Microsoft® Research

Faculty Summit 2010

Microsoft® Research
Faculty Summit 2010

Natural User Interfaces with speech

Alex Acero
Speech Technology Group
Microsoft Research Redmond

Kinect: Gesture Recognition with 3D camera

Kinect: Voice Control

Harry Potter and the Sorcerer's Stone available on Zune Marketplace

All titles available on Zune Marketplace
HARRY POTTER characters, names and related indicia are trademarks of and © Warner Bros. Entertainment Inc. Harry Potter Publishing Rights © JKR. Harry Potter and the Sorcerer's Stone © 2001 Warner Bros. Entertainment Inc. All Rights Reserved

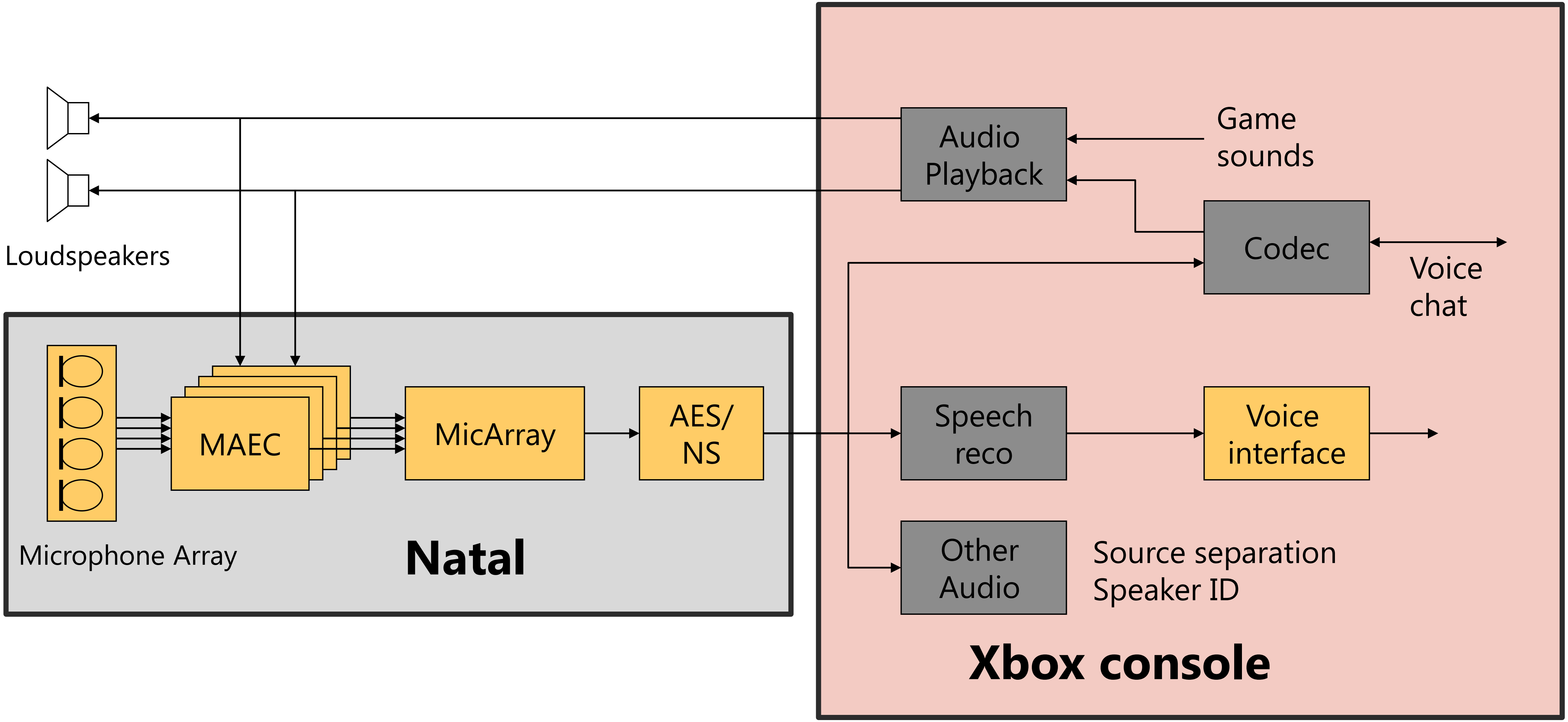
Kinect: Speech recognition

- Speech recognition
  - Complementary to gesture
  - Want to talk to your animal
  - Voice control without on-screen buttons
  - Access long lists
- From headsets to hands free
  - Needs relatively good quality audio!
  - Loud gaming sounds from Xbox
  - Noise and reverberation in the room

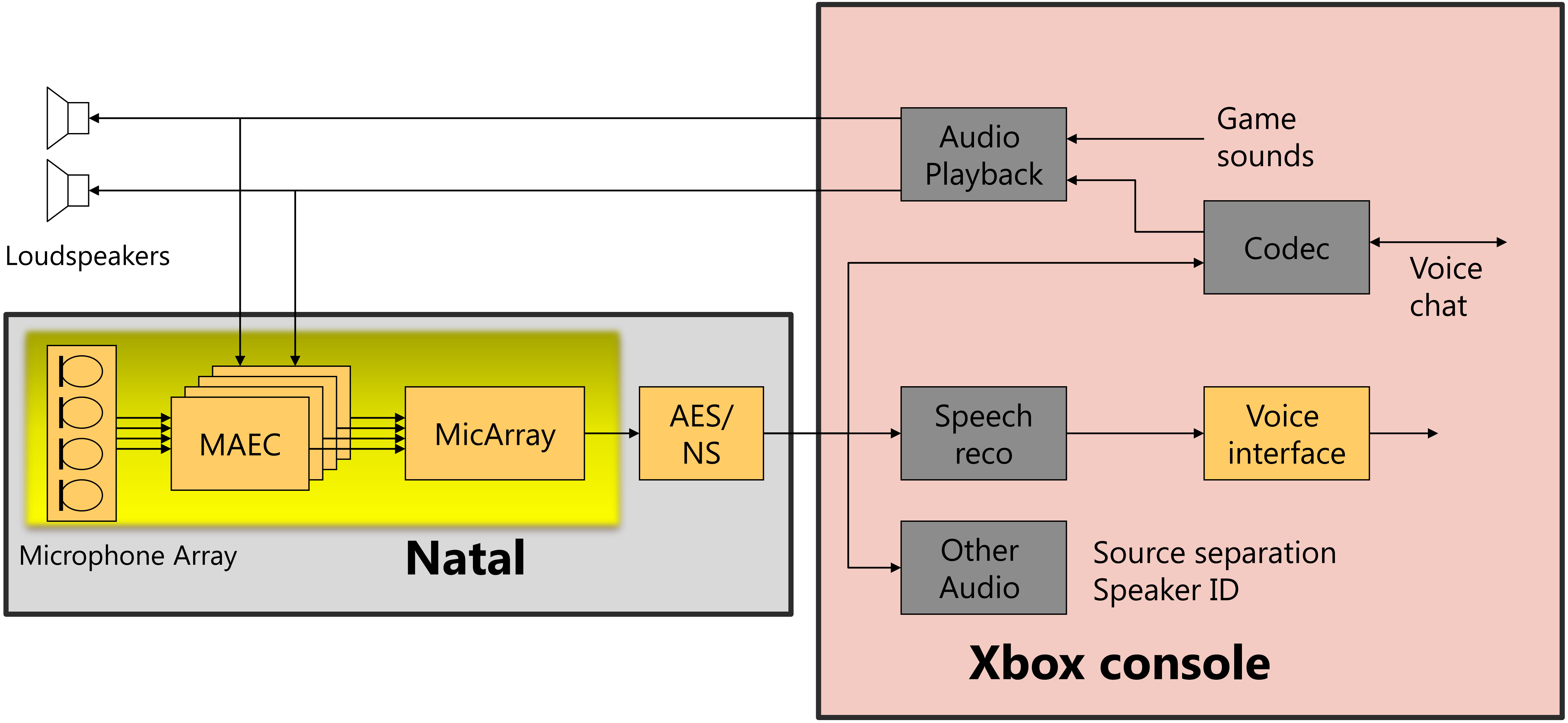
Outline

- Audio processing
- Voice Search
- Robust Voice Control
- Voice interfaces for the automobile
- Voice dialogs
- Error Correction
- Other speech interfaces

Audio Stack

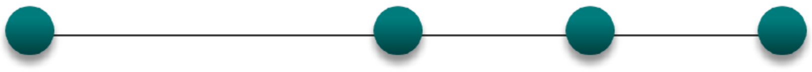

Loudspeakers
Audio Playback
Game sounds
Codec
Voice chat
MAEC
MicArray
AES/ NS
Speech reco
Voice interface
Other Audio
Source separation
Speaker ID
Microphone Array
Natal
Xbox console

