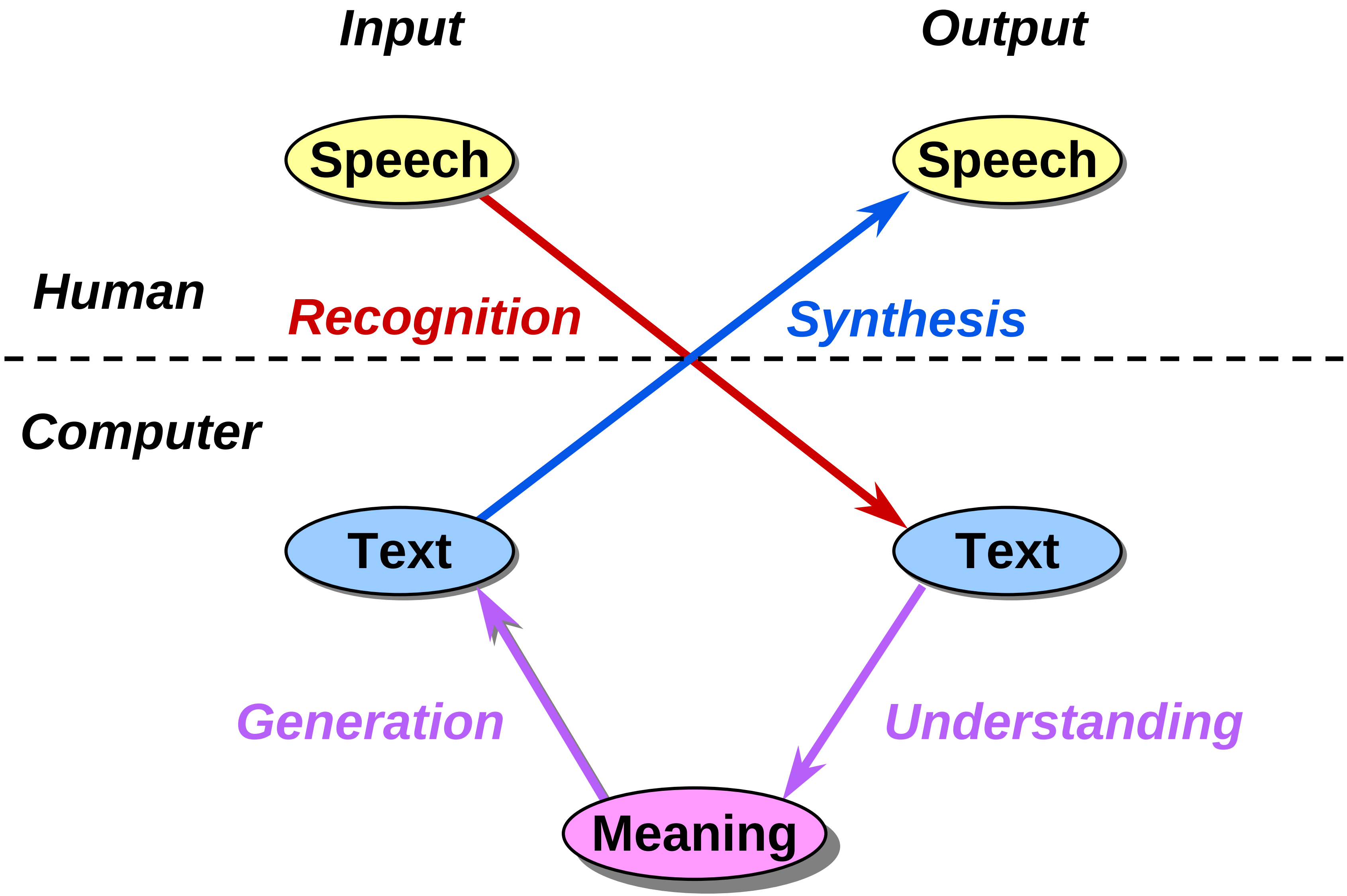
Introduction to Automatic Speech Recognition

