

Spoken dialogue systems: challenges, and opportunities for research

Jason D. Williams




ASRU – December 2009

What is a spoken dialog system?

A spoken dialogue system is a computer agent that interacts with people by understanding spoken language.

Spoken dialogue systems come in many flavours

Input	Output	Example
Speech	Speech	Telephone technical support [1] 
Speech + ?GUI	Speech + ?GUI	In-car music control, navigation
Speech + GUI	Speech + GUI	Tutoring
Speech + GUI	Speech + GUI	Language learning
Speech + GUI	?Speech + GUI	TV program guide
Speech + GUI	?Speech + GUI	Mobile search interface
Speech + vision	Speech + robot/agent	Eldercare
Speech + vision	Speech + robot/agent	Automated receptionist

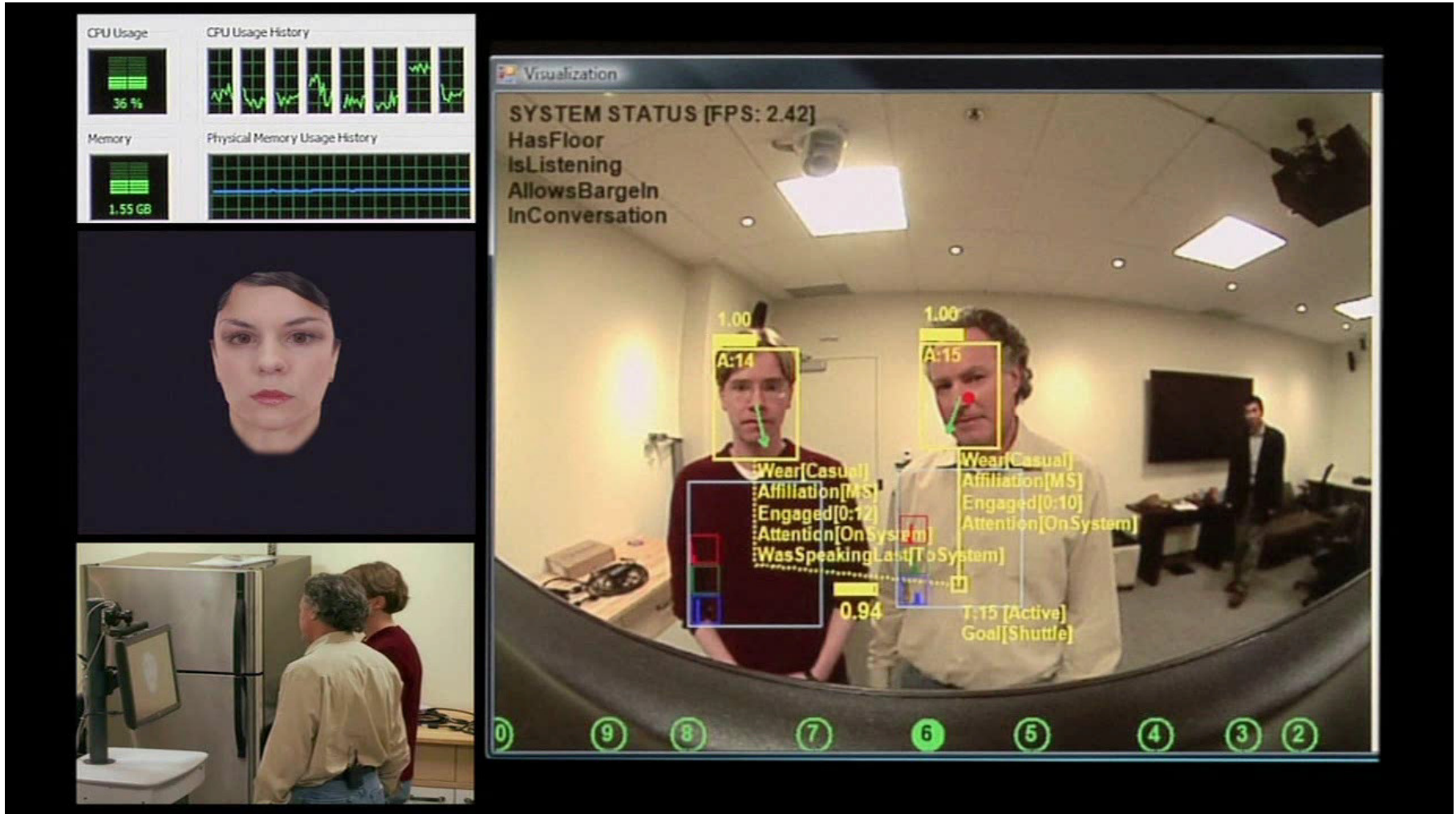
[1] Recording of a deployed dialog system, AT&T

In-car spoken dialogue system



Source: IBM

Automated receptionist



Bohus, D., Horvitz, E. (2009). Models for Multiparty Engagement in Open-World Dialog, in Proceedings of SIGdial'09, London, UK

Outline

- Key challenges for building dialogue systems
- Areas of current research
- Views on their potential for commercial success

Challenges

1. Channel errors (ASR, SLU, turn-taking)
2. Curse of history
3. Lack of a single optimization metric
4. Theory of mind problem