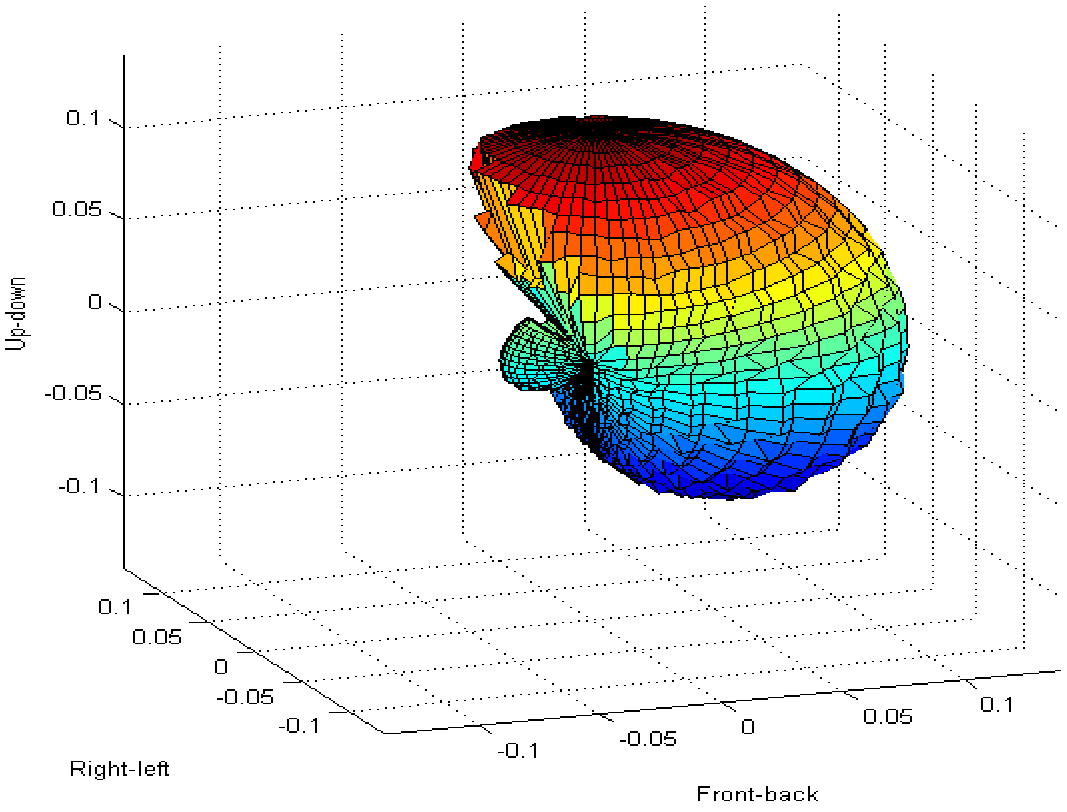
Audio Stack


Loudspeakers
Audio
Playback
Game
sounds
Codec
Voice
chat
MAEC
MicArray
AES/
NS
Speech
reco
Voice
interface
Other
Audio
Source separation
Speaker ID
Microphone Array
Natal
Xbox console

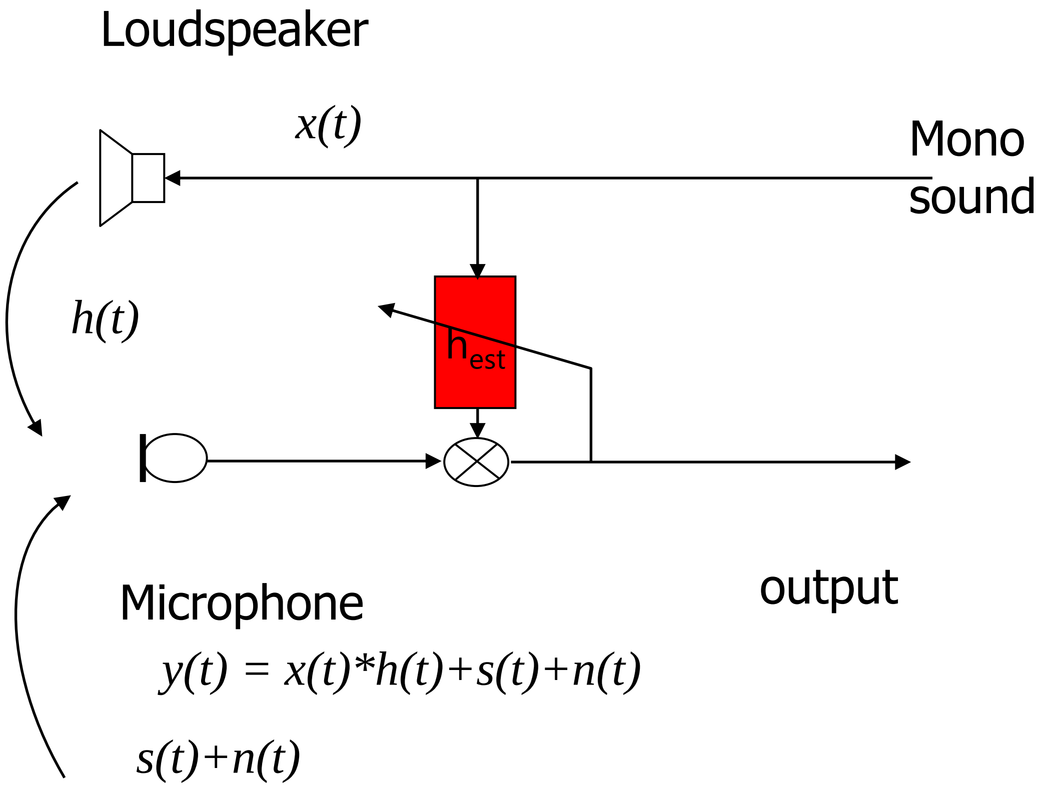
Directional Microphones

- Acoustical design
  - Using the enclosure shape to increase the microphones directivity
- Optimized microphone array geometry
  - Non-equal spacing, covers the entire bandwidth



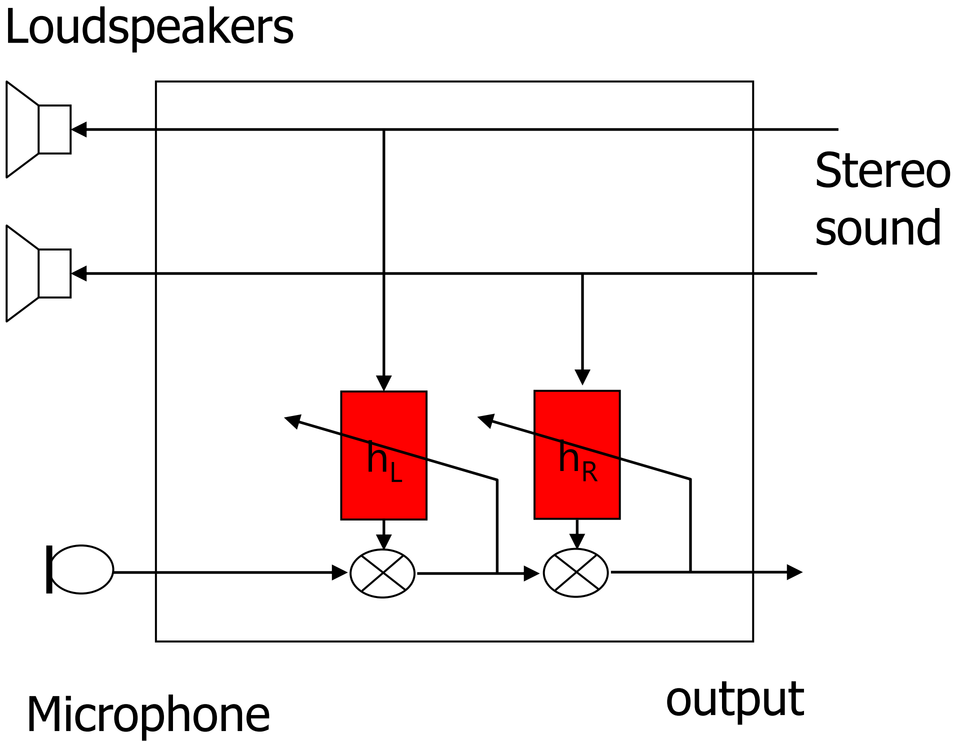


Mono Acoustic Echo Cancellation



- Acoustic echo cancellation
  - Mono AEC – part of each speakerphone

Multichannel Acoustic Echo Cancellation



Loudspeakers

Stereo sound

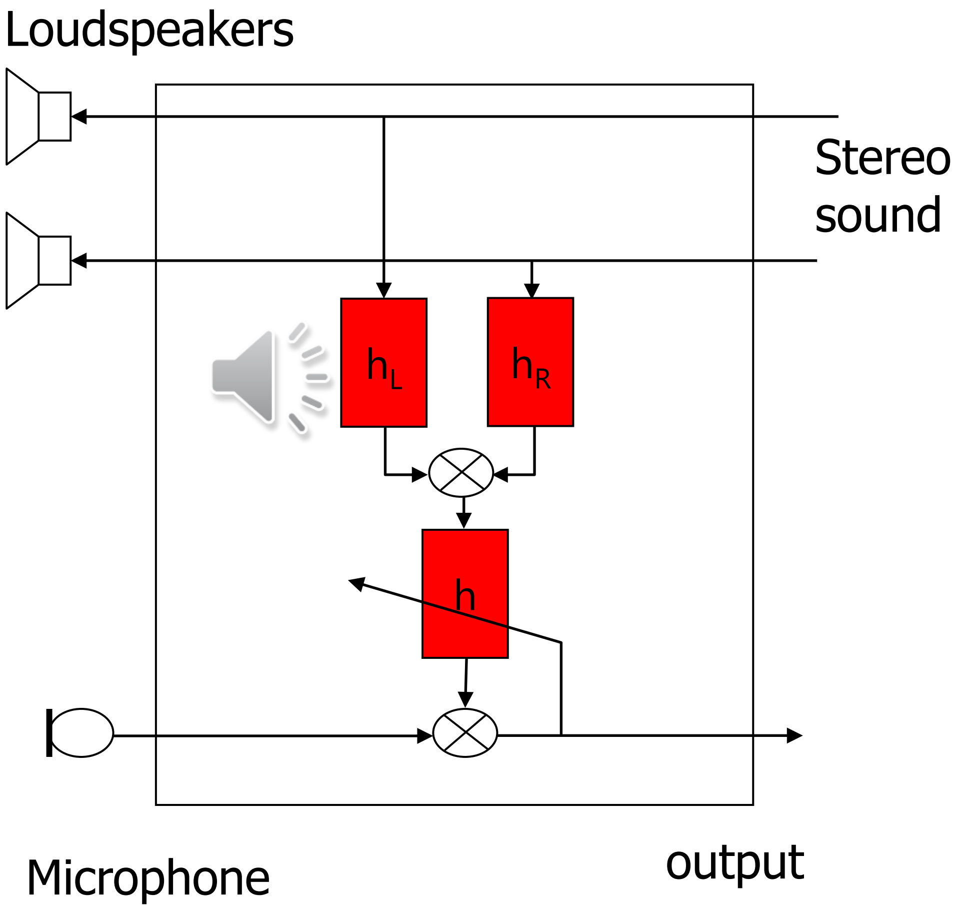
Microphone

output

- Acoustic echo cancellation
  - “Stereo AEC has a non-uniqueness problem that presents a fundamental limitation” (Sondhi et al. Bell Labs, 1995)

