Speech technologies for interactive mobile applications – a primer

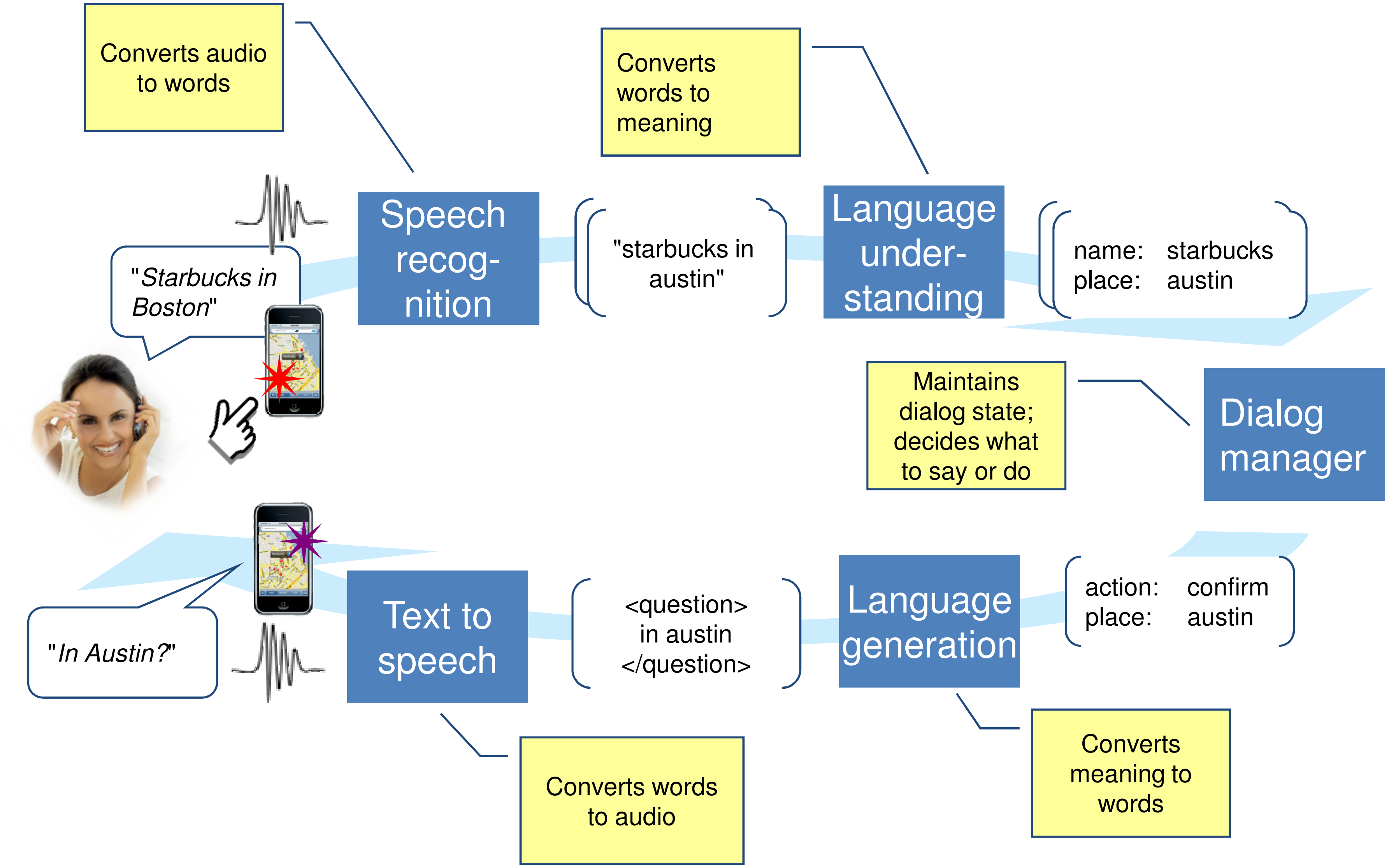
Jason D. Williams

AVIxD / IxD workshop – August 2011

Roadmap

- A quick tour of the toolbox
- Technology and usage issues (that influence design)
- A few pointers on getting started

What is a spoken dialog system?



Converts audio to words

Speech recognition

"starbucks in austin"