- Lectures: Jim Glass & guest lecturers
- Introduction to ASR
  - Problem definition
  - State of the art examples
- Course overview
  - Lecture outline
  - Assignments
  - Term Project
  - Grading

Communication via Spoken Language



Virtues of Spoken Language

Natural:	**Requires no special training**
Flexible:	**Leaves hands and eyes free**
Efficient:	**Has high data rate**
Economical:	**Can be transmitted/received inexpensively**

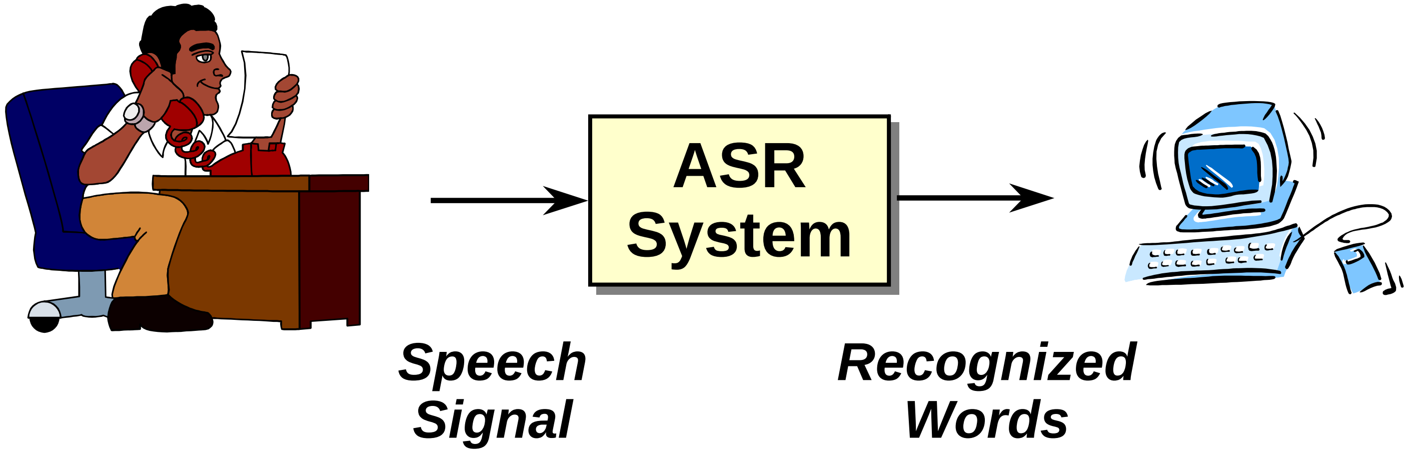
Speech interfaces are ideal for information access and management when:

- The information space is broad and complex,
- The users are technically naive, or
- Only telephones are available.

Diverse Sources of Constraint for Spoken Language Communication

Acoustic:	**human vocal tract**
Phonetic:	**let us pray** **lettuce spray**
Phonological:	**gas shortage** **fish sandwich**
Phonotactic:	**blit vnuk**
Syntactic:	**I am flying to Chicago tomorrow** **tomorrow I flying Chicago am to**
Semantic:	**Is the baby crying** **Is the bay bee crying**
Contextual:	**It is easy to recognize speech** **It is easy to wreck a nice beach**

Automatic Speech Recognition



ASR System

Speech Signal

Recognized Words

- An ASR system converts the speech signal into words
- The recognized words can be
  - The final output, or
  - The input to natural language processing