Multichannel Acoustic Echo Cancellation

Ivan Tashev 2008

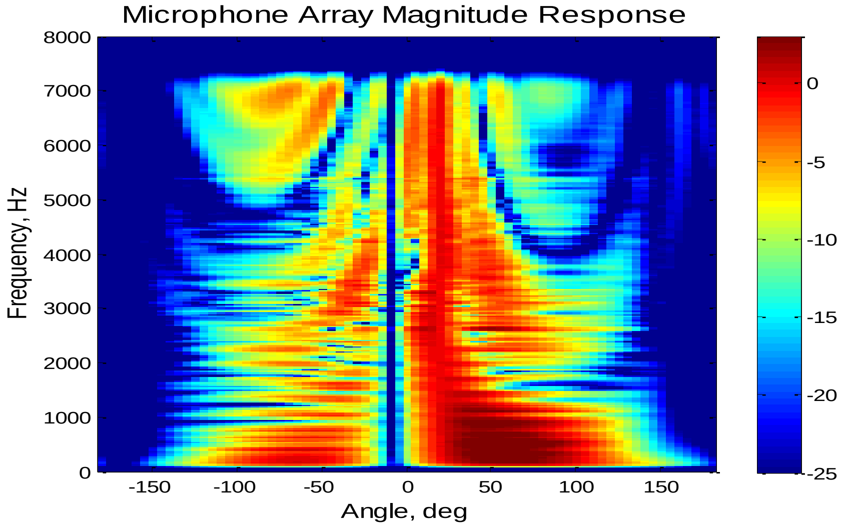


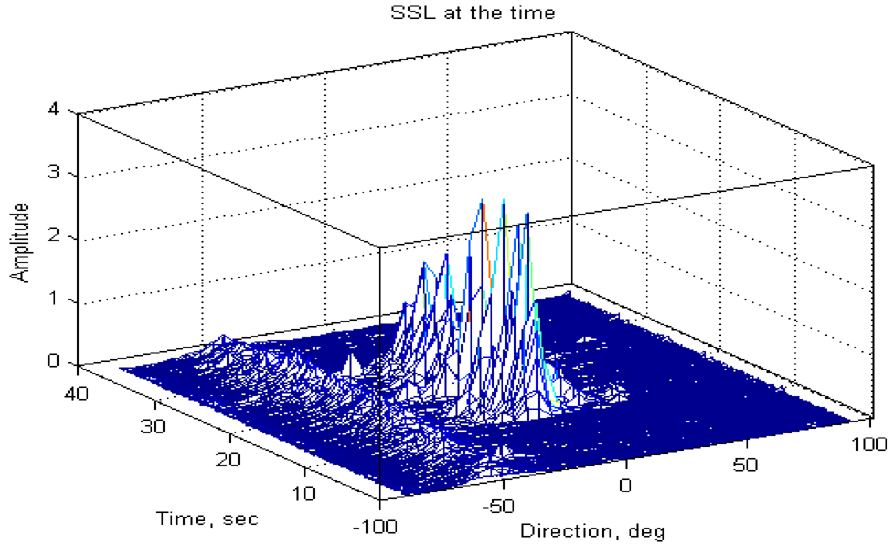
- Acoustic echo cancellation
  - “Stereo AEC has a non-uniqueness problem that presents a fundamental limitation” (Sondhi et al. Bell Labs, 1995)
- Multichannel AEC
  - Use calibration pulses, lock mixing filters, use one adaptive filter
  - Reduces 15-20 dB echo
  - Entire audio pipeline: ~35 dB

Microphone array processing

Ivan Tashev 2008

- Adaptive beamformer
  - Acts as a steerable directional microphone
  - Can suppress interferers as well
  - Reduces 3-6 dB noise
- Spatial filtering
  - Sound source localization per frequency bin
  - Suppresses sounds outside desired direction range
  - Suppresses 6-12 dB noise



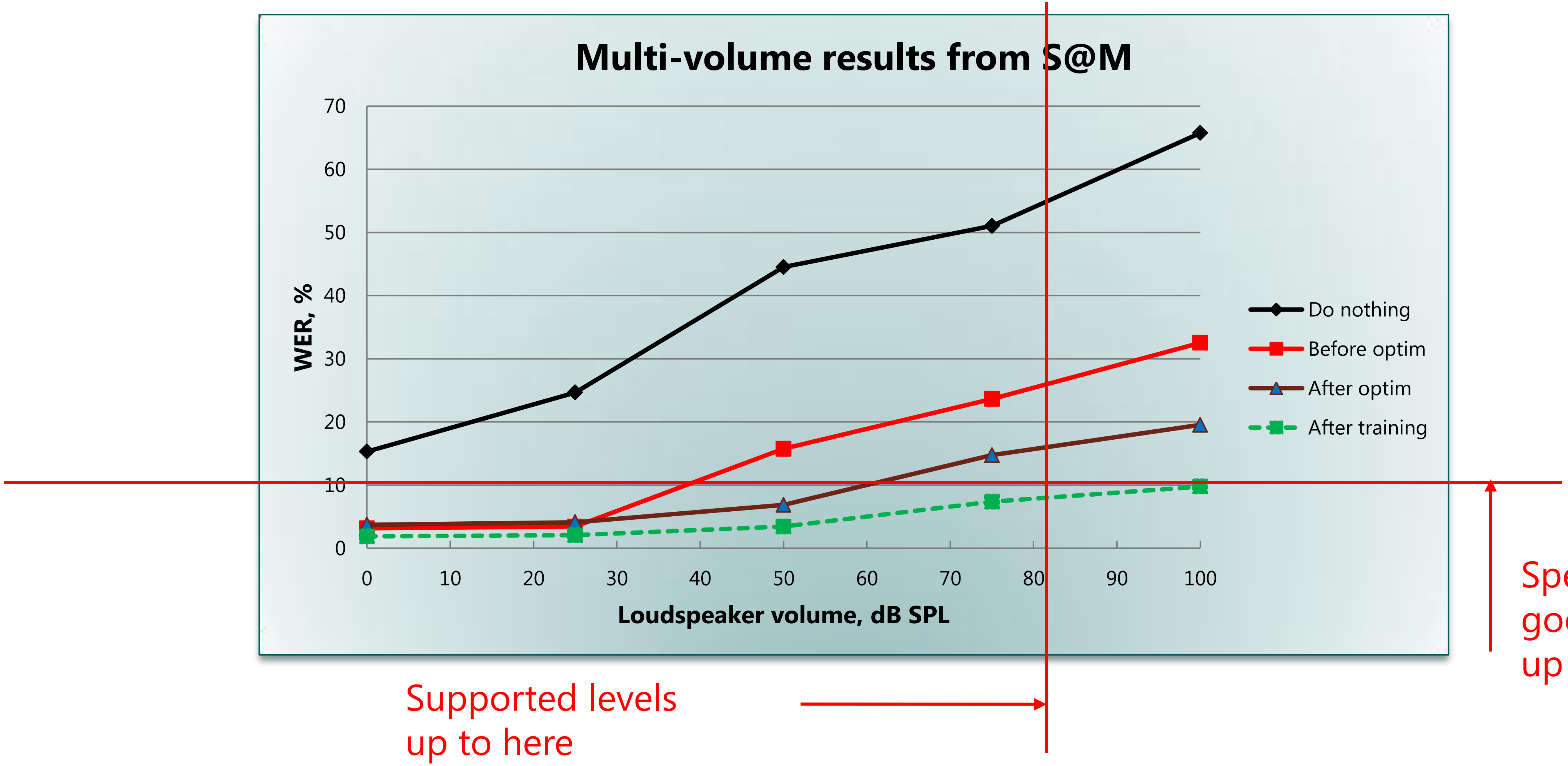


End-to-end optimization

Ivan Tashev 2008

- A chain of optimal processing blocks is suboptimal
- Optimization criterion:
  - Perceptual Evaluation of Sound Quality (PESQ)
- 25 parameters for optimization
  - Time constants, thresholds
- Parallelized processing on cluster
  - Large data corpus
- Results with speech recognizer

