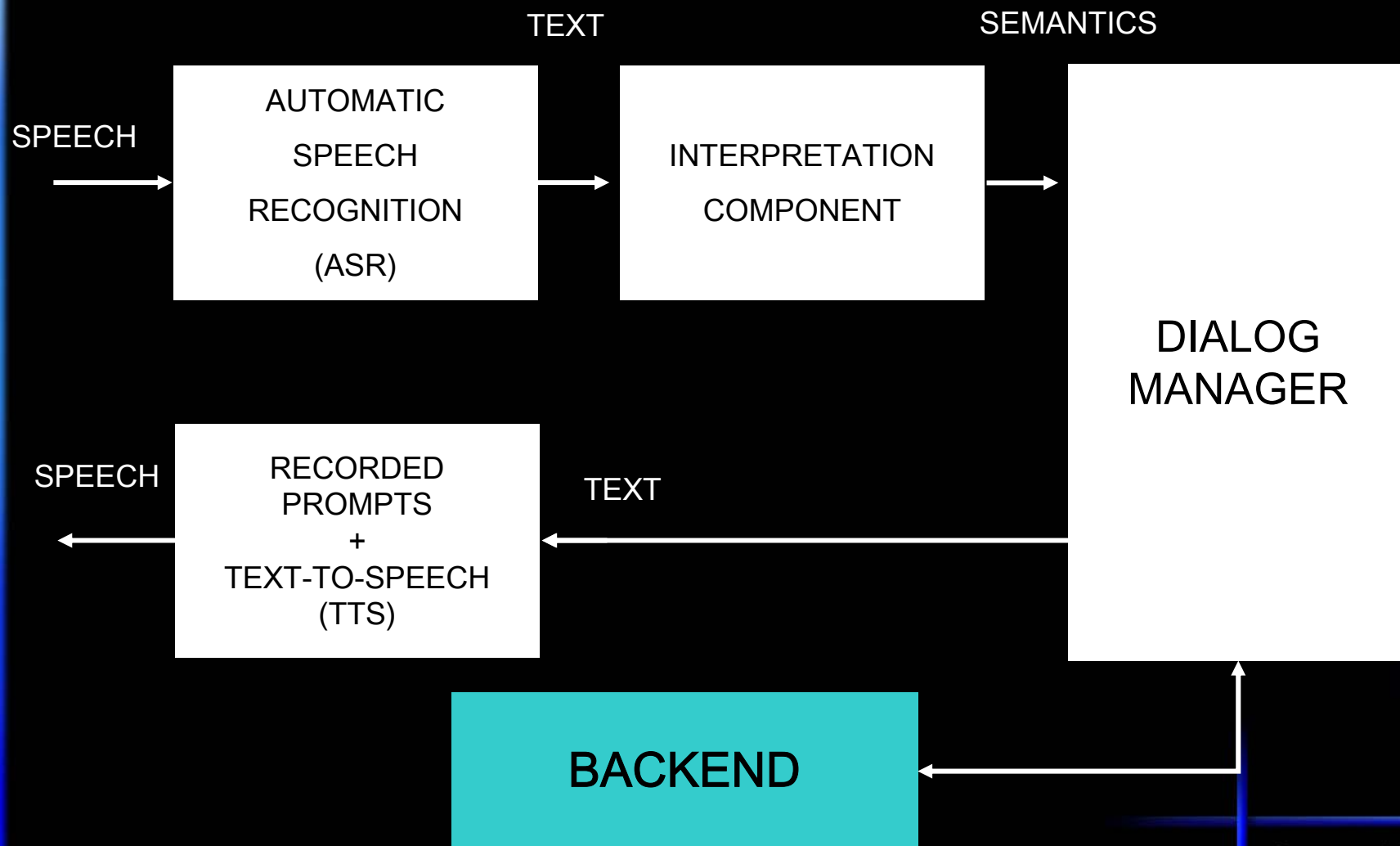
Dynamic Generation of VoiceXML



Dynamic Generation of VoiceXML



Dialog Management



Yesterday

- IVR / Touch Tone

Today

- Spoken Dialog Systems
 - The technology
 - **The Art**

Building commercial dialog systems

- **SADL: Speech Application Development Lifecycle**

- Requirement Analysis
- Specification
- Usability testing
- Integration
- Tuning
- Partial deployment
- Full deployment
- Continuous monitoring

AUTOMATIC EVALUATION

User Interface Design

- **Speech interfaces are new**
- **Speech interfaces are social**
- **Speech interfaces are somewhat human and somewhat synthetic**

Interaction: The psychology

- People treat computers like real people

- How do we know?

- ~Experiments; Cliff Nass, Byron Reeves

- Example 1: Politeness

- “People are polite to computers: When they are asked to evaluate a computer's performance, they tend to assess the one they are using more positively than others -- just as people tend to praise other people more to their faces than behind their backs.”*

- Nass+Reeves – The Media Equation

Prompt Scaffolding

- **Scaffolding: Structure which helps the caller learn.**

I know that you know that I know...