Converts words to meaning

Language understanding

name: starbucks
place: austin

Maintains dialog state; decides what to say or do

Dialog manager

action: confirm
place: austin

Language generation

Converts meaning to words

<question> in austin </question>

Text to speech

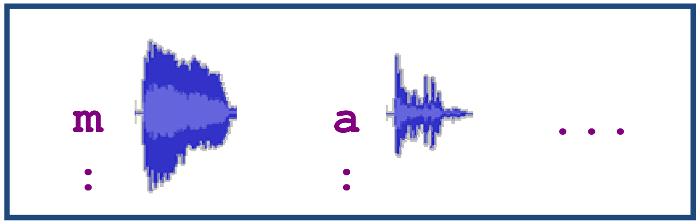
Converts words to audio

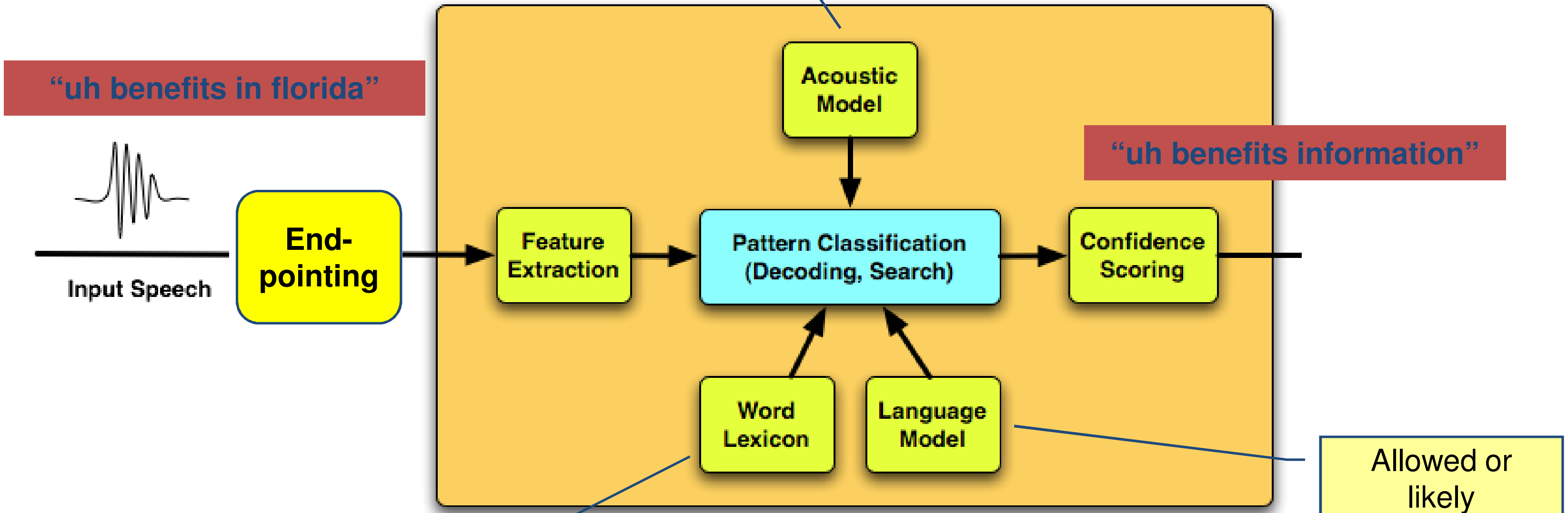
"*Starbucks in Boston*"

"*In Austin?*"

Speech recognition

Audio to phonemes





Allowed or likely sequences of words

Phonemes to words

```
tomato:  t o m a ^ o
         t o m A ^ o
```

```
[ (?(i live in) [
   alaska~0.001
   california~0.2
   arkansas~0.01
   ...]))]
```

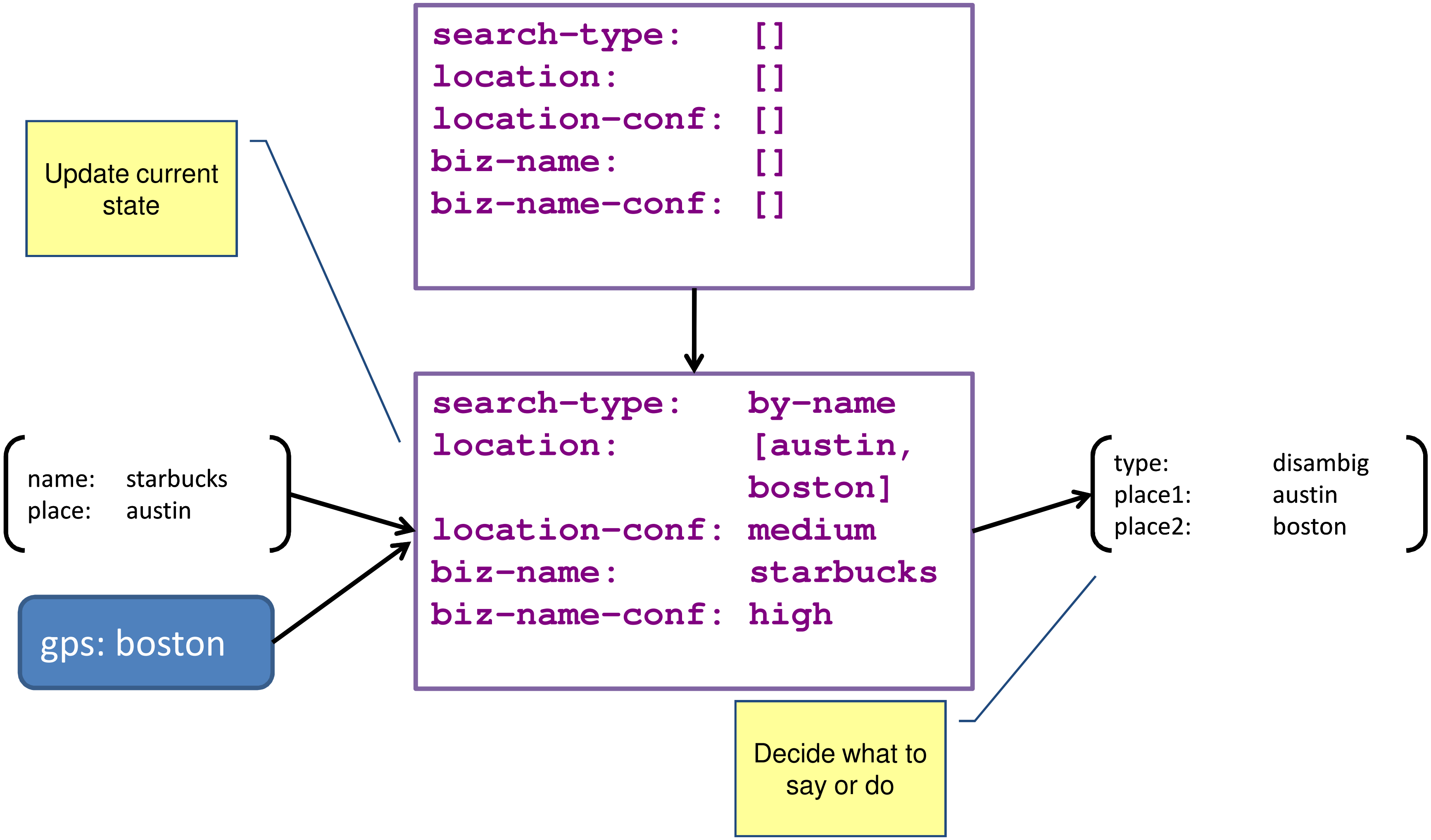
Language understanding

- Example input/output:

TRANSCRIPTION (input)	INTERPRETATION (output)
f_m_l_a claim	<call_request "fmla"> <action "claim">
payroll	<ambig_key "payroll">
customer service	[none]
i'd [fragment] ~	[none]
f_m_l_a agent	<call_request "fmla"> <action "operator">
employment	<call_request "employment">
[side_speech]	[none]
more examples	<command "moreoptions">
medical benefits	<ambig_key "healthplan">

- Can be done with rules, or by a pattern classifier

Dialog manager



Update current state

```
search-type:    []
location:       []
location-conf:  []
biz-name:       []
biz-name-conf:  []
```

```
search-type:    by-name
location:       [austin,
                boston]
location-conf:  medium
biz-name:       starbucks
biz-name-conf:  high
```