End-to-end optimization



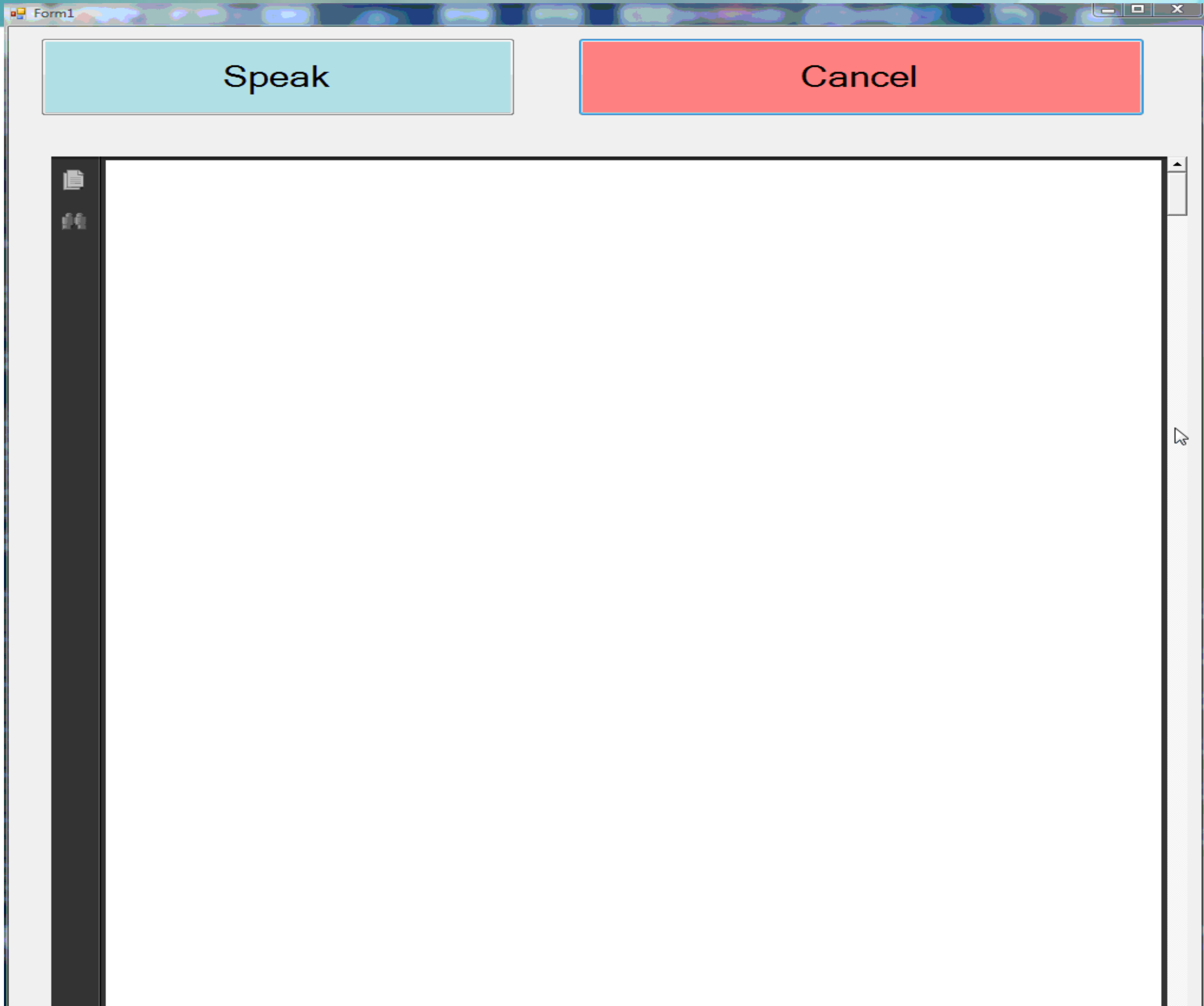
Speech NUI good up to here

Supported levels up to here

Outline

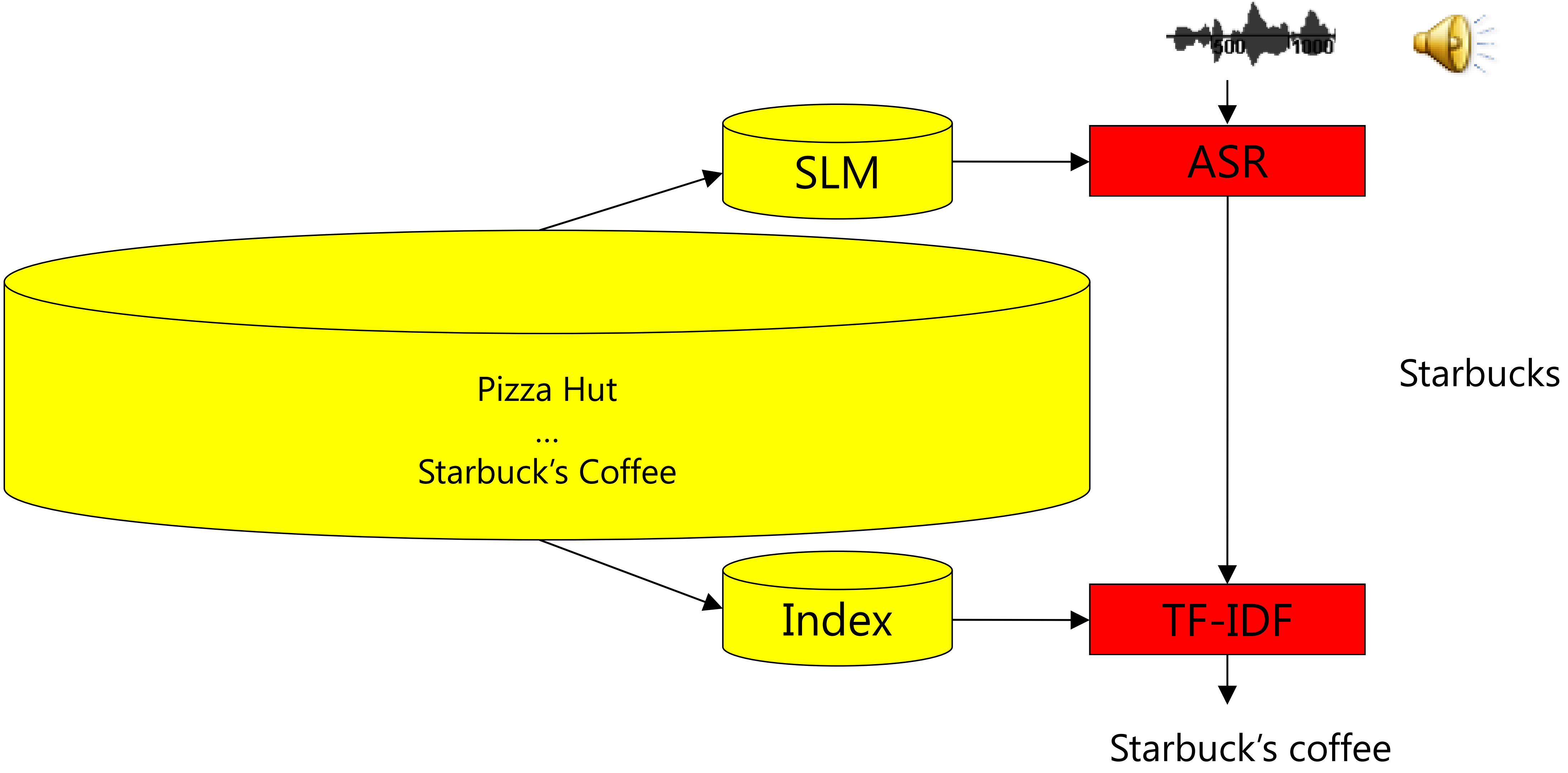
- Audio processing
- Voice Search
- Robust Voice Control
- Voice interfaces for the automobile
- Voice dialogs
- Error Correction
- Other speech interfaces