“What’s the approximate length of your package, in inches?
Please round off to the nearest inch.”

“What’s the approximate width, in inches?”

“What’s the height?”

Voice Casting

- Same words, spoken by different people...

Professional

Amateur

Professorial

Industrial

Accent (New York?)

Motherly

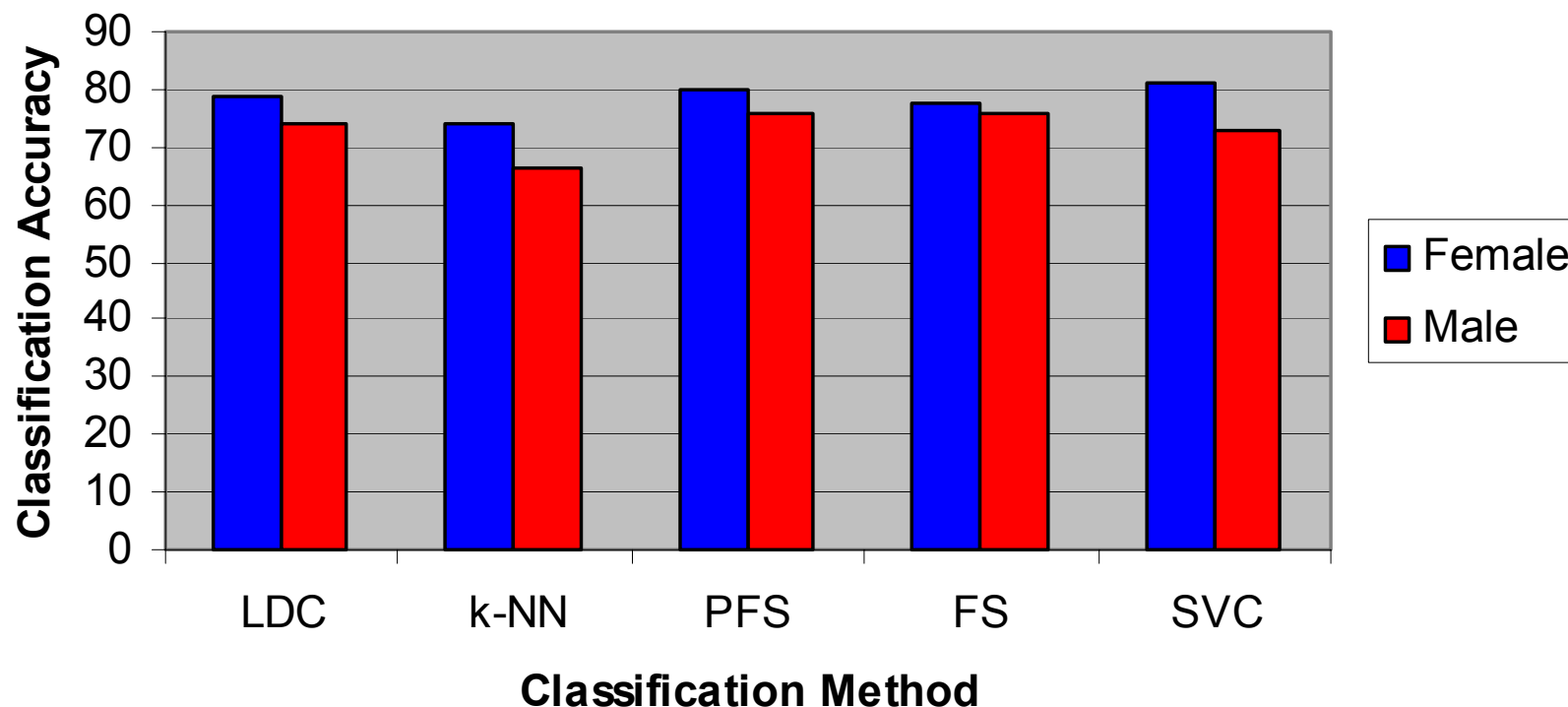
Mature

Disc Jockey

Emotion Classification



Negative vs Non Negative Emotion Classification



From: Lee, C.M., Narayanan, S., Pieraccini R., "Recognition of Negative Emotions from the Speech Signal," Proc. of ASRU01 IEEE Workshop, Madonna di Campiglio, Italy, Dec. 2001.