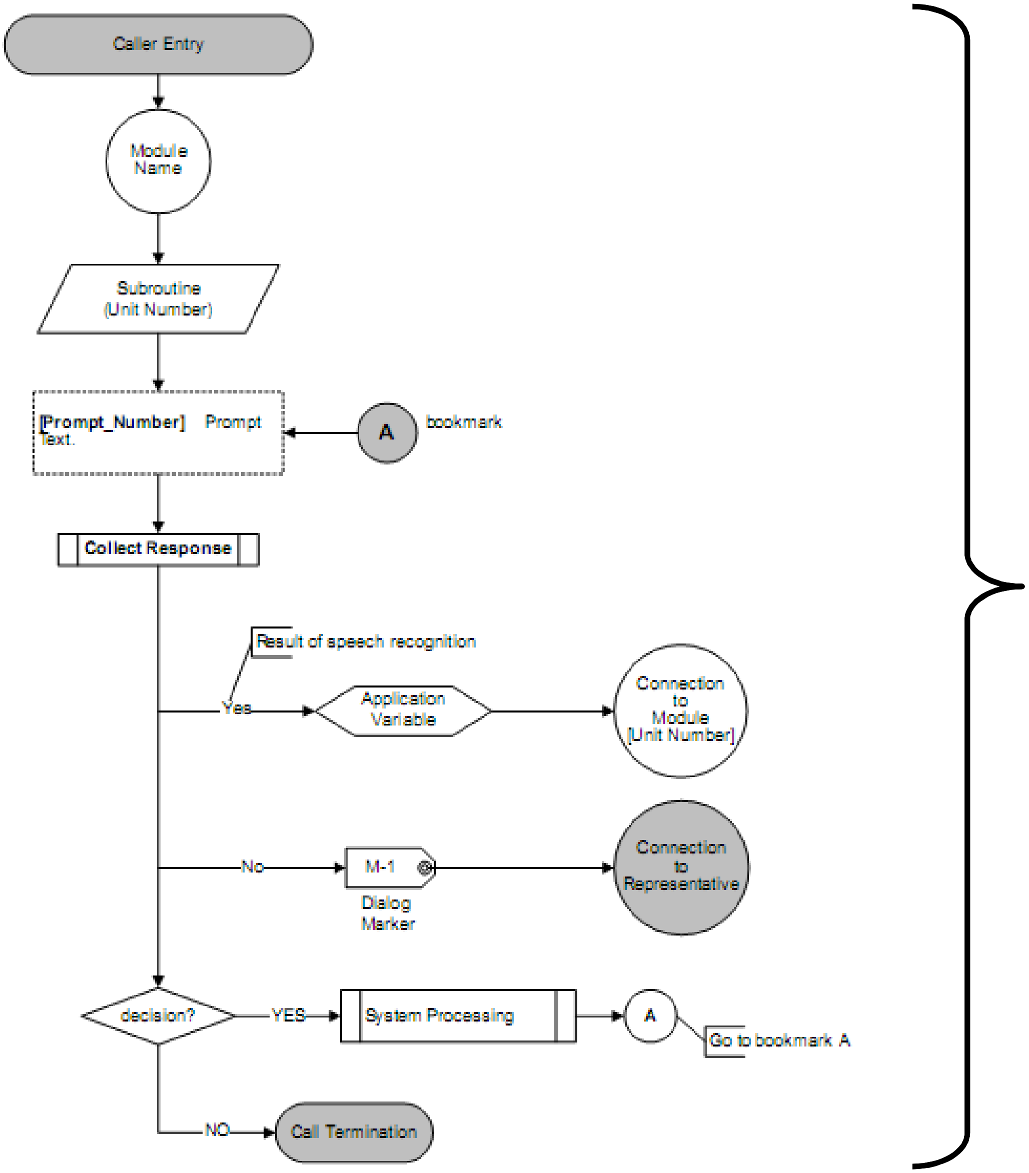
name: starbucks
place: austin

gps: boston

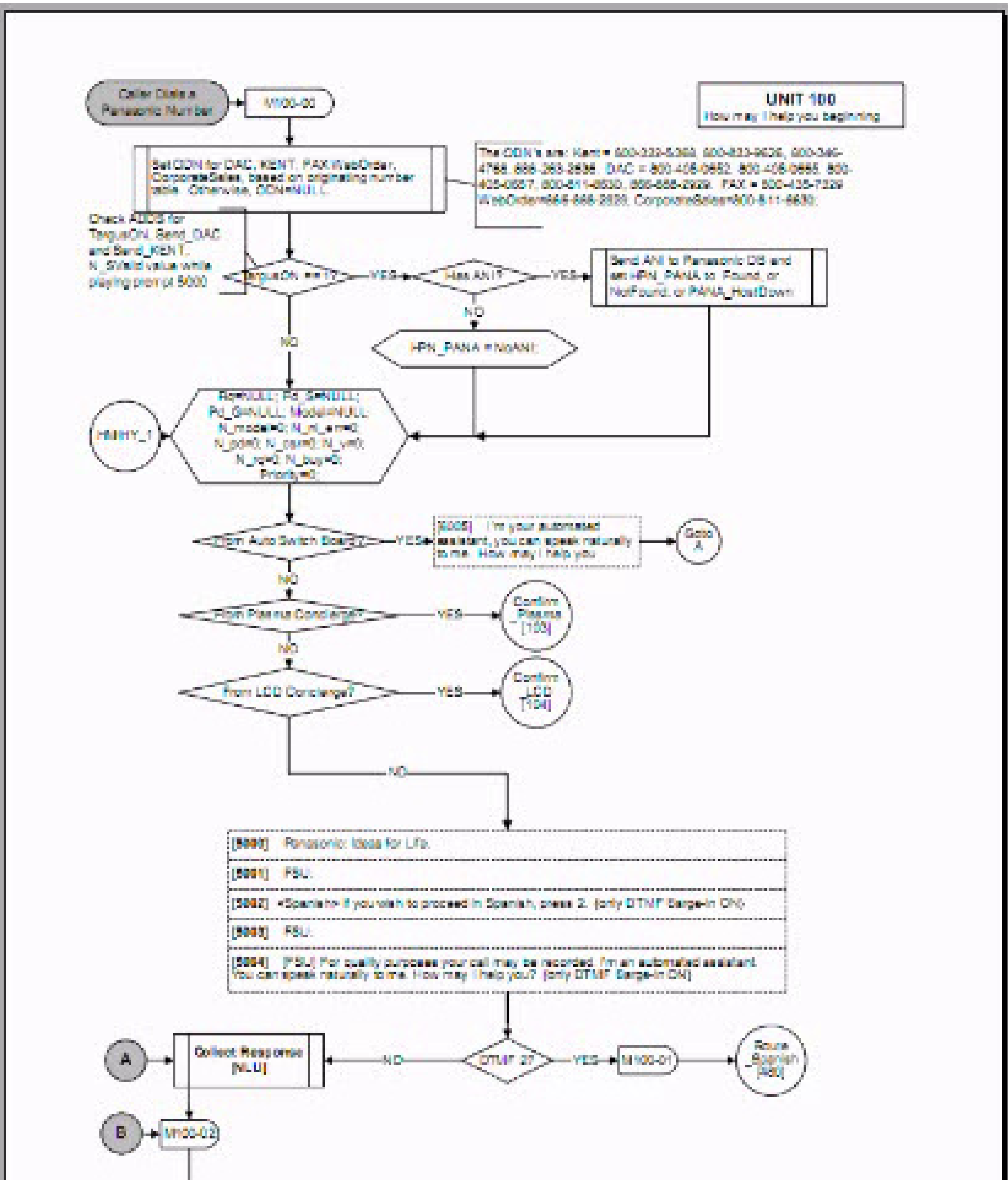
type: disambig
place1: austin
place2: boston