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%

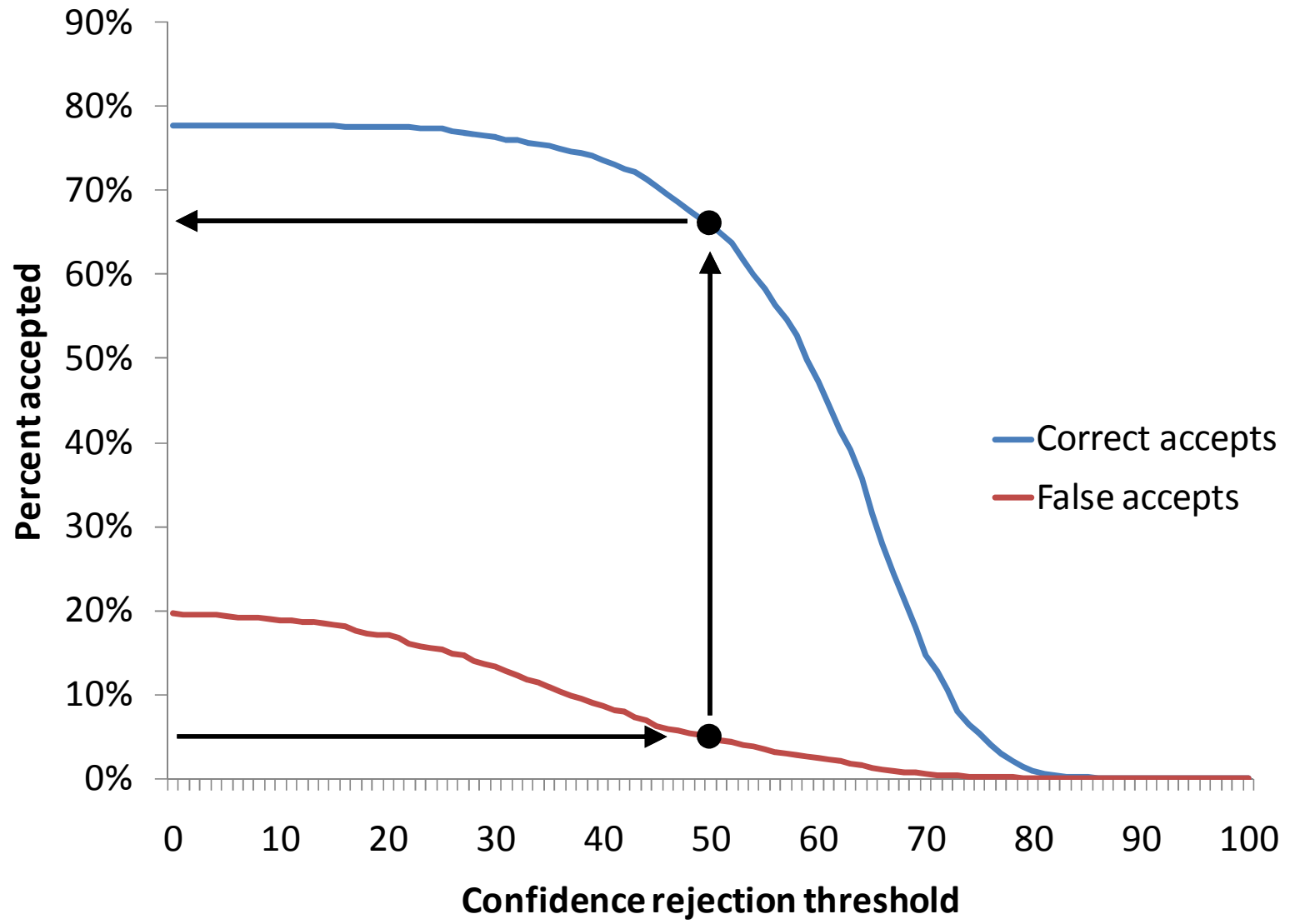
Source: Two different deployed commercial applications running two different speech recognizers

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 errors are hard to detect



ASR/SLU errors are common

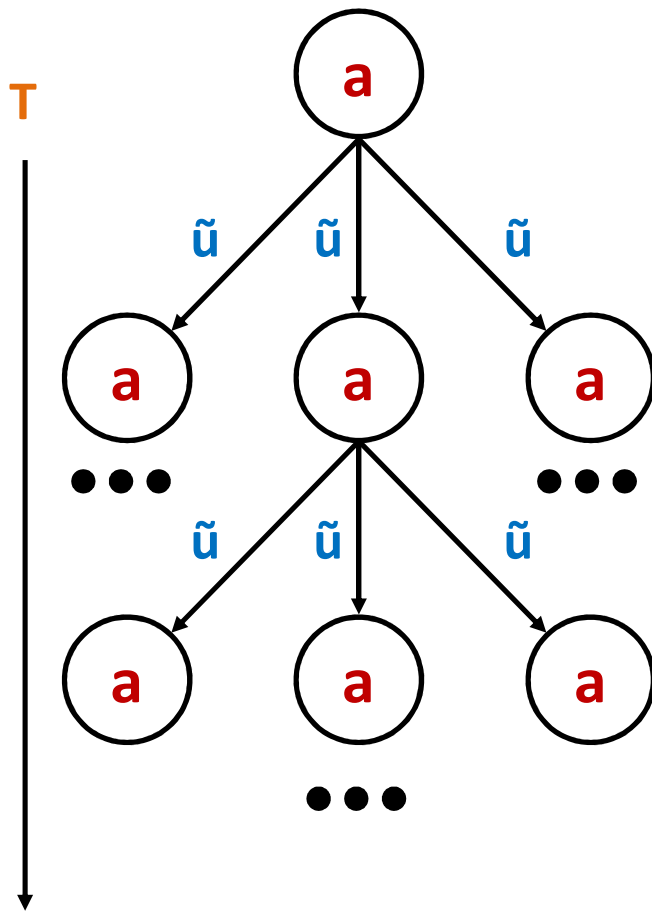
Grammar	Yes/no	City & state	How may I help you?
In-grammar/ in-domain accuracy	99.8%	85.1%	89.5%
% in-grammar/ in-domain	92.3%	91.0%	86.8%
Overall accuracy	92.1%	77.6%	77.7%
Accepted utts (False accepts)	89.6% (1.8%)	60.3% (4.9%)	73.3% (8.3%)

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

Curse of history (1/2)

$A = \{\text{ask}(\text{first-name}), \text{confirm}(\text{last-name}=\text{williams}), \dots\}$

$\tilde{U} = \{\text{YES}, \text{JASON}, \text{WILLIAMS}, \dots\}$



$\sim A^{\tilde{U}^T}$ possible
assignments

Typical system:

$$A = 10^{10}$$

$$\tilde{U} = 10^{10}$$

$$T = 10$$

Curse of history (1/2)

$$F(\tilde{u}_0, a_1, \tilde{u}_1, a_2, \tilde{u}_2, a_3, \tilde{u}_3, \dots, a_t, \tilde{u}_t) = a_{t+1}$$

Often it's more convenient to separate the *tracking* problem from the *action selection* problem:

Dialog state $s_t \approx (\tilde{u}_0, a_1, \tilde{u}_1, a_2, \tilde{u}_2, a_3, \tilde{u}_3, \dots, a_t, \tilde{u}_t)$

State tracking $s_{t+1} = G(s_t, a_t, \tilde{u}_n)$

Action selection $F(s_{t+1}) = a_{t+1}$

Now the problem is what to track in the dialog state s , and how to make use of it when choosing actions

Curse of history (2/2)

Many speech and language problems:

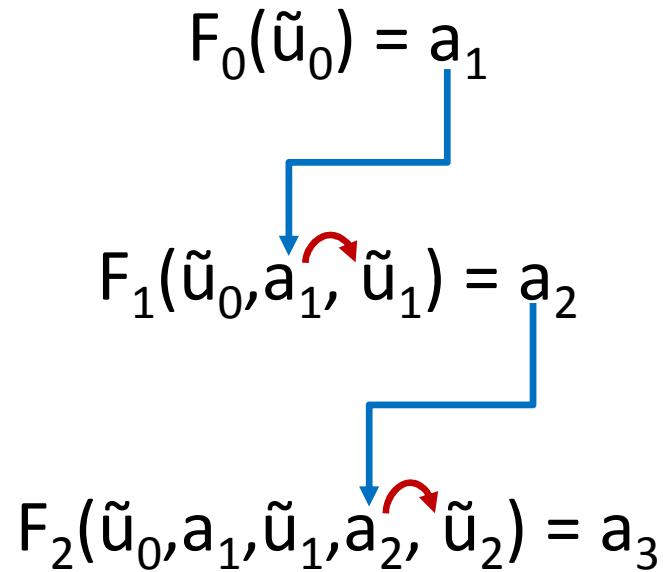
$$F(\text{input}) = \text{output}$$

Assumptions:

- $P(\text{input-data}, \text{output-data})$ is fixed
- Choice of F doesn't change $P(\text{input-data}, \text{output-data})$

These do not hold in spoken dialogue systems

Curse of history (2/2)



- Changing the dialogue system changes the distribution of the data
- Complete evaluations must be on real people – can't report end-to-end results on a common corpus

Lack of a single optimization metric

Candidate metric	Issues
Maximize user satisfaction	<ul style="list-style-type: none">• How to measure in <i>real</i> systems?• Agents – gethuman.com
Maximize task completion	<ul style="list-style-type: none">• Task in user's head is hidden• When is a hang-up a success?
Minimize dialogue length	<ul style="list-style-type: none">• Hang up on user immediately?
Maximize channel accuracy	<ul style="list-style-type: none">• Endless confirmations• High rejection rates
Maximize "stickiness": repeat usage	<ul style="list-style-type: none">• For ad-driven services, makes sense• For other services, probably less so
Maximize financial benefit to operator	<ul style="list-style-type: none">• Undoubtedly what companies use• But hard to mimic in research

Evaluations are hard

Each domain/system/operator has unique metrics that seem appropriate

Spoken dialog challenge [1] has 3 tracks:

1. Enter a system
2. Enter a simulated user
3. Evaluate the results

Bottom line: there is no accepted analog to WER, concept accuracy, BLEU Score, MOS, etc. for dialog systems.

[1] <http://www.dialrc.org/sdc>

The "theory of mind" problem

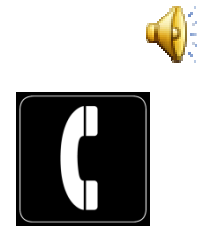
A real human



Graphical user interface



Spoken dialog system



What can she/he/it understand?

Anything I can explain

Only the buttons I can press

The contents of the grammar

How do I know what it can understand?

I'm used to speaking to people

I can see the buttons

I have to make a **conscious decision**

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

Recent results from research

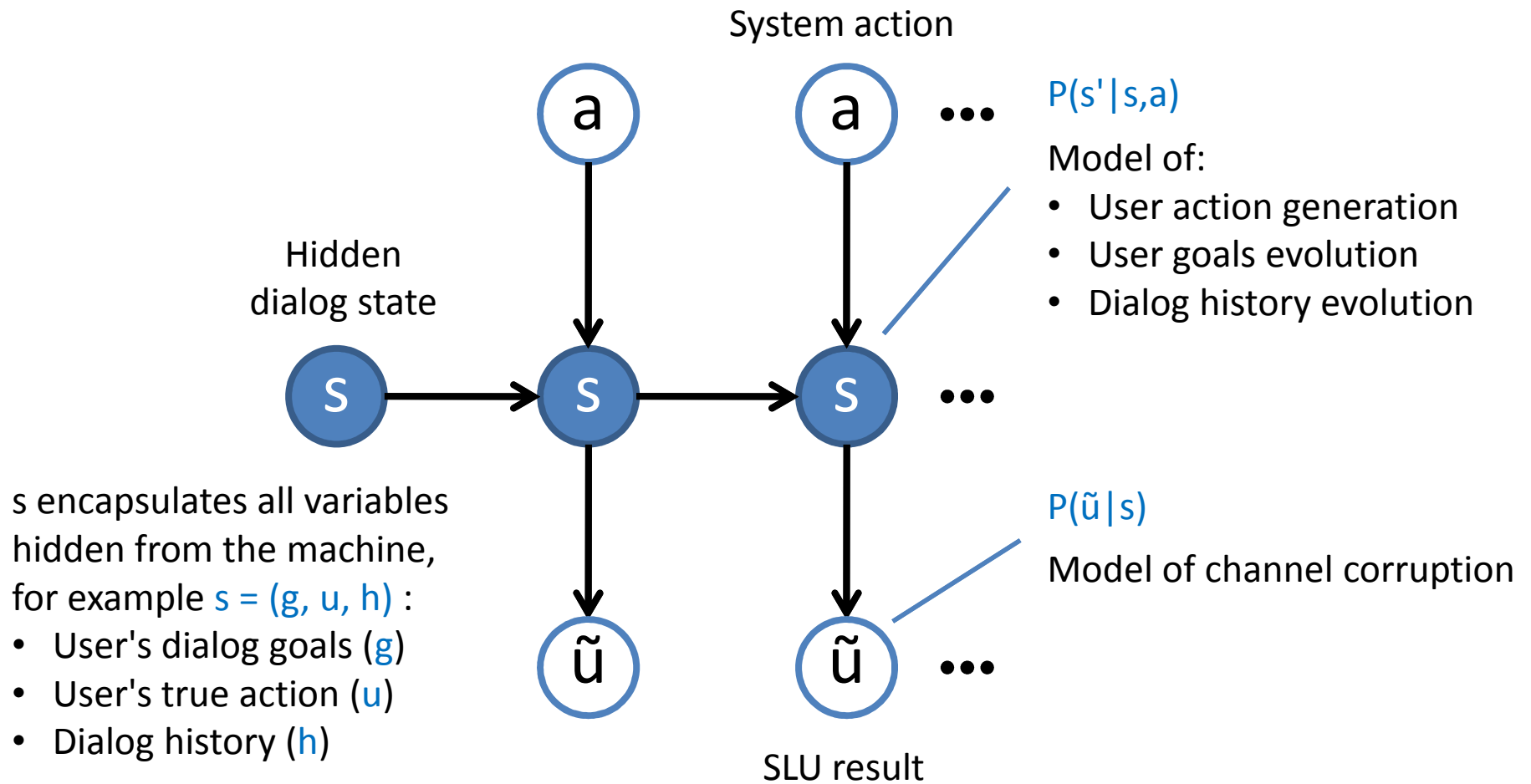
Spoken dialogue systems: challenges, and opportunities for research