Yesterday

- IVR / Touch Tone

Today

- Spoken Dialog Systems
 - The technology
 - The Art

Tomorrow

- Multimodal / Wireless

What's Coming

- **2.5G and 3G wireless networks**
 - Always on data networks
 - Merged voice/data communications
 - Reasonable bandwidths / costs
- **Continued evolution of devices**
 - More processing
 - More memory
 - Better displays
- **Still tedious data entry**
 - Graffiti
 - Soft Keyboard

Issues with Multi-Modal applications

- **User Interface Issues**

- Visual Centric vs. Dialog Centric view
- Speech only and multi-modal UI patterns
- Situationalization

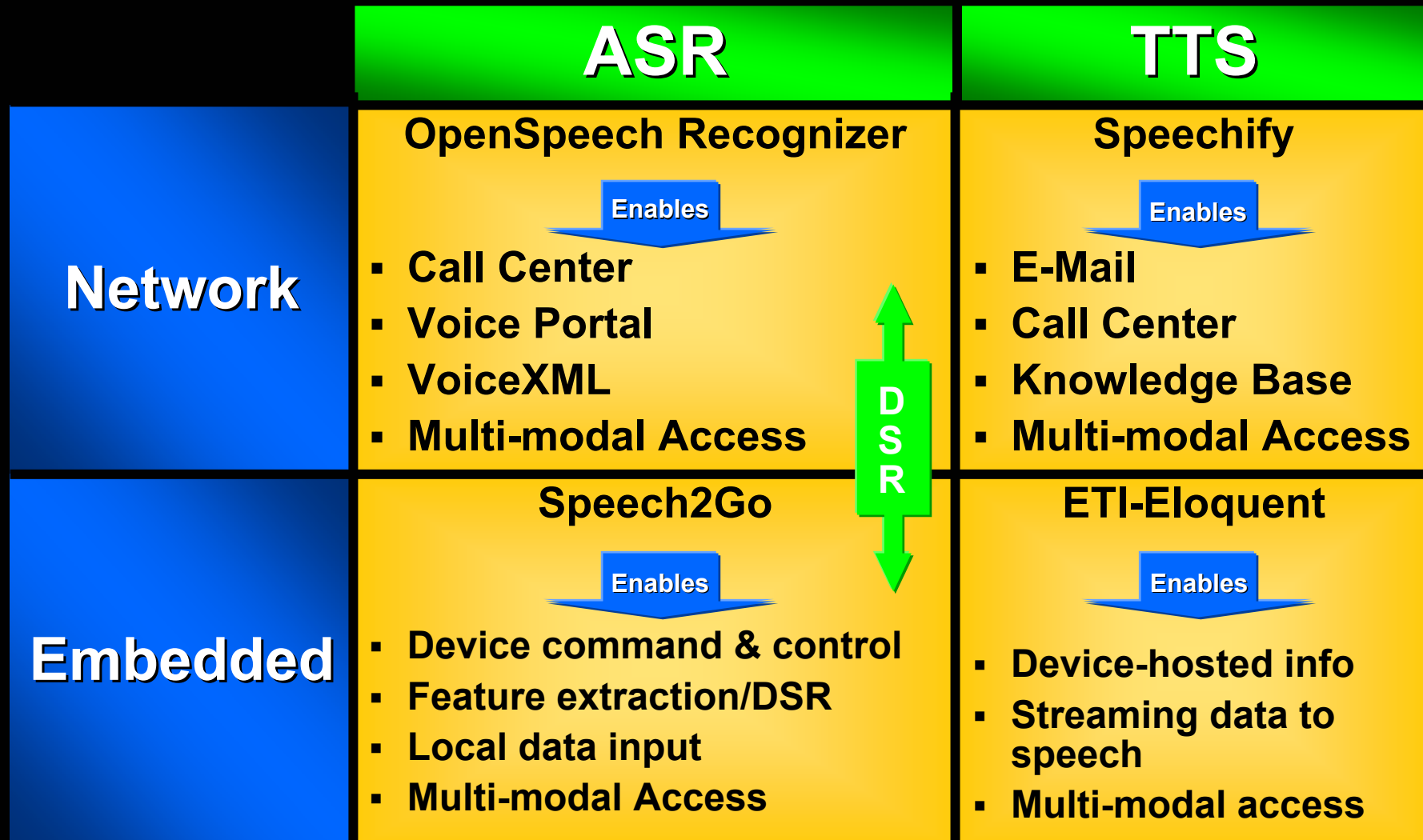
- **Speech Technology**

- Network vs. embedded resources
- Distributed Speech Recognition

- **Architectural Issues**

- Tight and loose control of UI
- Standards

Speech Technology Quadrant



Yesterday

- **IVR / Touch Tone**

Today

- **Spoken Dialog Systems**
 - **The technology**
 - **The Art**

Tomorrow

- **Multimodal / Wireless**

What we learned

What we learned

- **Evolution of Human machine communication**
 - IVR -> Spoken Dialog -> Multi-modal
- **Technology**
 - Speech Recognition – continuing improvement
 - Lack of data makes building new applications more difficult
 - Network and Embedded Resources
 - Architectures, Platforms and Standards
- **Art**
 - User Interface design makes technology usable