Decide what to say or do

How dialog systems are designed today



Typical commercial spoken dialog system contains ~100 pages of flowchart


UNIT 100
How may I help you beginning
Caller Dials a Panasonic Number
M100-00
Set DDN for DAC, KENT, FAX WebOrder, CorporateSales, based on originating number table. Otherwise, DDN=NULL.
Check ADDS for TargusON, Send_DAC and Send_KENT, N_SValid value while playing prompt 5000
TargusON == 1?
YES
Has ANI?
YES
Send ANI to Panasonic DB and set HPN_PANA to Found, or NotFound, or PANA_HostDown
NO
HPN_PANA = NoANI
NO
HMIHY_1
Po=NULL; Po_S=NULL; Po_G=NULL; Model=NULL; N_model=0; N_nl_err=0; N_pd=0; N_csr=0; N_vel=0; N_po=0; N_buy=0; Priority=0;
From Auto Switch Board?
YES
Goto A
NO
From Plasma Concierge?
YES
Confirm _Plasma [103]
NO
From LCD Concierge?
YES
Confirm _LCD [104]
NO
DTMF 2?
YES
M100-01
Route _Spanish [400]
NO
A
Collect Response [NLU]
B
M100-02

The DDN's are: Kent = 800-222-5369, 800-833-9626, 800-346-4768, 888-263-2698. DAC = 800-405-0652, 800-405-0665, 800-405-0657, 800-811-8630, 866-888-2929. FAX = 800-435-7329 WebOrder=866-888-2626, CorporateSales=800-811-8630.

[5005] I'm your automated assistant, you can speak naturally to me. How may I help you

[5000] Panasonic: Ideas for Life.

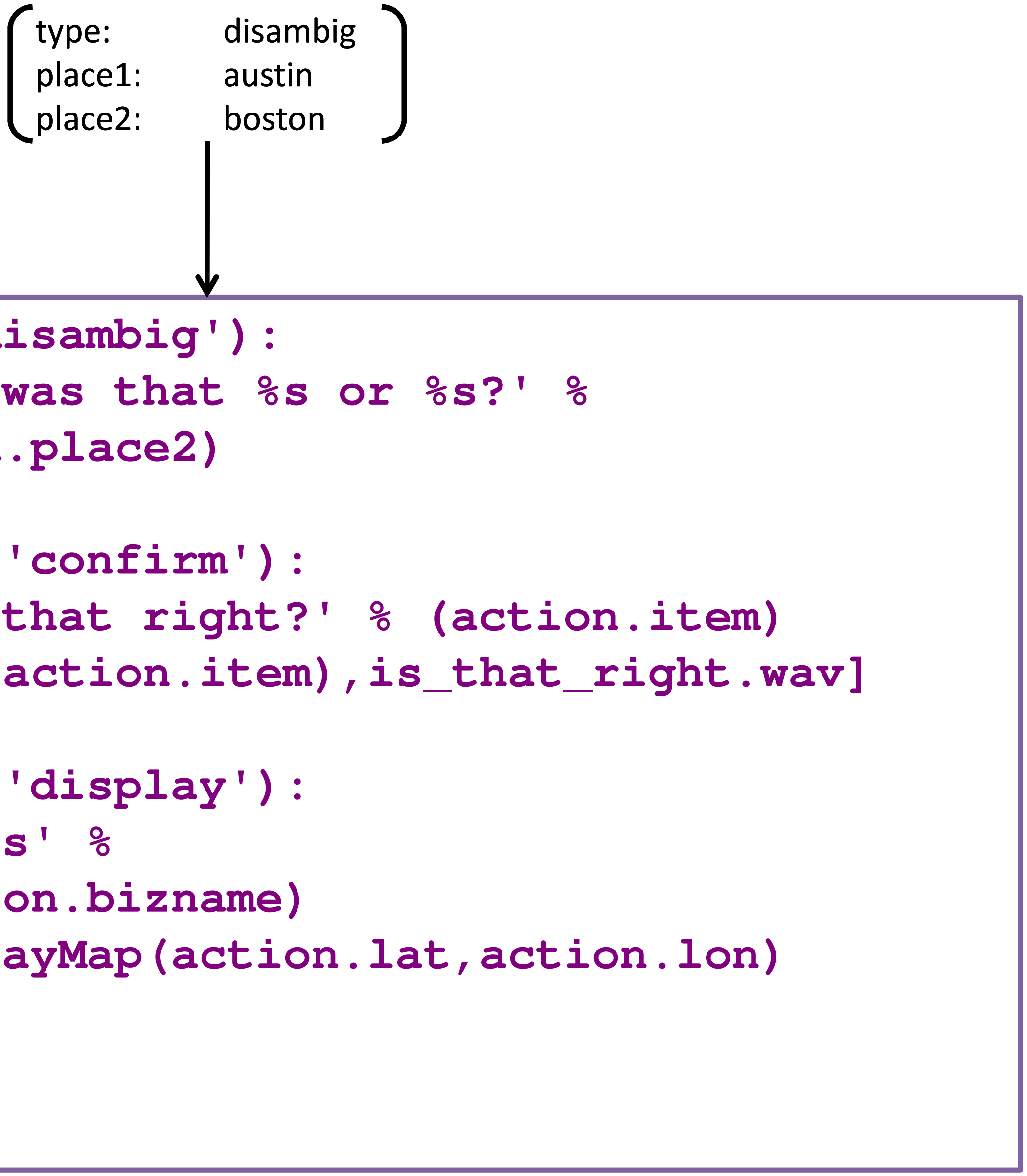
[5001] FSU.