Application Areas for Speech Based Interfaces

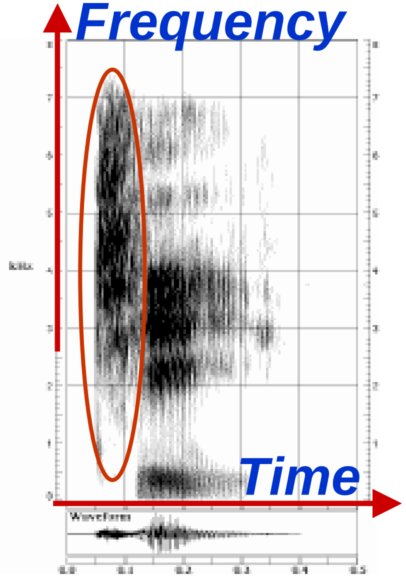
- **Mostly input (recognition only)**
  - Simple command and control
  - Simple data entry (over the phone)
  - Dictation
- **Interactive conversation (understanding needed)**
  - Information kiosks
  - Transactional processing
  - Intelligent agents

Basic Speech Recognition Challenges

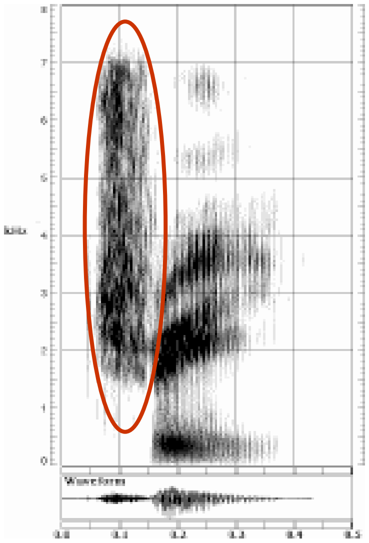
- Co-articulation
- Speaker independence
  - Dialect variations
  - Non-native speakers
- Spontaneous speech
  - Disfluencies
  - Out-of-vocabulary words
- Language modelling
- Noise robustness

Phonological Variation Example

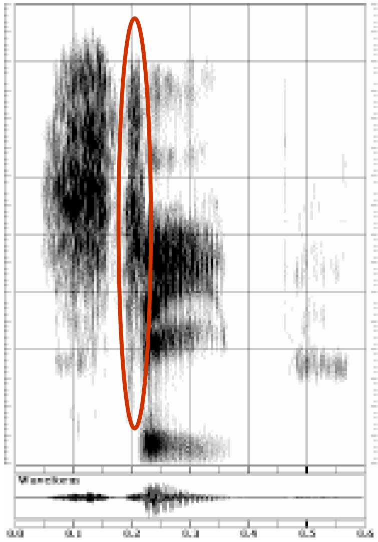
- **The acoustic realization of a phoneme depends strongly on the context in which it occurs**



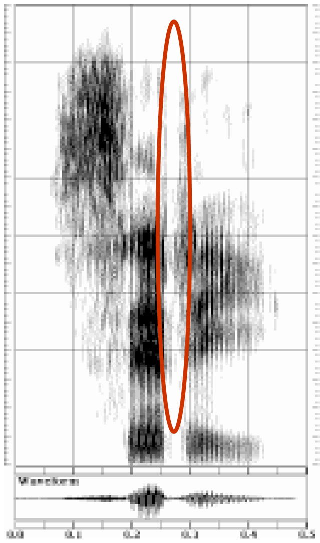
TEA



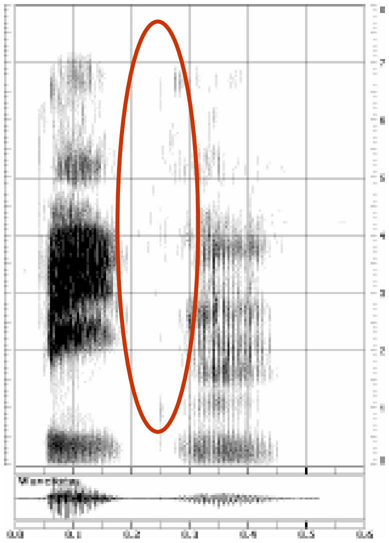
TREE



STEEP



CITY



BEATEN

Examples Contrasting
Read and Spontaneous Speech (Navigation Domain)