Voice search for FAQ

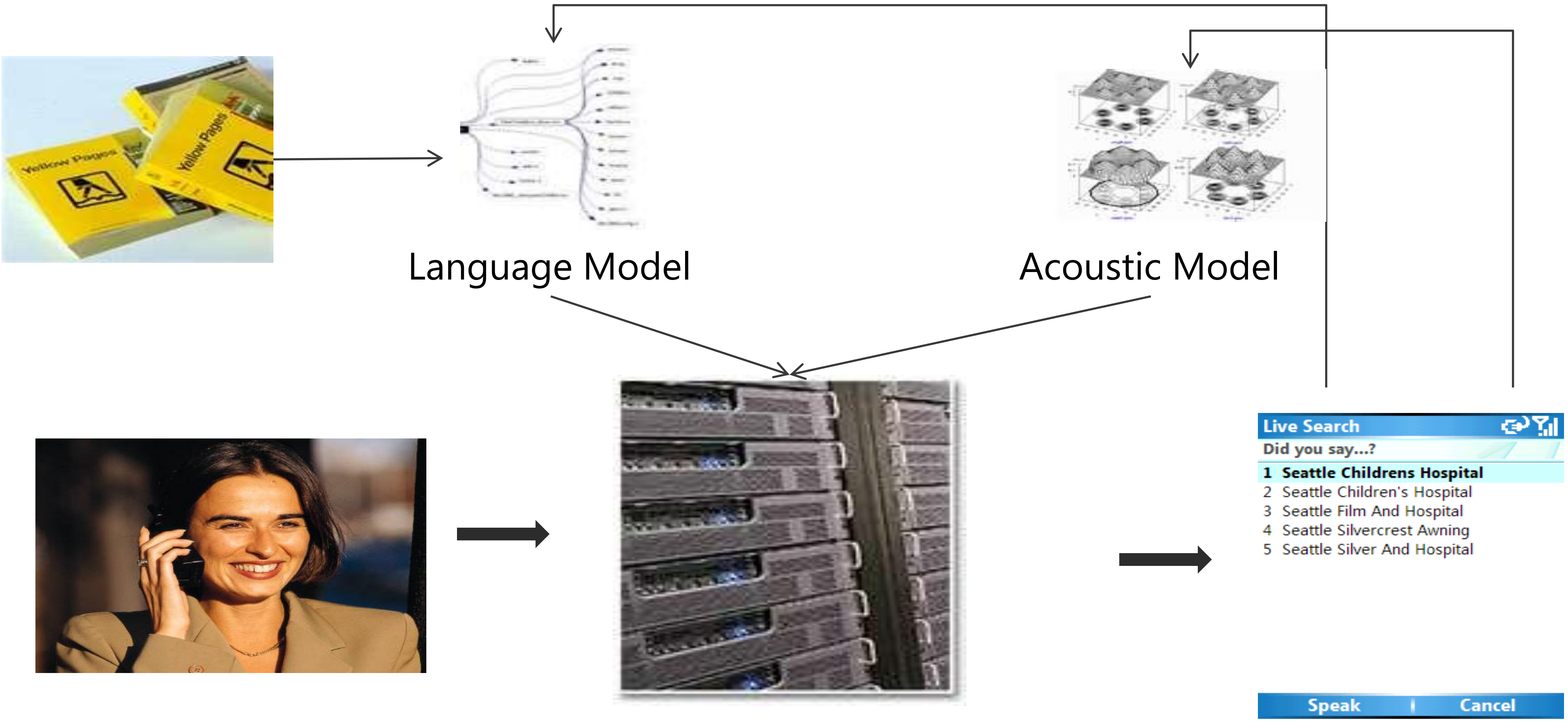


Voice Search architecture

Geoff Zweig, Xiao Li, Patrick Nguyen 2007



Click-Driven Automated Feedback



Language Model

Acoustic Model

Did you say...?

1 **Seattle Childrens Hospital**
2 Seattle Children's Hospital
3 Seattle Film And Hospital
4 Seattle Silvercrest Awning
5 Seattle Silver And Hospital

Outline

- Audio processing
- Voice Search
- Robust Voice Control
- Voice interfaces for the automobile
- Voice dialogs
- Error Correction
- Other speech interfaces

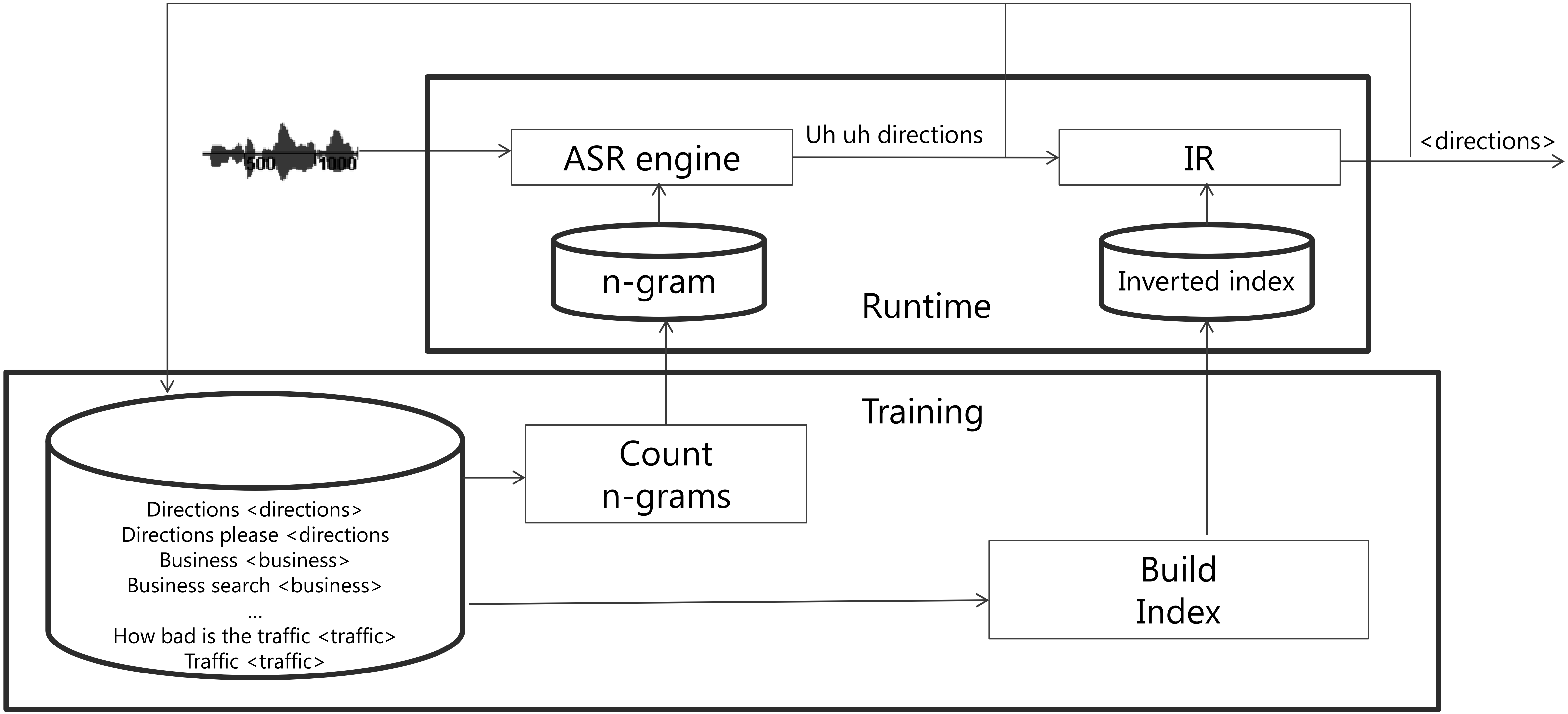
Building Accurate Voice UI is hard

- Traditional Context Free Grammar (CFG):

```
<one-of>
     <item> business search </item>
     <item> search </item>
     <item> biz search </item>
     <item> driving directions </item>
     <item> directions </item>
     <item> traffic </item>
     <item> tell me my choices </item>
     <item> What are my options </item>
     …
</one-of>
```

- Easy to write but fragile

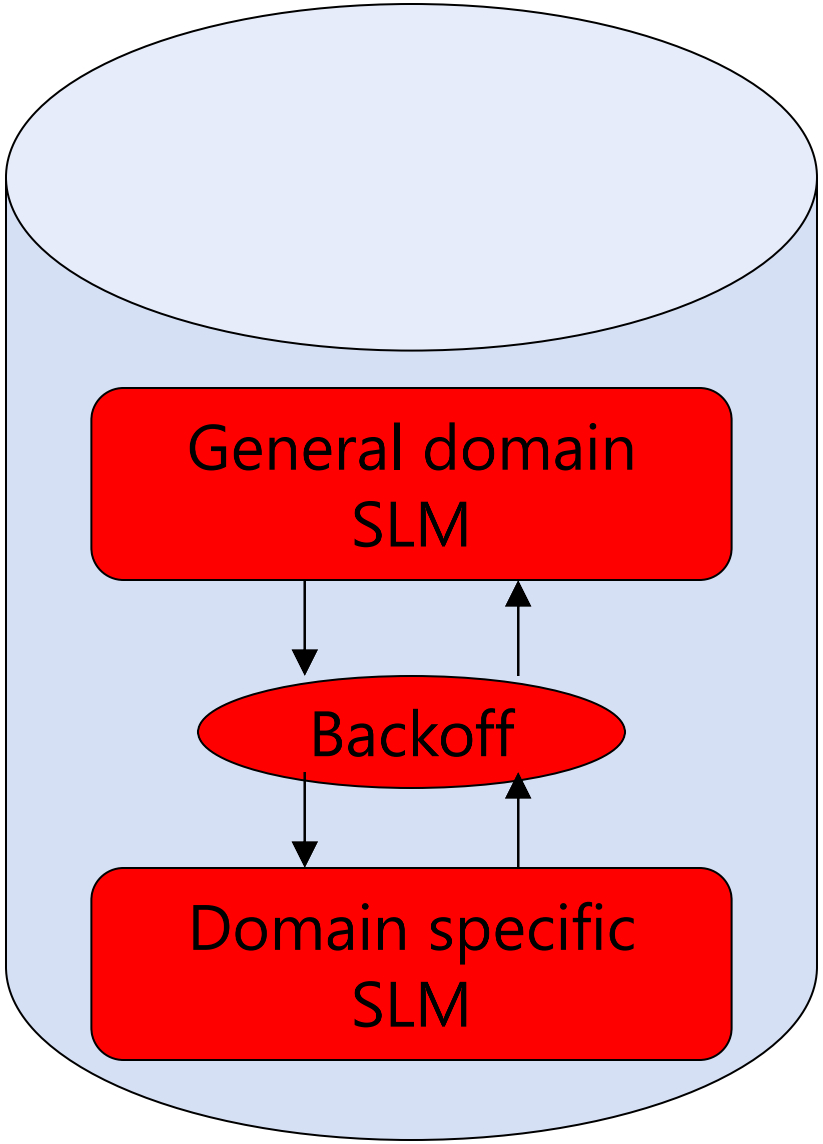
Data driven speech understanding



Example-based SLM

- Interpolation of
  - Large general domain bigram model
  - Small domain specific bigram model

  through backoff state
- Robust SLM with little in-domain data



Information Retrieval (TF-IDF)

- TF-IDF: No need for training data
- If training data is available we can learn a classifier instead
  - Linear classifier. Score for class *i*:

$$S_i = \sum_{j=1}^{N} \lambda_{ij} f_j$$

  - Binary feature f_i: Does word “ticket” occur in class “Reservations”?
  - Weights λ_{ij} are trained through Maximum Entropy

Outline

- Audio processing
- Voice Search
- Robust Voice Control
- Voice interfaces for the automobile
- Voice dialogs
- Error Correction
- Other speech interfaces

SMS in Cars (Ford SYNC)

- SMS are commonly used
- But sending SMS while driving is dangerous
  - and illegal in many countries
- Ford SYNC reads SMS using TTS
- Most SMS only require short replies

FORD SYNC Canned SMS

I’ll BE LATE

MEETING CANCELLED

CAN'T TALK RIGHT NOW

CALL ME

WHERE R YOU?