[5002] <Spanish> If you wish to proceed in Spanish, press 2. (only DTMF Barge-In ON)

[5003] FSU.

[5004] [FSU] For quality purposes your call may be recorded. I'm an automated assistant. You can speak naturally to me. How may I help you? (only DTMF Barge-In ON)

Language generation



type:	disambig
place1:	austin
place2:	boston

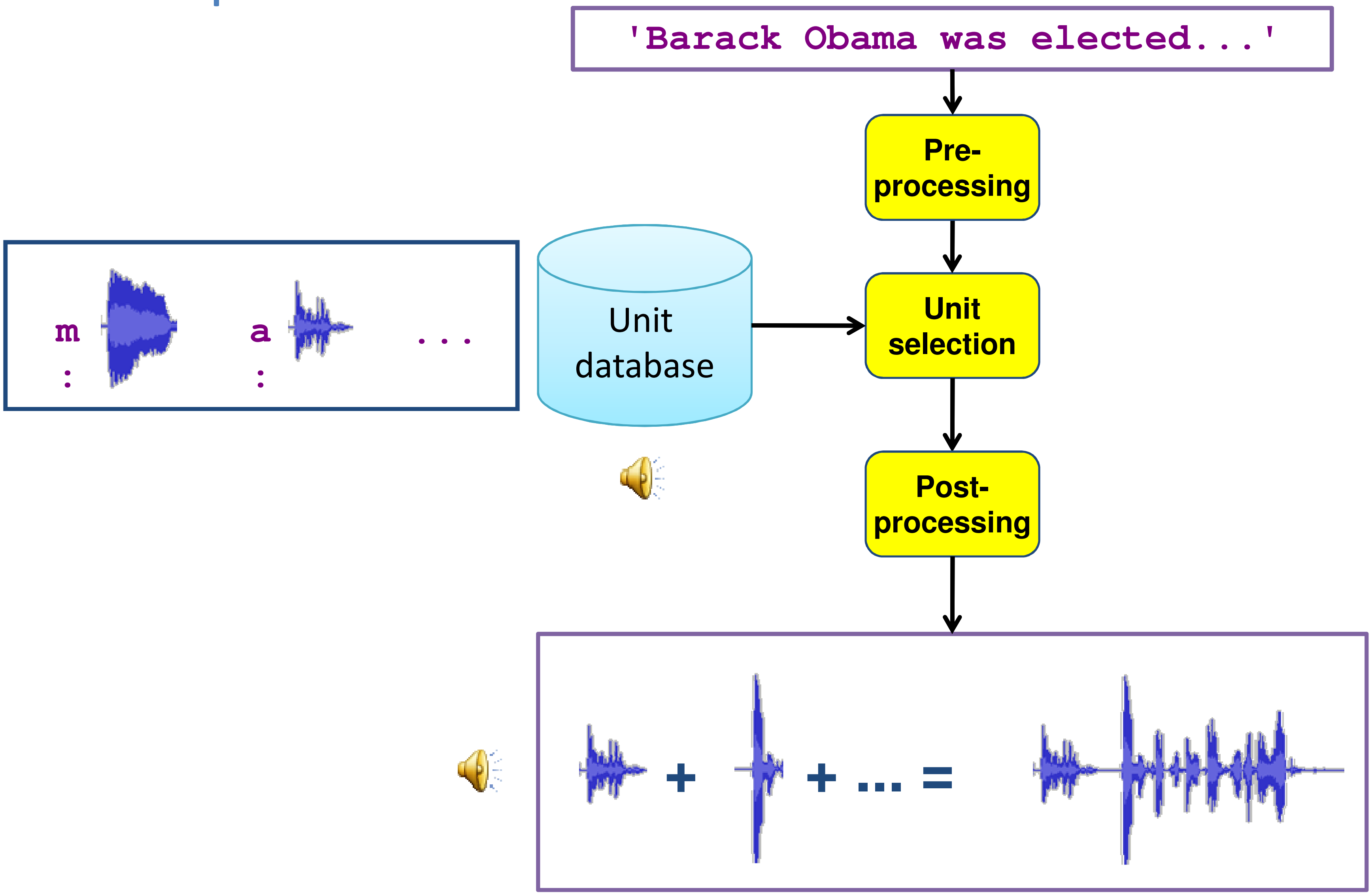
```
if (action.type == 'disambig'):
  out.text = 'Sorry, was that %s or %s?' %
(action.place1,action.place2)

elif (action.type == 'confirm'):
  out.text = '%s, is that right?' % (action.item)
  out.wavs = [GetWav(action.item),is_that_right.wav]

elif (action.type == 'display'):
  out.text = '%s in %s' %
(action.business,action.bizname)
  out.display = DisplayMap(action.lat,action.lon)

...
```

Text-to-speech



'Barack Obama was elected...'