Filled and unfilled pauses:	**read, spontaneous**
Lengthened words:	**read, spontaneous**
False starts:	**read, spontaneous**

Sometimes Real Data will Dictate Technology Requirements (City Name Domain)

Technology Required	Example
Simple word spotting	Um, Braintree
Complex word spotting	Eh yes, Avis rent-a-car in Boston Hello, please Brighton, uh, can I have the number of Earthscape, in, uh, on Nonantum Street
Speech understanding	Woburn, uh, Somerville. I'm sorry

Parameters that Characterize the Capabilities of ASR Systems

Parameters	Range
Speaking Mode:	Isolated word to continuous speech
Speaking Style:	Read speech to spontaneous speech
Enrollment:	Speaker-dependent to speaker-independent
Vocabulary:	Small (<20 words) to large (>50,000 words)
Language Model:	Finite-state to context-sensitive
Perplexity:	Small (<10) to large (>200)
SNR:	High (>30dB) to low (<10dB)
Transducer:	Noise-cancelling microphone to cell phone

ASR Trends*: Then and Now

	before mid 70's	mid 70’s - mid 80’s	after mid 80’s
Recognition Units:	whole-word and sub-word units	sub-word units	sub-word units
Modeling Approaches:	heuristic and ad hoc	template matching	mathematical and formal
	rule-based and declarative	deterministic and data-driven	probabilistic and data-driven
Knowledge Representation:	heterogeneous and complex	homogeneous and simple	homogeneous and simple
Knowledge Acquisition:	intense knowledge engineering	embedded in simple structure	automatic learning

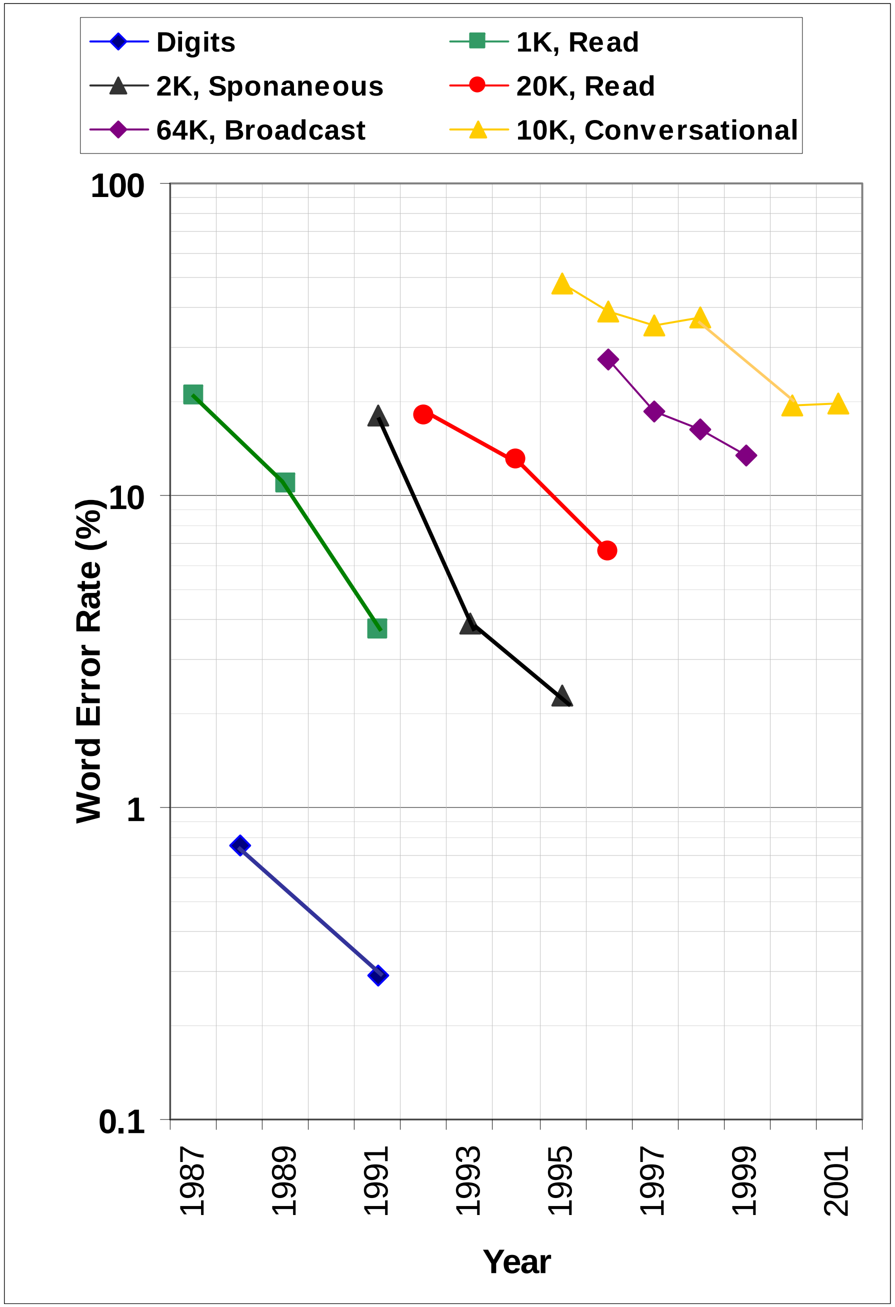
* There are, of course, many exceptions.