I NEED MORE DIRECTIONS

THANKS

I AGREE

I DISAGREE

I’M STUCK IN TRAFFIC

C U IN 5(10,15,20) MINUTES

I LOVE YOU

TOO FUNNY

WHAT DO YOU THINK?

ON MY WAY

YOU ARE THE BEST

CALL U LATER

YES

NO

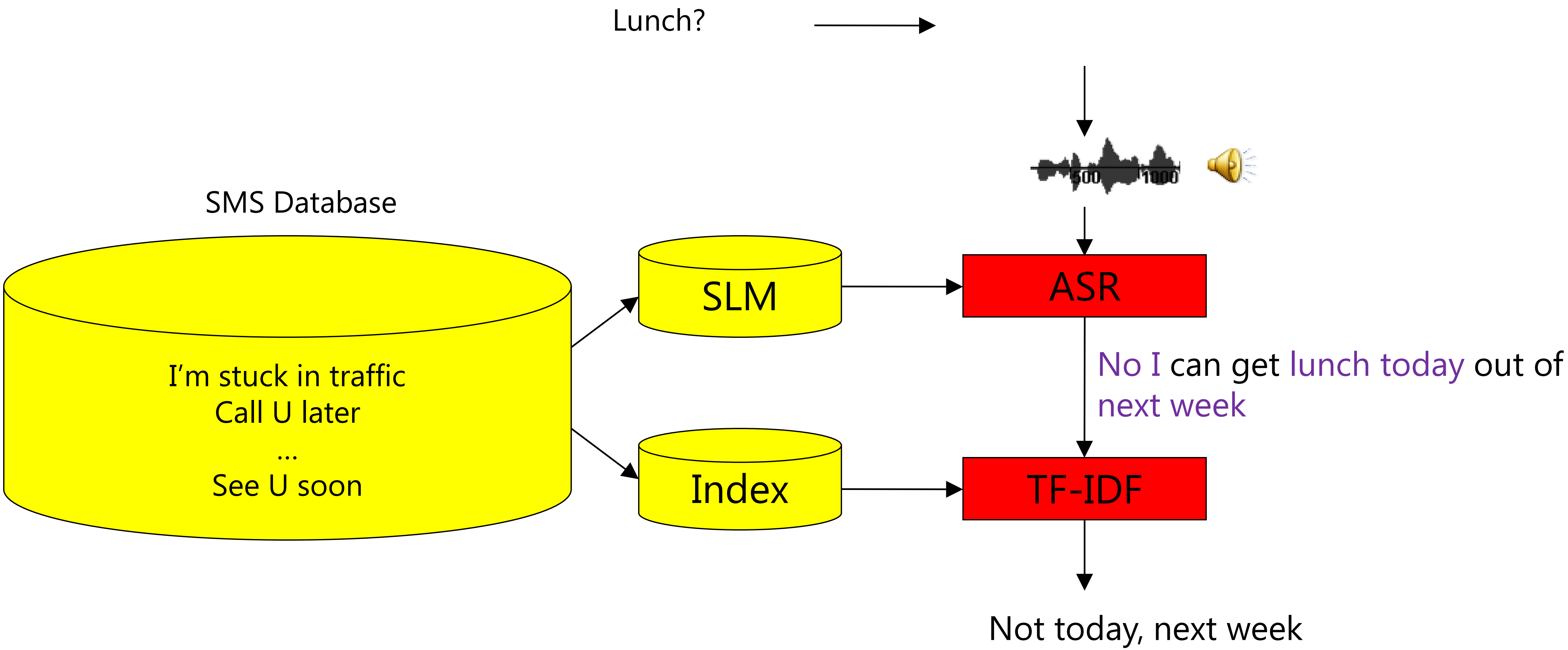
WHY?

TELL ME MORE

CAN'T WAIT TO SEE YOU

SMS Dictation using voice search

YC Ju, 2009



Lunch?

SMS Database

I'm stuck in traffic
Call U later
...
See U soon

SLM

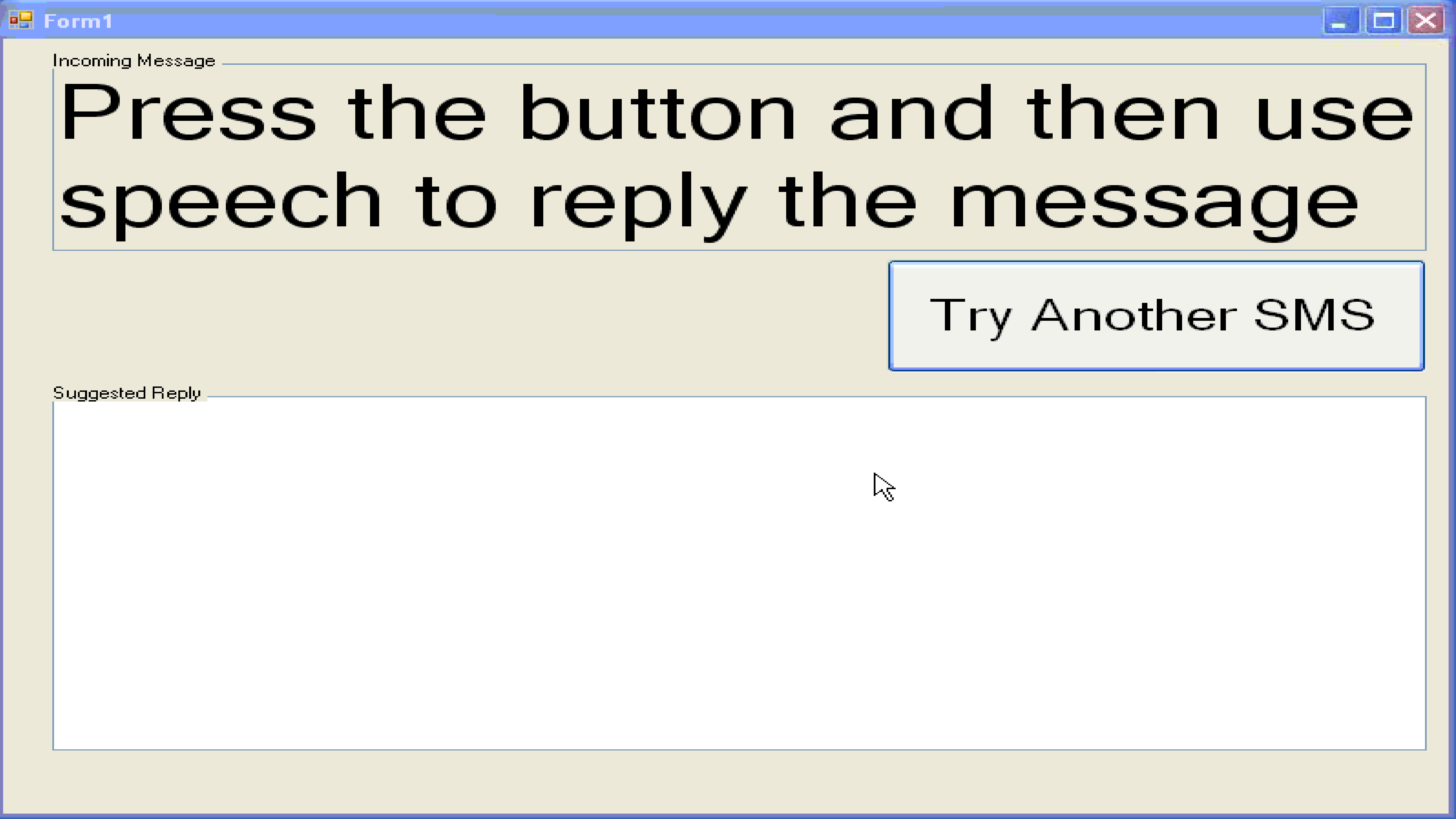
Index

ASR

TF-IDF

No I can get lunch today out of next week

Not today, next week



Incoming Message

Press the button and then use speech to reply the message

Suggested Reply

CommuteUX

Ivan Tashev, Mike Seltzer, YC Ju, 2009

Outline

- Audio processing
- Voice Search
- Robust Voice Control
- Voice interfaces for the automobile
- Voice dialogs
- Error Correction
- Other speech interfaces

Problems with directed dialogs

Who manages the Dialog?