Pre-processing

Unit selection

Post-processing

Unit database

m

:

a

:

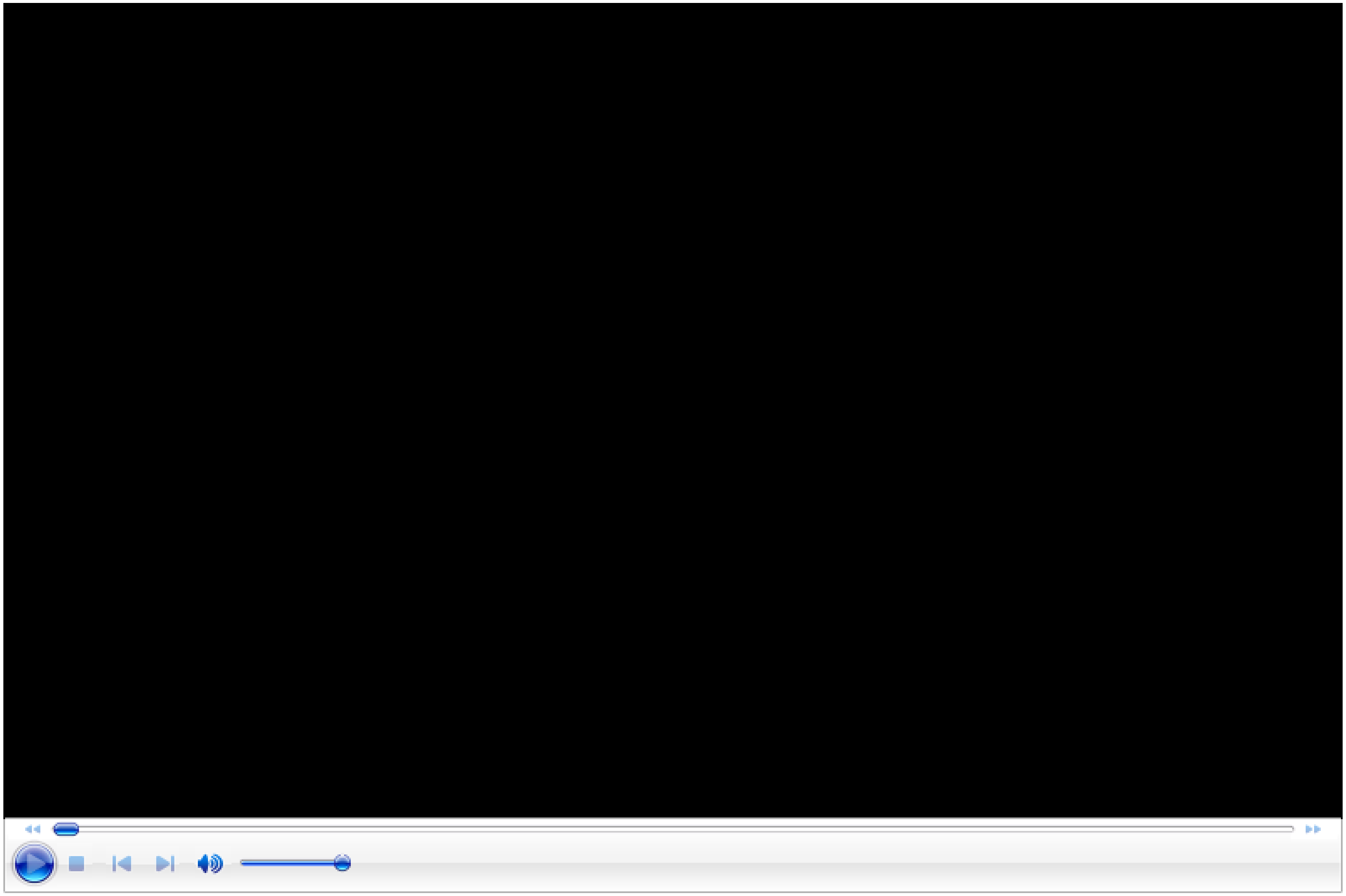
. . .

+ + ... =

Putting it all together: example

- Uni-modal telephone-based spoken dialog system
  - SLM speech recognition
  - Language understanding (~250 categories)
  - Speaker verification (~300K users)

An example multi-modal system



Two central challenges for deploying speech technology

AVIxD / IxD workshop – August 2011

ASR/SLU errors are common

Grammar	Yes/no	City & state	How may I help you?

Source: Two different deployed commercial applications running two different speech recognizers

ASR/SLU errors are common

Grammar	Yes/no	City & state	How may I help you?
In-grammar/ in-domain accuracy	99.8%	85.1%	89.5%

Source: Two different deployed commercial applications running two different speech recognizers

ASR/SLU errors are common

Grammar	Yes/no	City & state	How may I help you?
In-grammar/ in-domain accuracy	99.8%	85.1%	89.5%
% in-grammar/ in-domain	92.3%	91.0%	86.8%
Overall accuracy	92.1%	77.6%	77.7%

Source: Two different deployed commercial applications running two different speech recognizers

ASR/SLU errors are common

Grammar	Yes/no	City & state	How may I help you?
In-grammar/ in-domain accuracy	99.8%	85.1%	89.5%
% in-grammar/ in-domain	92.3%	91.0%	86.8%
Overall accuracy	92.1%	77.6%	77.7%
Accepted utts (False accepts)	89.6% (1.8%)	60.3% (4.9%)	73.3% (8.3%)

Source: Two different deployed commercial applications running two different speech recognizers

The "theory of mind" problem

	A real human	Graphical user interface	Spoken dialog system
			☎
What can it understand?	Anything you can explain	Only the buttons you can press	The contents of the grammar
How do I know a task *can't* be performed?	"I can't do that."	There is no button for it.	It's not in the grammar

Users must think simultaneously about what language the system can understand, and what the system can do – they must form a "theory of mind" about the dialog system

Responses to "How may I help you?"