Speech Recognition: Where Are We Now?

- **High performance, speaker-independent speech recognition is now possible**
  - **Large vocabulary (for cooperative speakers in benign environments)**
  - **Moderate vocabulary (for spontaneous speech over the phone)**
- **Commercial recognition systems are now available**
  - **Dictation (e.g., ~~Dragon~~, IBM, ~~L&H~~, ~~Philips~~) Scansoft**
  - **Telephone transactions (e.g., AT&T, Nuance, Philips, SpeechWorks, TellMe, etc.)**
- **When well-matched to applications, technology is able to help perform real work**

Examples of ASR Performance

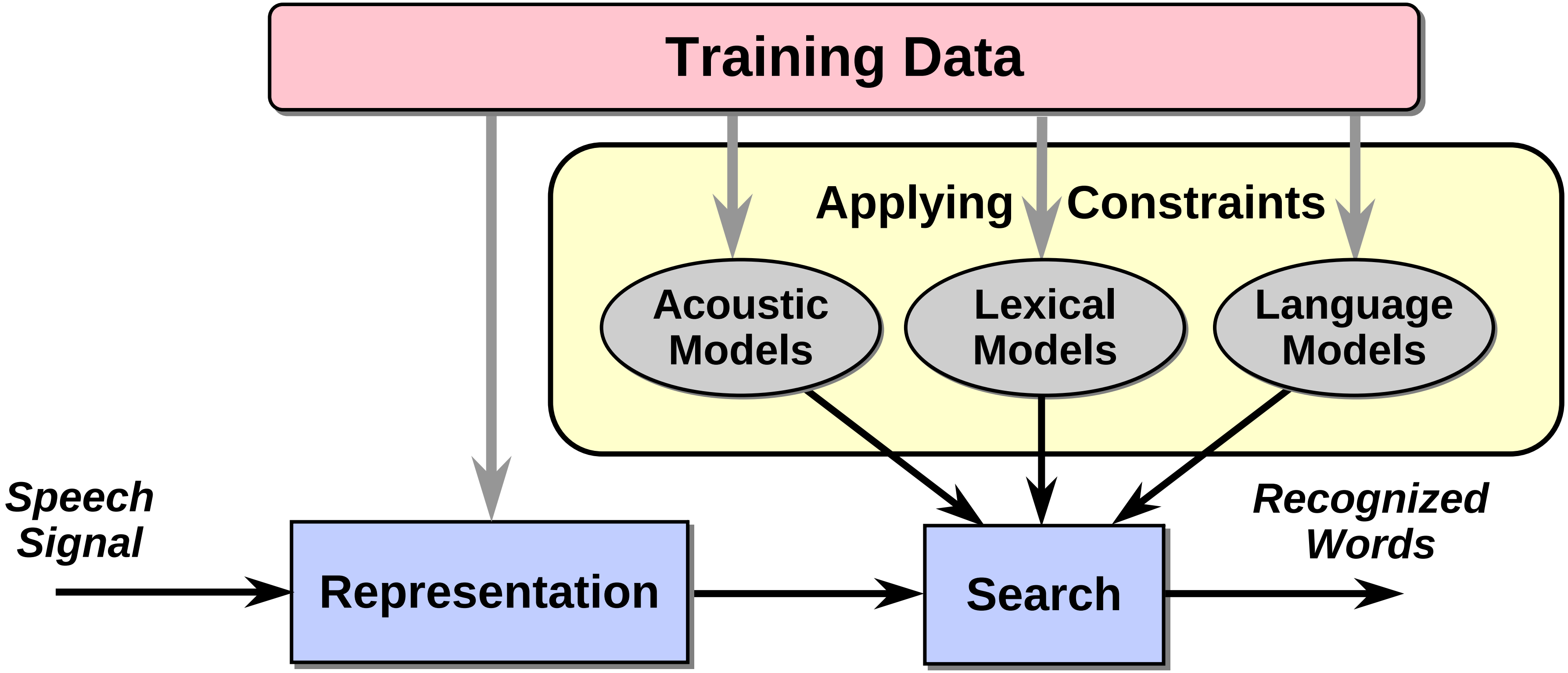
- **Speaker-independent, continuous-speech ASR now possible**
- **Digit recognition over the telephone with word error rate of 0.3%**
- **Error rate cut in half every two years for moderate vocabulary tasks**
- **Error for spontaneous speech more than twice that of read speech**
- **Conversational speech, involving multiple speakers and poor acoustic environment, remains a challenge**
- **Tens of hours of training data to port to a different domain**
- **Statistical modeling using automatic training achieves significant advances**



Important Lessons Learned

- Statistical modeling and data-driven approaches have proved to be powerful
- Research infrastructure is crucial:
  - Large amounts of linguistic data
  - Evaluation methodologies
- Availability and affordability of computing power lead to shorter technology development cycles and real-time systems
- Performance-driven paradigm accelerates technology development
- Interdisciplinary collaboration produces enhanced capabilities (e.g., spoken language understanding)

Major Components in a Speech Recognition System



- **Speech recognition is the problem of deciding on**
  - **How to *represent* the signal**
  - **How to *model* the constraints**
  - **How to *search* for the most optimal answer**

Demo: Continuous Dictation

- **IBM ViaVoice running on a ThinkPad**
- **Trained for a quiet office (classroom performance not optimal)**