Recent results from research

1. Tracking multiple dialogue states
2. Reinforcement learning
3. Incremental processing

Tracking multiple dialogue states: method

Aim: better robustness to errors



rec

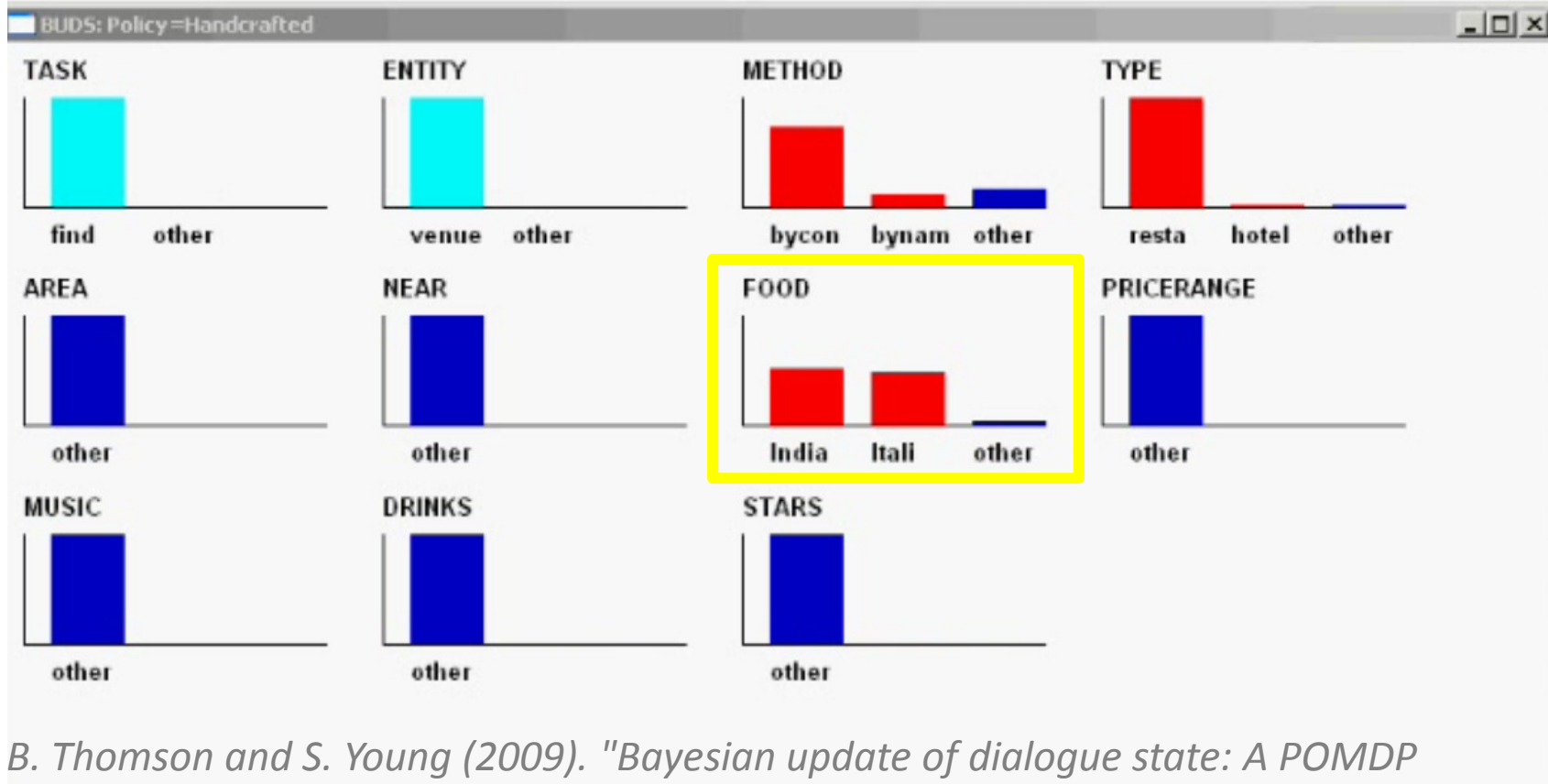
M LOOKING FOR AN INDIAN WHAT NEAR IN ITALIAN RESTAURANT </s>)5 / 10 (<s> HI I'M LOOKING FOR A
N INDIAN WHAT NEAR THE AGAIN ITALIAN RESTAURANT </s>)6 / 10 (<s> HI I'M LOOKING FOR AN INDIAN
WHAT YOU IN THE IN ITALIAN RESTAURANT </s>)7 / 10 (<s> HI I'M LOOKING FOR AN INDIAN WHAT YOU N
EAR THE ITALIAN RESTAURANT </s>)8 / 10 (<s> HI I'M LOOKING FOR AN INDIAN WHAT NEAR THE IN REST
AURANT </s>)9 / 10 (<s> HI I'M LOOKING FOR AN INDIAN WHAT KNOW THE IN ITALIAN RESTAURANT </s>
)10 / 10

<s> HI I'M LOOKING FOR AN INDIAN WHAT NEAR THE ITALIAN RESTAURANT </s>

Status ■ Time 10.7s Score ■ ■ HMM sil NAct 2586 Mode CSSA

Output

**Let me confirm, you are looking
for a venue and that serves
Indian food?**



B. Thomson and S. Young (2009). "Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems." *Computer Speech and Language*, To appear.

Tracking multiple dialogue states: results

		Task completion rates	
	Domain	Single state	Multiple states
[1] Higashinaka et al	Room reservation	88%	91%
[2] Henderson & Lemon	Tourist info	67%	73%
[3] Young et al	Tourist info	66%	79%
[4] Thomson & Young	Tourist info	65%	84%

[1] Ryuichiro Higashinaka, Mikio Nakano, Kiyooki Aikawa, "Corpus-based Discourse Understanding in Spoken Dialogue Systems", ACL, pp240-247, 2003

[2] James Henderson and Oliver Lemon, "Mixture Model POMDPs for Efficient Handling of Uncertainty in Dialogue Management", ACL 2008

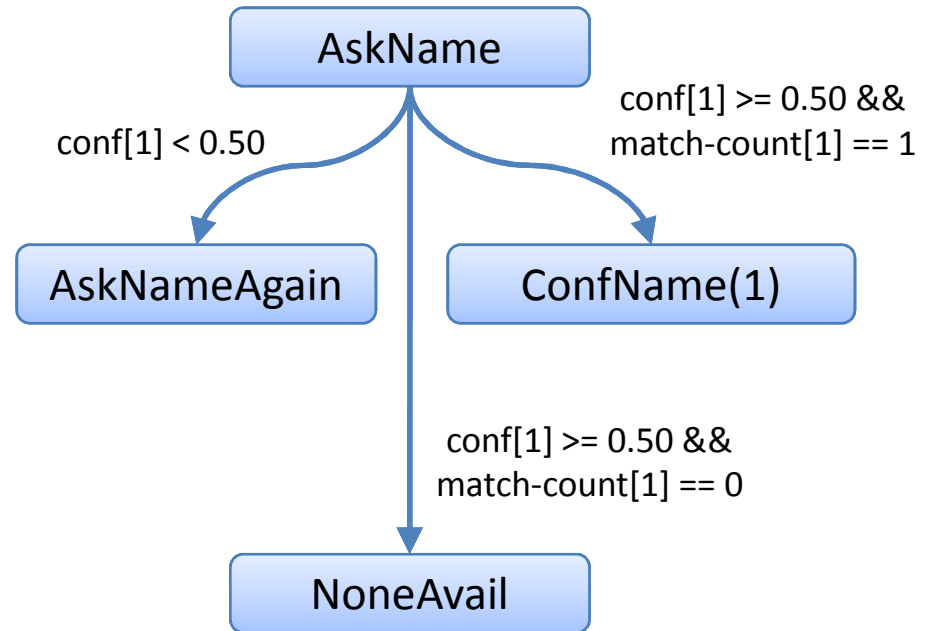
[3] S. Young, M. Gasic, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson and K. Yu (2009). "The Hidden Information State Model: a practical framework for POMDP-based spoken dialogue management." Computer Speech and Language, 24(2): 150-174.