Directed Dialog

- “Who would you like to contact?”
- Finite State Machine
- Simple CFG
- MSConnect

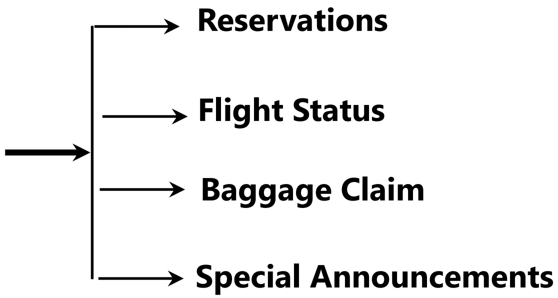
Initiative

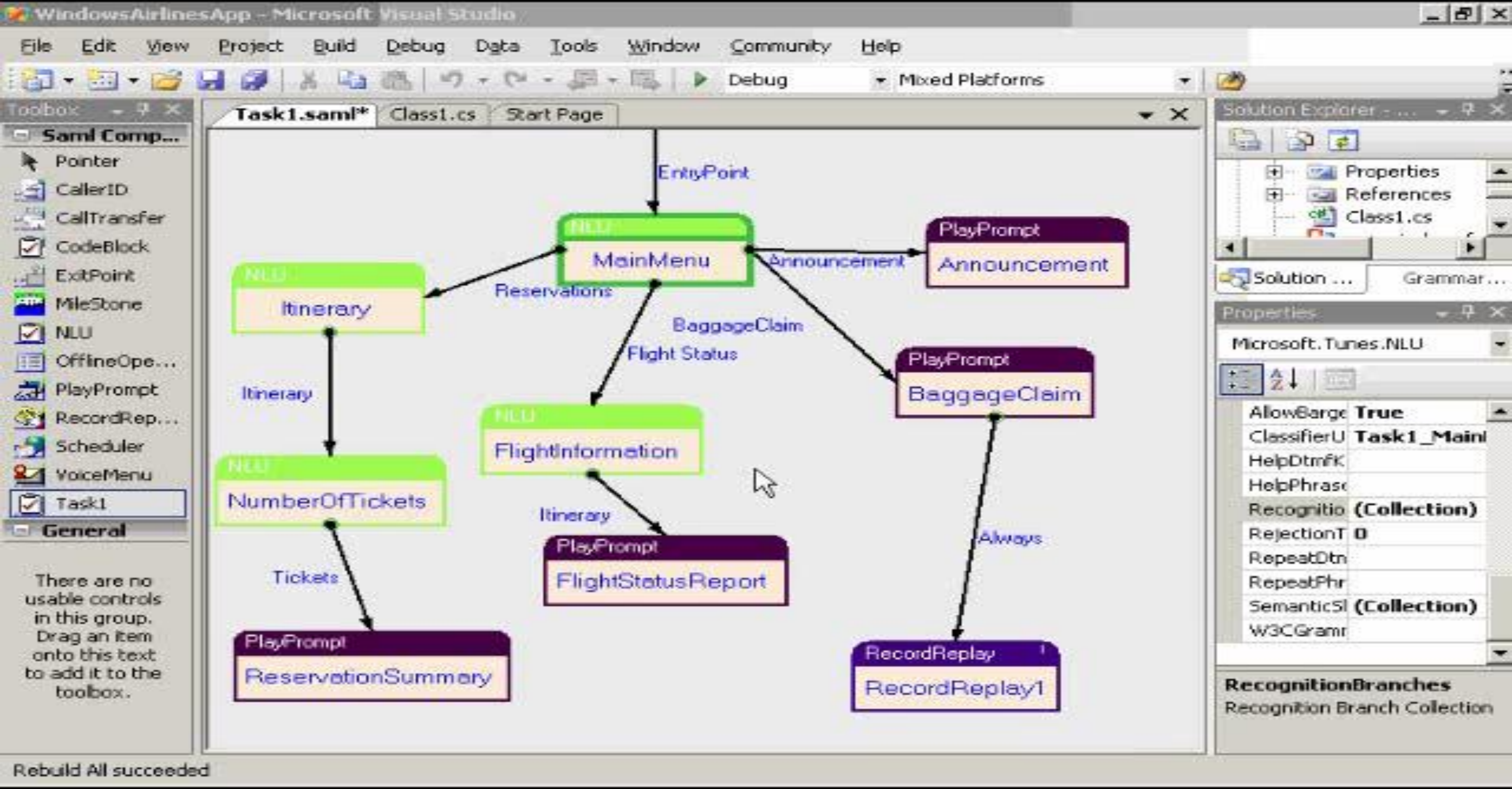
User Initiative Dialog

□ “What can I do for you?”

□ Ngrams

□ Windows Airlines





There are no usable controls in this group. Drag an item onto this text to add it to the toolbox.

Microsoft.Tunes.NLU

AllowBarge	True
ClassifierU	Task1_MainI
HelpDtmfK	
HelpPhrase	
Recognitio	(Collection)
RejectionT	0
RepeatDtn	
RepeatPhr	
SemanticSI	(Collection)
W3CGramr	

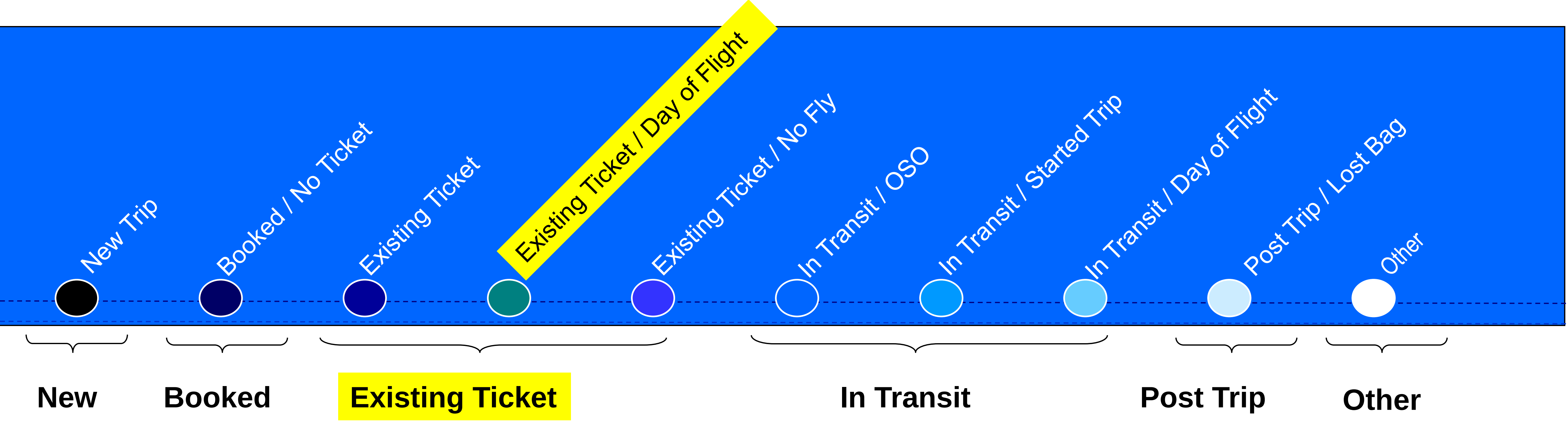
RecognitionBranches
Recognition Branch Collection

Rebuild All succeeded

User-initiative dialogs

- Pros:
  - Can result in a shorter call
  - Can feel more natural
  - Useful when too many choices
- Cons:
  - Requires expensive expertise
  - Could lead to user frustration: system appears human but caller can’t use full natural language

airline traveler journey: a trip



At each stage:

What are the callers *immediate needs*?

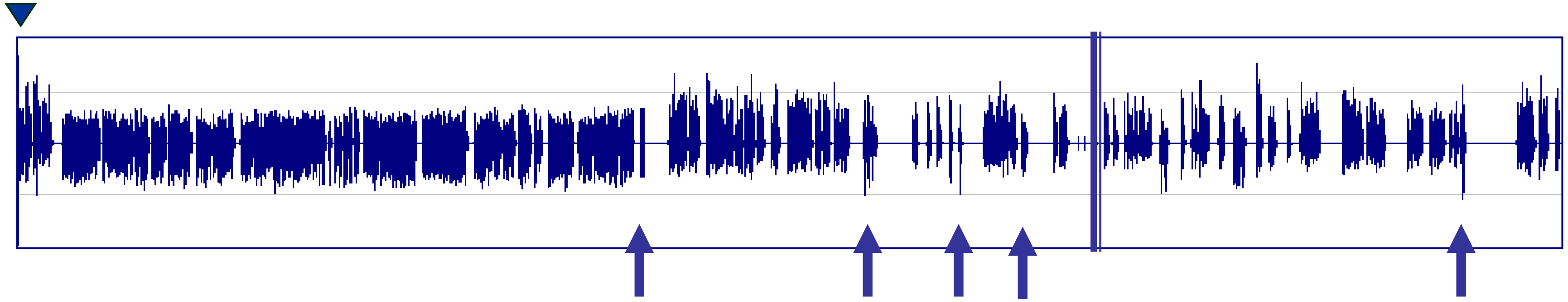
Which *set of tasks* do they want to perform?

How can we use what we already know to *shorten the process*?