- Silences and hesitations while users think
  - Leads to end-pointing problems
  - Leads to users confusing themselves
- "Robot" language (hence examples, "speak naturally")
  - Example 1
  - Example 2
- Recognition errors confused with competences
  - > "i need to sign up for a get off benefit" *[no parse]*
  - > "i would like to enroll in a get one" *[no parse]*
  - > "i would like to get help with my dental insurance" <HELP>
  - > "dental insurance" <INSURANCE>

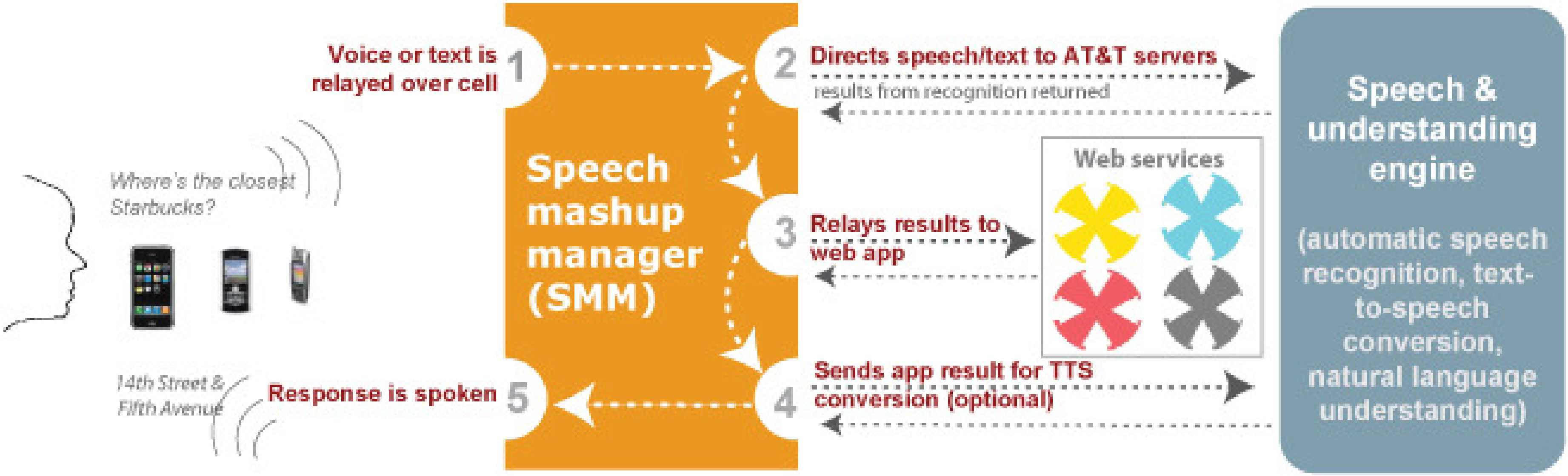
Source: Live calls, human resources dialog system, AT&T

Getting started with speech technology

AVIxD / IxD workshop – August 2011

How do I add ASR to my mobile application?

- ASR on mobile devices is usually a bad idea
  - ASR will kill your battery
  - Mobile processors are probably too small
  - Can't benefit from cross-user acoustic data
- Probably better to run ASR in the cloud
  - Stream audio to server; server sends back ASR result
  - Example: AT&T Speech Mash-up


Voice or text is relayed over cell
1
Where's the closest Starbucks?
Speech mashup manager (SMM)
2
Directs speech/text to AT&T servers
results from recognition returned
Speech & understanding engine
(automatic speech recognition, text-to-speech conversion, natural language understanding)
Web services
3
Relays results to web app
4
Sends app result for TTS conversion (optional)
5
Response is spoken
14th Street & Fifth Avenue

Language models: constrain/weigh word sequences

Usually, each "dialog state" has its own language model (LM)

Rule-based LMs

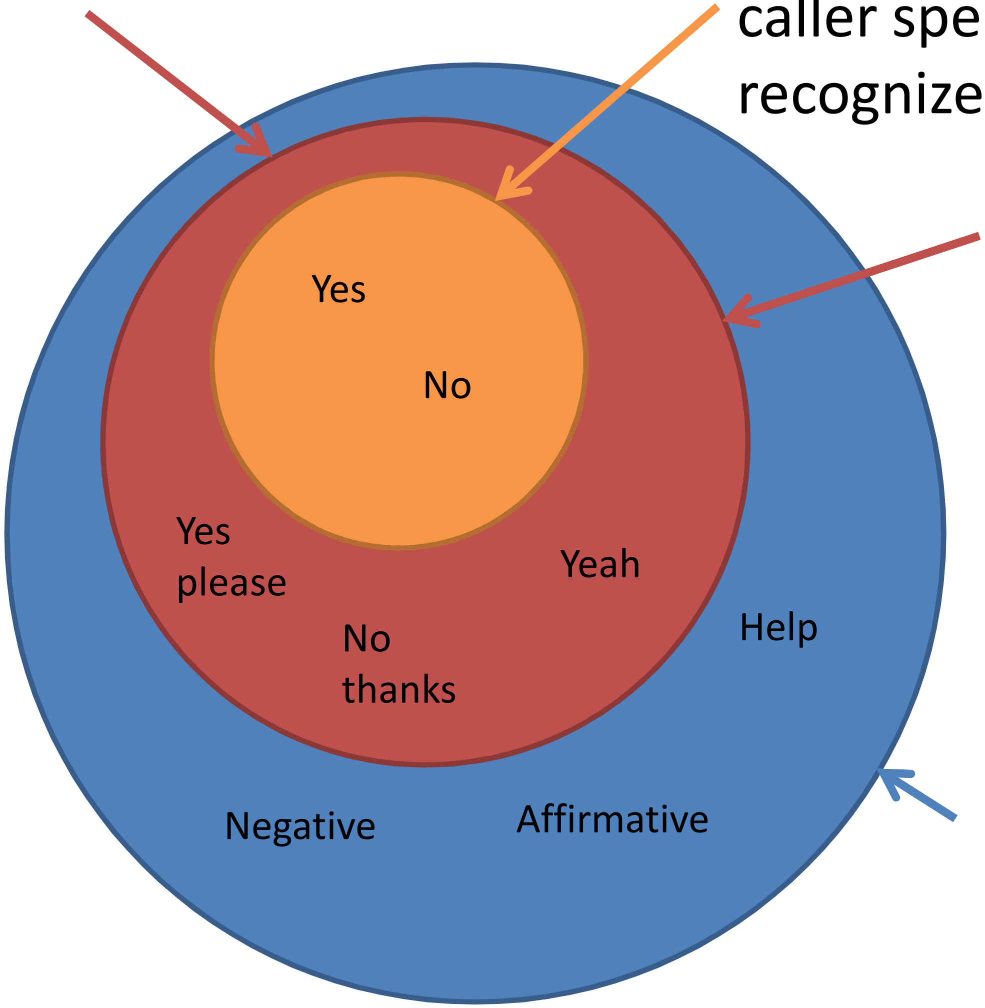
```
GET_STATE [
  (?(i live in) [
   alaska~0.001
   california~0.2
   arkansas~0.01
   ...
  ])]
```

Statisical LMs

```
i want 0.0035
want to 0.0023
want benefits 0.0034
want need 0.000002
```

What to include in the language model?