[4] B. Thomson and S. Young (2009). "Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems." Computer Speech and Language, To appear.

Reinforcement learning: background

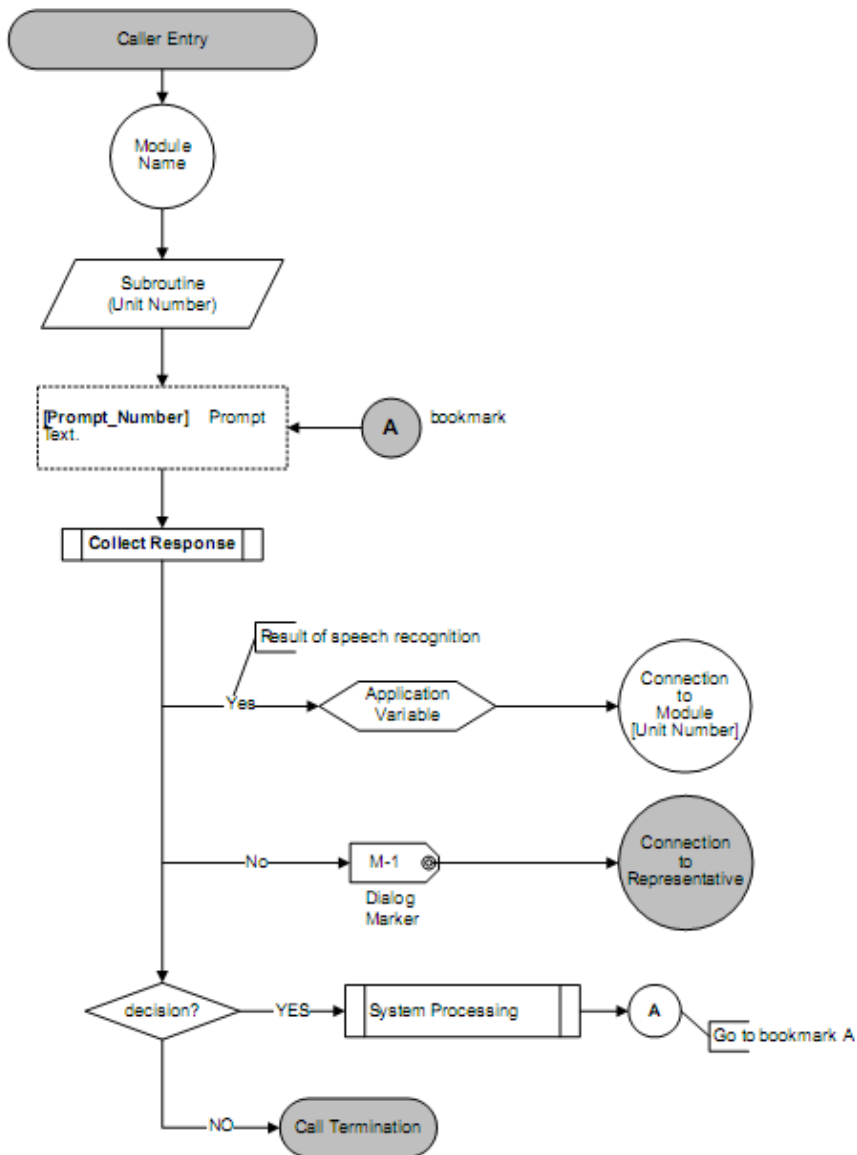
S =

reco[1]:	Jason Williams
conf[1]:	0.43
reco[2]:	Jay Wilpon
conf[2]:	0.05
reco[3]:	Jim Wilson
conf[3]:	0.01
name-tries:	2
confirmed-stat:	No
confirmed-tries:	0
confirmed-ID:	{}
match-count[1]:	1
match[1][1]:	jw4796
location[1][1]:	Florham Park
phone-types[1]:	{office, mobile}
phone-types[2]:	{office}
phone-types[3]:	{mobile}
caller-location:	New York
last-call:	Jay Wilpon



10s – 100s of dialog situations

Reinforcement learning: background



Typical commercial spoken dialog system contains ~100 pages of flowchart

Reinforcement learning: method

Aim: overcome curse of history – build more detailed dialog plans

Designer specifies a reward function with overall goals:

$$R(s,a)$$

Example:

Successful task completion: +20

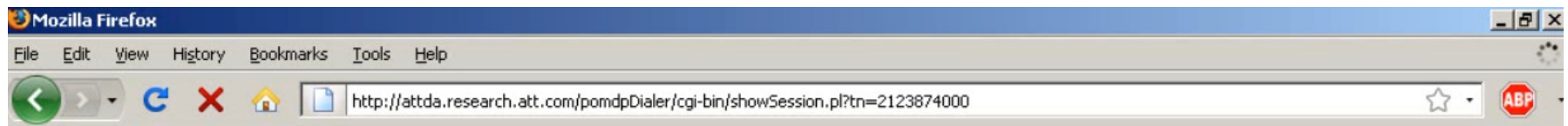
Unsuccessful task completion: -20

Any other dialog action: -1

Optimization chooses policy $\pi(s) = a$ to maximize:

$$E[\sum_t R_t(\pi(a), a)]$$

May require many training dialogues → user simulation



POMDP Dialer : call from 2123874000

<p>Previous system action</p> <p>Sorry, first and last name?</p> <hr/> <p>Recognition result</p> <p>50 jason williams florham_park nj jason williams florham_park nj usa</p>	<p>Belief State</p> <p>Remaining mass [0 partition(s)]</p> <hr/> <p>jason williams florham_park, nj (usa)</p> <p>jason fong columbia, md (usa)</p> <p>juan dong north_sydney, au (iaus)</p> <p>jason downing sacramento, ca (usa)</p> <p>jason kan englewood, co (usa)</p> <p>jason hendrix houston, tx (usa)</p> <p>zhesheng huang middletown, nj (usa)</p>	<div style="border: 2px solid yellow; padding: 5px;"> <p>State Features</p> <p>Best name [red bar]</p> <p>Best phone type [red bar]</p> <p>Phones available both</p> <p>Name confirmed? no</p> <p>Name is ambiguous? no</p> </div> <hr/> <p>Allowed Actions</p> <p>AskName Sorry, first and last name? AskPhoneType jason d williams florham_park new jersey. Say office, cell, or cancel.</p> <div style="border: 2px solid yellow; padding: 5px; margin-top: 10px;"> <p>Action Search</p> <p>Values at point 51 (distance 0.028)</p> <p>18.511 AskPhoneType</p> <p>17.806 ConfirmPhoneType</p> <p>17.546 AskName</p> </div> <hr/> <p>Output system action</p> <p>jason d williams florham_park new jersey. Say office, cell, or cancel.</p>
--	---	--

Jason D. Williams. 2008. *The best of both worlds: Unifying conventional dialog systems and POMDPs.* roc Interspeech, Brisbane, Australia.