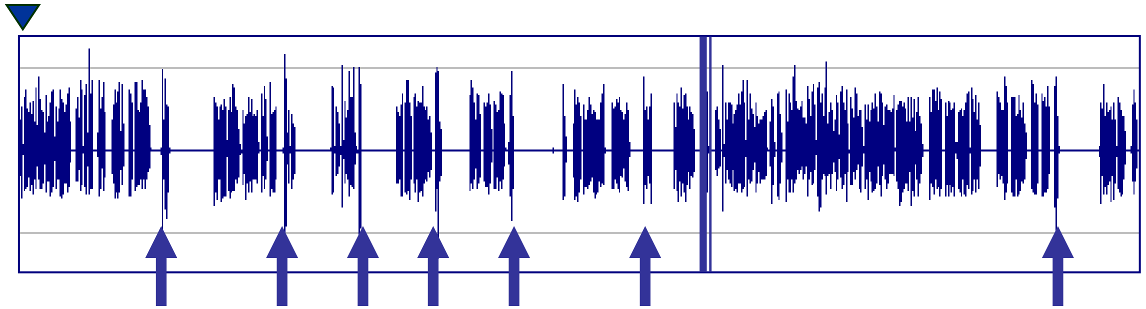
Design for the user

Tellme circa 2000

UNITED

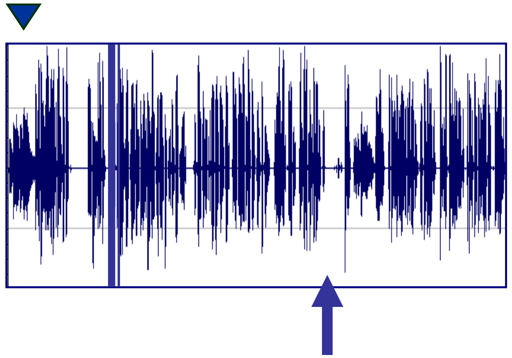


nwa



1:40 min

AmericanAirlines



0:42 min

[Stop]

TellMe

Situated interactions

Dan Bohus 2009

Outline

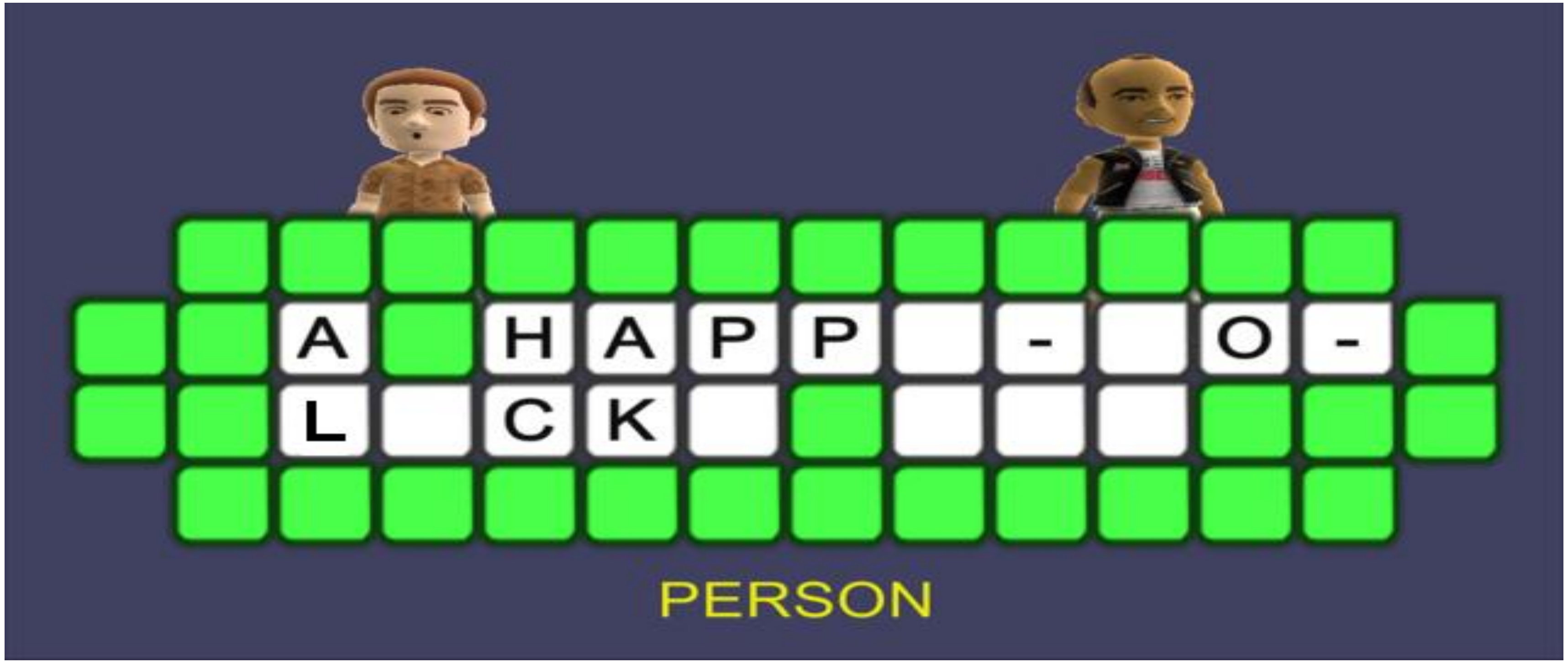
- Audio processing
- Voice Search
- Robust Voice Control
- Voice interfaces for the automobile
- Voice dialogs
- Error Correction
- Other speech interfaces

Outline

- Audio processing
- Voice Search
- Robust Voice Control
- Voice interfaces for the automobile
- Voice dialogs
- Error Correction
- Other speech interfaces

Utterance Verification in Games

YC Ju 2010



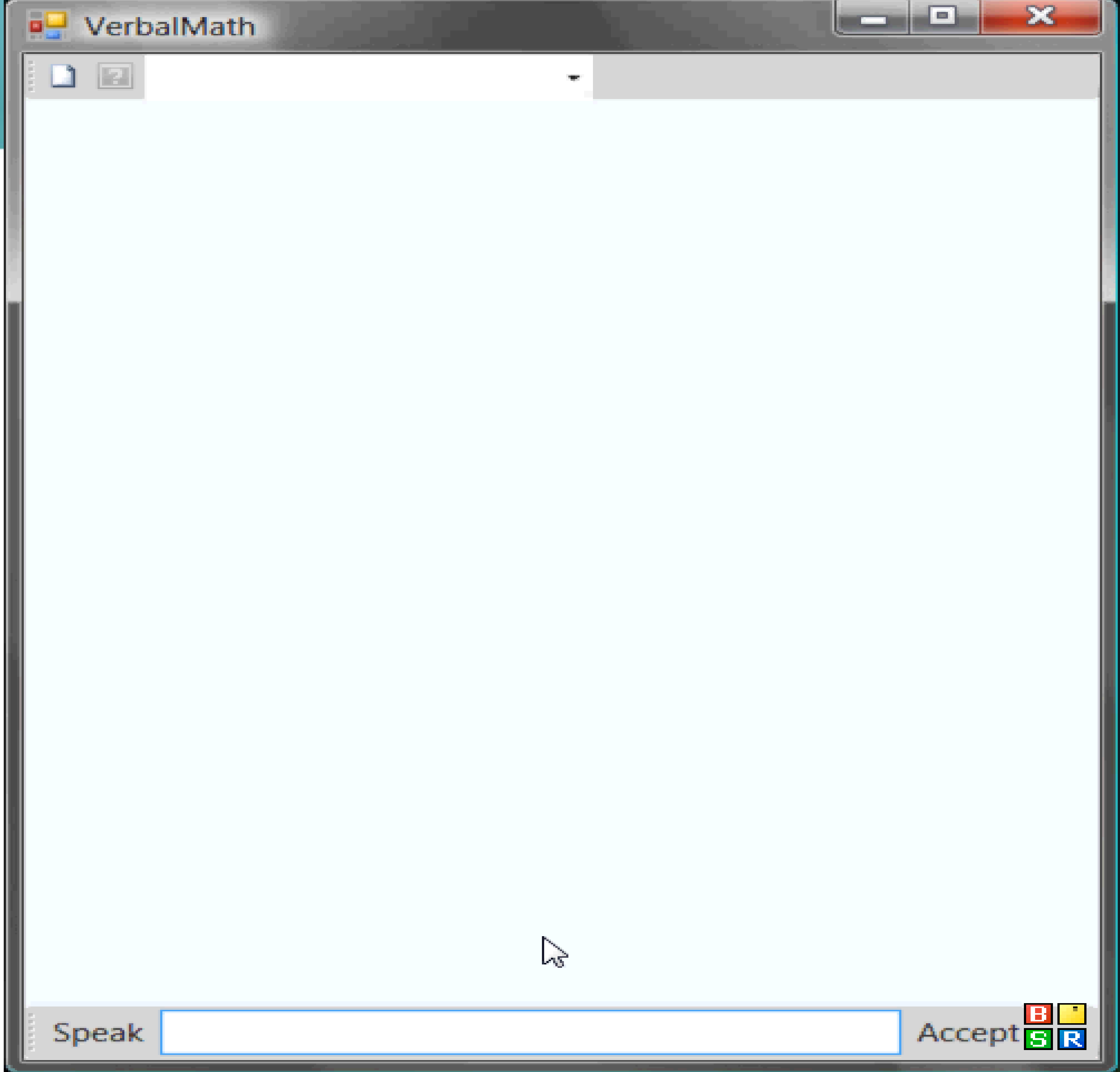
- Engines will typically assign similar scores to “A Happy Go Lucky **Guy**” and “A Happy Go Lucky **Man**”
- **Word-dependent** utterance verification

Speech in Education

Xiaolong Li, 2007

VerbalMath

Xiao Li, 2008



Summary

- Speech for gaming applications require clean audio
- Robust voice control requires flexible grammars
- Voice interface is an interdisciplinary field:
  - Use context
  - Think about the user and collect real data

Thank you

Microsoft®