What the user says

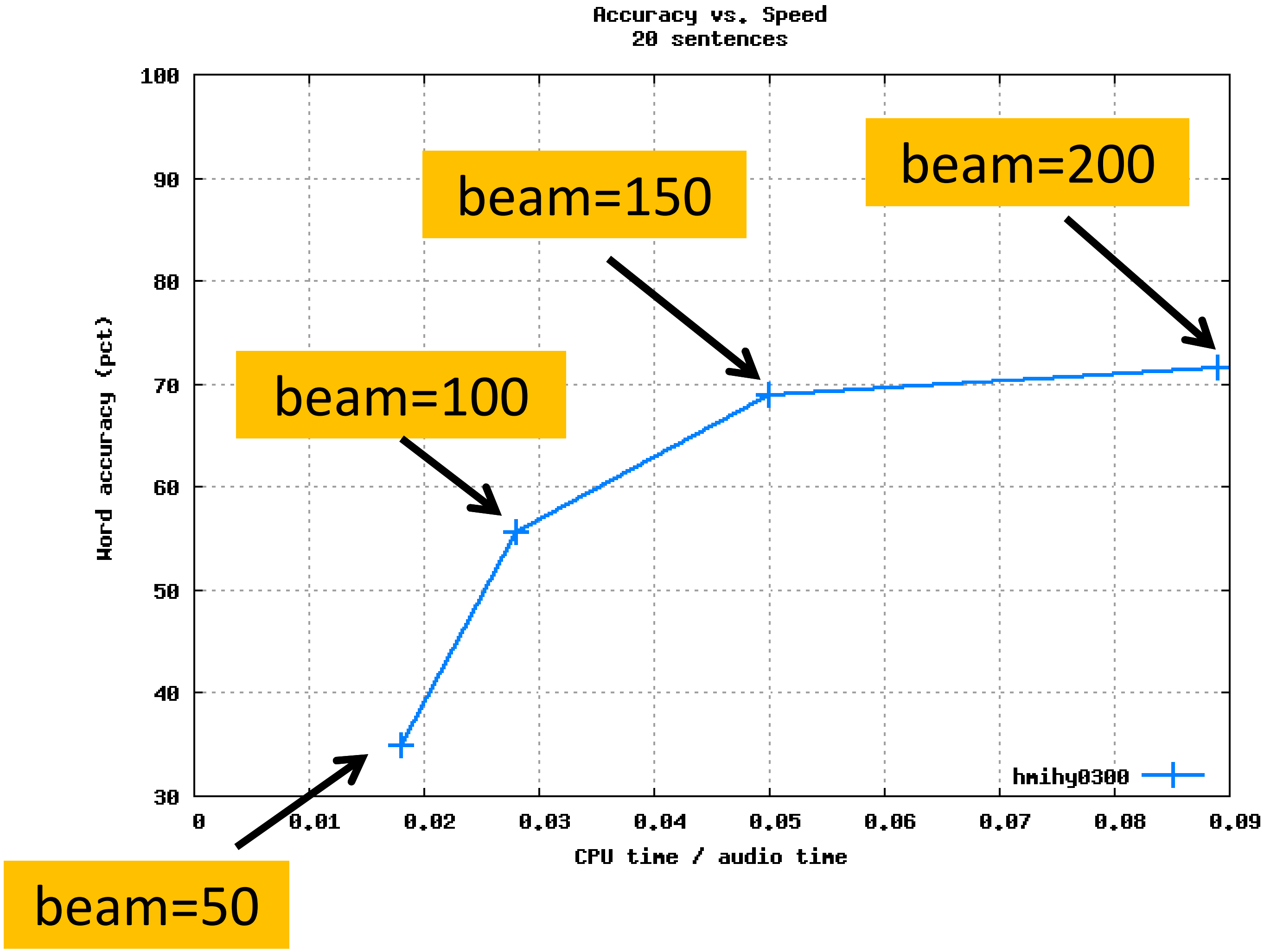


Too narrow: some caller speech can't be recognized

Just right: only things caller actually says can be recognized

Too broad: things caller will never say could be recognized, increasing misrecognitions

Setting parameters: beam width



beam=50

beam=100

beam=150

beam=200

Another parameter: language model scale factor

LOW

HIGH

Too low: good acoustic matches but words make no sense; low accuracy

Just right: a little acoustic information is ignored in favor of words that make sense; best accuracy

Too high: word sequences make sense but too much acoustic information is ignored; low accuracy

Example "mac store santa monica california"

LM Scale	Result

Example "mac store santa monica california"

LM Scale	Result
0	mm aq ck storr sade a mon in ck a california

Example "mac store santa monica california"

LM Scale	Result
0	mm aq ck storr sade a mon in ck a california
0.01	maxtor sadd a monnin ke california

Example "mac store santa monica california"

LM Scale	Result
0	mm aq ck storr sade a mon in ck a california
0.01	maxtor sadd a monnin ke california
1	maxtor sad a monica california
2	maxtor sata monica california

Example "mac store santa monica california"

LM Scale	Result
0	mm aq ck storr sade a mon in ck a california
0.01	maxtor sadd a monnin ke california
1	maxtor sad a monica california
2	maxtor sata monica california
3-19	mac store santa monica california

Example "mac store santa monica california"

LM Scale	Result
0	mm aq ck storr sade a mon in ck a california
0.01	maxtor sadd a monnin ke california
1	maxtor sad a monica california
2	maxtor sata monica california
3-19	mac store santa monica california
20	maxtor pharmacies

Example "mac store santa monica california"

LM Scale	Result
0	mm aq ck storr sade a mon in ck a california
0.01	maxtor sadd a monnin ke california
1	maxtor sad a monica california
2	maxtor sata monica california
3-19	mac store santa monica california
20	maxtor pharmacies
25+	restaurants

Summary

Brief tour of the toolbox

Challenges for building ASR systems

A few pointers for getting started

But what about design?!

That's next!

Speech technologies for interactive mobile applications – a primer

Thanks!

Jason D. Williams

at&t

AVIxD / IxD workshop – August 2011