Reinforcement Learning: results

	Domain	Task completion	
		Baseline	RL
[1] Singh et al, 2002	Tourist info	20-64%	88%
[2] Lemon et al, 2006	Tourist info	68%	82%
[3] Frampton & Lemon, 2008	Tourist info	82%	91%
[4] Young et al, 2009	Tourist info	64%	79%
[5] Thomson & Young, 2009	Tourist info	84%	75%
[6] Cuayahuitl et al, 2010	Flight booking	94%	95%

[1] S Singh, DJ Litman, M Kearns, and M Walker, "Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system," Journal of Artificial Intelligence Research, 2002.

[2] Oliver Lemon, Kallirroi Georgila, James Henderson, "Evaluating Effectiveness and Portability of Reinforcement Learned Dialogue Strategies with real users: the TALK TownInfo Evaluation", IEEE/ACL Spoken Language Technology, 2006.

[3] Matthew Frampton and Oliver Lemon. 2008. Using dialogue acts to learn better repair strategies. Proc ICASSP 2008.

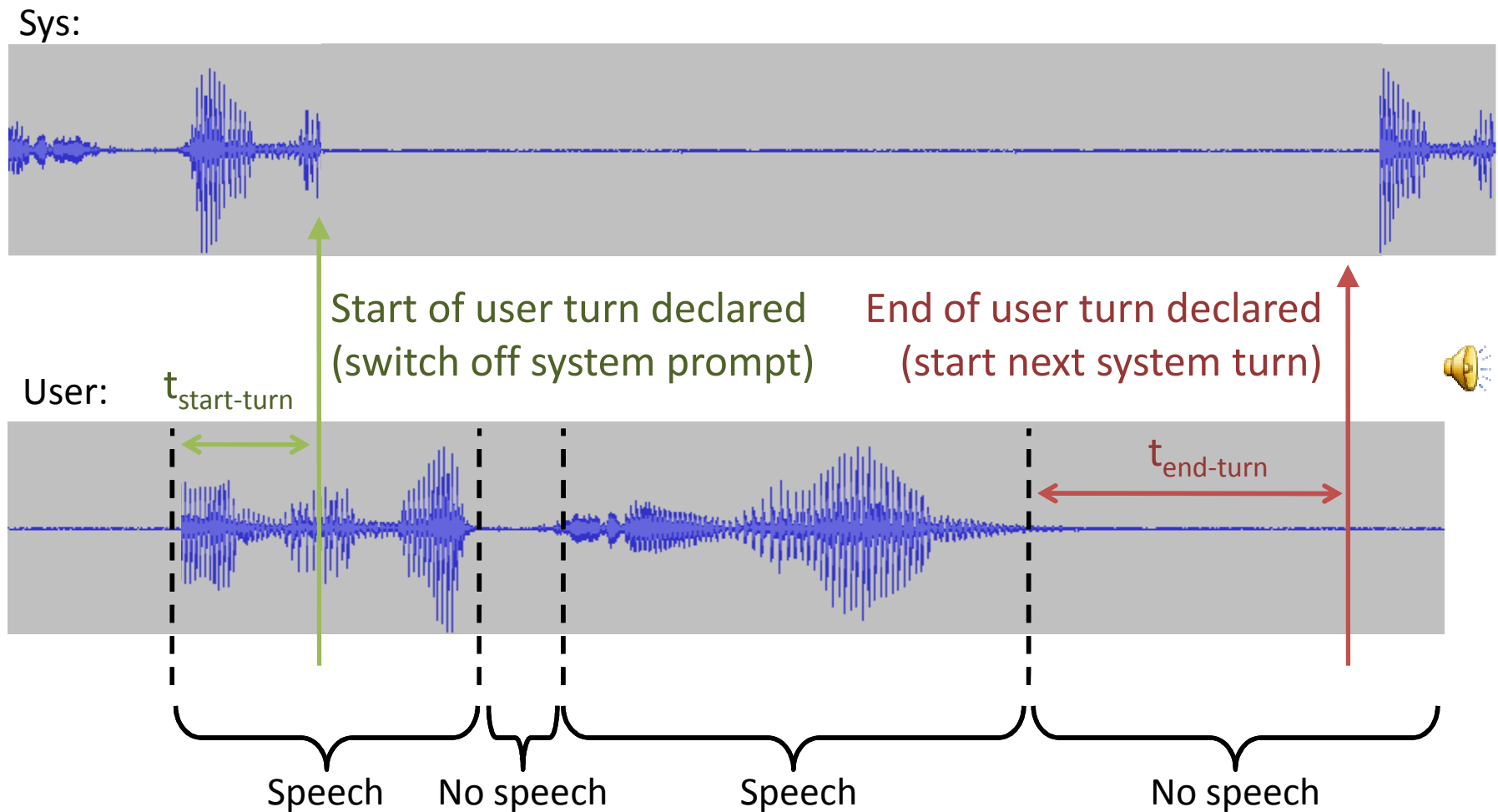
[4] S. Young, M. Gasic, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson and K. Yu (2009). "The Hidden Information State Model: a practical framework for POMDP-based spoken dialogue management." Computer Speech and Language, 24(2): 150-174.

[5] B. Thomson and S. Young (2009). "Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems." Computer Speech and Language, To appear.

[6] Heriberto Cuayahuitl, Steve Renals, Oliver Lemon, Hiroshi Shimodaira, "Evaluation of a hierarchical reinforcement learning spoken dialogue system", Computer Speech and Language, (to appear)

Incremental processing: background

Current systems assume a simple turn-taking model



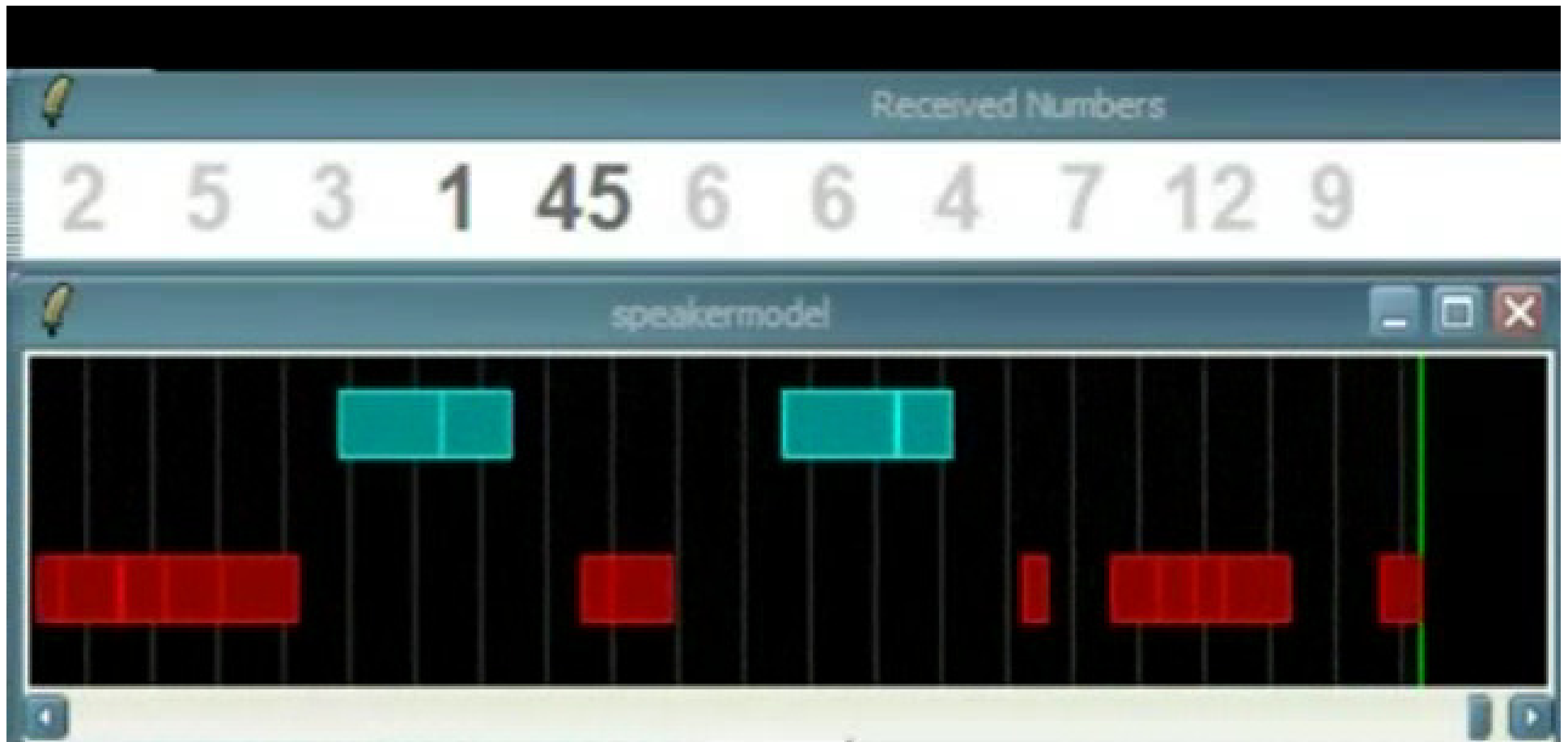
Incremental processing: method

Aim: more natural turn-taking; overcome channel errors

Many threads of work; some themes:

- Continuous ASR and continuous parsing
- Building semantics incrementally
- Predicting end of user speech with richer features
- Predicting when partial results are "stable"
- Predicting when to interrupt the user
- Reference resolution on partial results
- Incorporating ASR N-Best lists
- Metrics for evaluations

... and extending the dialog manager to handle incremental input and output



G Skantze and D Schlangen. (2009). Incremental dialogue processing in a micro-domain. Proc EACL. Athens, Greece.



David DeVault, Kenji Sagae, and David Traum. 2009. Can I finish? Learning when to respond to incremental interpretation results in interactive dialogue. Proc SIGDIAL, London, UK.

Incremental processing: results

	Evaluation type	Results
[1] Ferrer et al., 2002	Static corpus	Reduce end-pointing errors: 9% to 5%
[2] Aist et al., 2007	Observer	20% faster dialogs; higher user satisfaction
[3] DeVault et al, 2009	Static corpus	Collaborative completion of user speech
[4] Skantze & Schlangen, 2009	Interactive	Higher user satisfaction; more human-like

[1] L. Ferrer, E. Shriberg, and A. Stolcke. 2002. Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody in human-computer dialog. Proc. ICSLP, Denver, Colorado.

[2] Gregory Aist, James Allen, Ellen Campana, Carlos Gomez-Gallo, Scott Stoness, Mary Swift and Michael Tanenhaus. 2007. Incremental dialogue system faster than and preferred to its nonincremental counterpart. Proc 29th Cognitive Science Society (CogSci-07), Nashville, Tennessee.

[3] David DeVault, Kenji Sagae, and David Traum. 2009. Can I finish? Learning when to respond to incremental interpretation results in interactive dialogue. Proc SIGDIAL, London, UK.

[4] G Skantze and D Schlangen. (2009). Incremental dialogue processing in a micro-domain. Proc EACL. Athens, Greece.

Prospects for commercial use

Spoken dialogue systems: challenges, and opportunities for research

Life "in the wild" is different

Same system (Let's Go), standard conditions

Metric	Paid subjects	Real users	Ref
WER	17-43%	68%	[1]
Task completion	81%	67%	[2]
DTMF usage	Low	5X more	[2]
First utterance content	Place or destination	Bus route number	[2]
False barge-in	2%	20%	[2]

[1] Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, Maxine Eskenazi. Let's Go Public! Taking a Spoken Dialog System to the Real World. Interspeech 2005.

[2] Hua Ai, Antoine Raux, Dan Bohus, Maxine Eskenazi, Diane Litman. 2008. Comparing Spoken Dialog Corpora Collected with Recruited Subjects versus Real Users. Proc SigDial.

When research systems go public

"Our initial system was designed for mixed-initiative, fairly open-ended dialogs. Although good for exploring and experimenting with natural spoken language interactions, this approach makes the system more fragile in the presence of less-than-optimal conditions... We modified our baseline system towards a very conservative and cautious approach to dialog... we opted for this solution so as to maximize the chances of task success."

Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, Maxine Eskenazi. *Let's Go Public! Taking a Spoken Dialog System to the Real World*. Interspeech 2005.

For more on this issue:

[Bruce Balentine: *It's Better to be a Good Machine than a Bad Person: Speech Recognition and Other Exotic User Interfaces at the Twilight of the Jetsonian Age.*](#)

Data and dialogue systems

In our community it is often said:

"There's no data like more data."

For dialogue systems, this might better be:

"There's no data like continuous access to real people
with real problems."

Interviews with commercial practitioners

- Interviewed 10 practitioners at 9 companies
- All had at least 5 years continuous experience building spoken dialog systems
- All engaged day-to-day with building telephony-based dialog systems
- Anonymous (participants and company names)
 - Understood that results would be used for this talk
- Telephone interviews, 30-60 minutes
- Battery of standardized questions

Scope and limitations of survey

This survey hopes to...

- Identify significance of problems we're trying to solve
- Identify barriers to commercial use
- Guide work to commercialize
- Highlight where education and outreach is needed
- Bound cost-savings associated with these techniques

This survey cannot...

- Comment on other kinds of dialog systems – only telephone-based
- Explain research methods in-depth to the practitioners
- State statistical significance of opinions
- Evaluate broader understanding research provides of human-human communication in general

Multiple dialog states

There is a technique in research which accumulates information from all of the N-Best lists across all recognitions to yield a combined, whole-dialog confidence score. For example, if the same city were recognized twice with low local confidence, after the second recognition it would have a much higher global confidence. This same technique can synthesize together multiple N-Best lists to find the most likely user goals. What is the potential for success of this approach in commercial systems?

6/10 practitioners already aware **Pros:**

Potential	N
High	7
Moderate	3
Low	0

- Right now I'm reading up on this and trying to figure out how we can use it in our applications
- We built an application which uses a more basic version of this approach – it “really helped” task completion
- If this could re-rank N-Best lists, it “could improve dialog quite a lot”
- Our apps are context free with no memory; if we could improve on this, I'm sure it would help performance

Source: Anonymous interviews with 10 industry practitioners

Multiple dialog states

Cons:

- The problem is comprehensibility to testers and designers, so the cost and obstacles may not be paid back in the benefit.
- Might make sense for large-scale, centrally managed systems, but not in one-off smaller systems (N=2)

Conclusions for researchers:

- This is a radically different approach to design vs. industry
- Need to figure out how to communicate this approach to practitioners – including concepts, engineering, APIs, etc.
- Start addressing large-scale problems hard recognition tasks (e.g., business search)
 - Scalability
 - Relationship to search/question-answering

Reinforcement learning

There is a technique in research which tries to learn the best action to take in each dialog state automatically. The idea is for a designer to specify, at each dialog state, a small set of possible actions. Then the designer specifies an overall objective function, such as +10 points for successful completion of the dialog, and -1 point for each question asked. Then the machine tries taking each action in each state, and works out which combination is optimal. What is the potential for success of this approach in commercial systems?

5/10 practitioners already aware

Potential	N
High	5
Moderate	4
Low	1

Pros:

- Designers often have to make many guesses about how people will react. This takes out some of the guesswork.
- “Extremely high” potential for success, especially if this could link abandonment to specific prompts and interactions.
- "I could see how this could save some time" because I wouldn't have to define all the arcs in the callflow.

Reinforcement learning

Cons:

- One obstacle is the level of skills required to do the optimization: doing this without a PhD right now is impossible.
- To get a client to "sign off", you need to make it clear what they're signing off on – documenting all the different versions could be very tedious to produce.
- If there are many paths, how do we know that all paths make sense? Would some paths be crazy?
- See also "VUI Completeness" (Paek and Pierracini. 2008. Automating spoken dialogue management design using machine learning: An industry perspective. Speech Communication.)

Conclusions for researchers:

- Need to be able to assure that all possible user experiences are acceptable
- Incorporation of business rules is crucial (cf Williams, 2008)

Incremental processing

There is a technique in research which processes speech incrementally. Currently most industrial ASR engines make simple speech/no speech decisions for end-pointing. Techniques in research process speech incrementally, accounting for both the system speech output so far and the content of the user's speech. One concrete benefit of this approach is that it could help improve end-pointing by reducing false barge-in and help the system respond to users faster. What is the potential for success of this approach in commercial systems?

1/10 practitioners already aware

Potential	N
High	7
Moderate	2
Low	1

Pros:

- Current approaches are "ridiculously crude"
- I'm "really supportive" of better turn-taking models and I think this line of work "holds real promise" with a "really high" chance of adoption.
- "Anything that can substantially enhance today's end-pointers will do the industry a serious amount of good – this is solving a real problem."

Source: Anonymous interviews with 10 industry practitioners

Incremental processing

Cons:

- The processing is a concern – it might incur some delay
- This relies on good ASR – but ASR is unreliable (and so are confidence scores)
- I don't know how this would really benefit the caller – would this really improve caller experience?

Conclusions for researchers:

- Can simple (application independent) forms of this be packaged for use immediately?
- How can the more sophisticated ideas be incorporated into industrial practices?

What does it cost to build a dialog system?

"Imagine an IVR banking application with about 30 dialog states - directed dialog, no SLM. A client developer will be building the interfaces to the back-end systems. The application receives about 50,000 calls/day, requiring about 350 ports. Roughly how much would this cost to design, build, test, and deploy from scratch?"

Cost to build (N=10):

Median: \$225,000

Mean: \$353,000

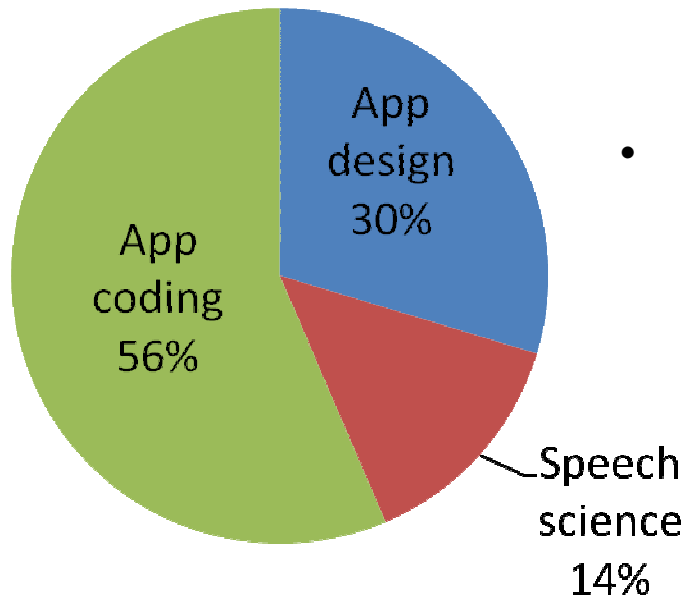
(Services only: not ASR software or hardware)

With SLM (N=9):

Median: +\$68,000

Mean: +\$112,000

Cost detail



Notes:

- 6 practitioners cited high cost of testing (15-50% overall)
- 5 practitioners cited high cost of project management (20-30%)

Very rough business case

A common motivation in research studies:

"The design and development of dialog systems is expensive."

Suppose: reduce coding and design by half: \$100,000

Sell tool for half of that: \$50,000

Number of systems to make \$1MM in revenue: 20/year

Number of systems to make \$10MM in revenue: 200/year

Challenging business case for a stand-alone company focused on
reducing cost of dialog design

*Is there an opportunity for the research community to produce a
tool, embed some of our methods, and in the process expose
them to real callers?*

Conclusions

Spoken dialogue systems: challenges, and opportunities for research

Conclusions

Wide range of applications

- Common thread is understanding spoken language
- May also output spoken language
- May also use GUI, robots, embodied agents, etc.

Building dialog systems is challenging

- Input errors are ubiquitous and hard to detect
- Curse of history
- No single evaluation metric
- Theory of mind

Conclusions

Research is making some headway

- Multiple dialog states
- Reinforcement learning
- Incremental processing

Reasonable prospects for commercial use

For telephone-based spoken dialog systems:

- Important issues remain; to gain adoption, need to take these concerns seriously
- Opportunity for research community to engage by building a dialog tool?

Thanks!

Jason D. Williams

Spoken dialogue systems: challenges, and
opportunities for research



ASRU – December 2009