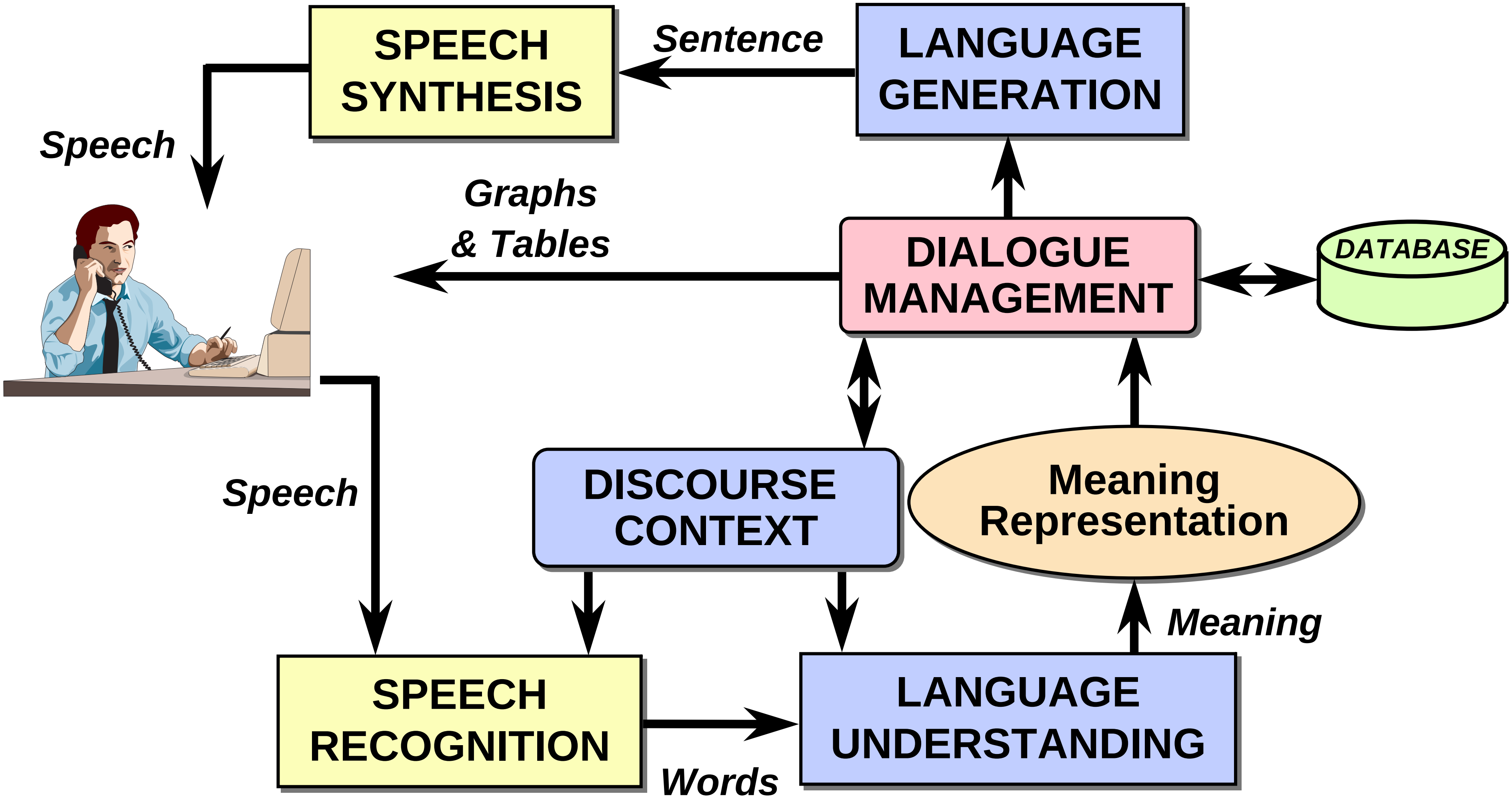
Demo: Simple Telephone Transactions

- **Developed by SpeechWorks International (there are others)**
- **Shipping cost information for Fedex (1-800-GO-FEDEX)**
  - **Provides information on:**
    * Package types
    * Source and destination zip codes
    * Weight, size, value
    * Service type
  - **Handles all US rate information calls**
- **Automated Brokerage System for E*Trade**
  - **Supports quotes and trades**
    * Using symbols or names
    * For stocks, options, and mutual funds
  - **Users can “barge in” at any time**
  - **Nationwide deployment for over 450,000 customers**

Conversational Interfaces: The Next Generation

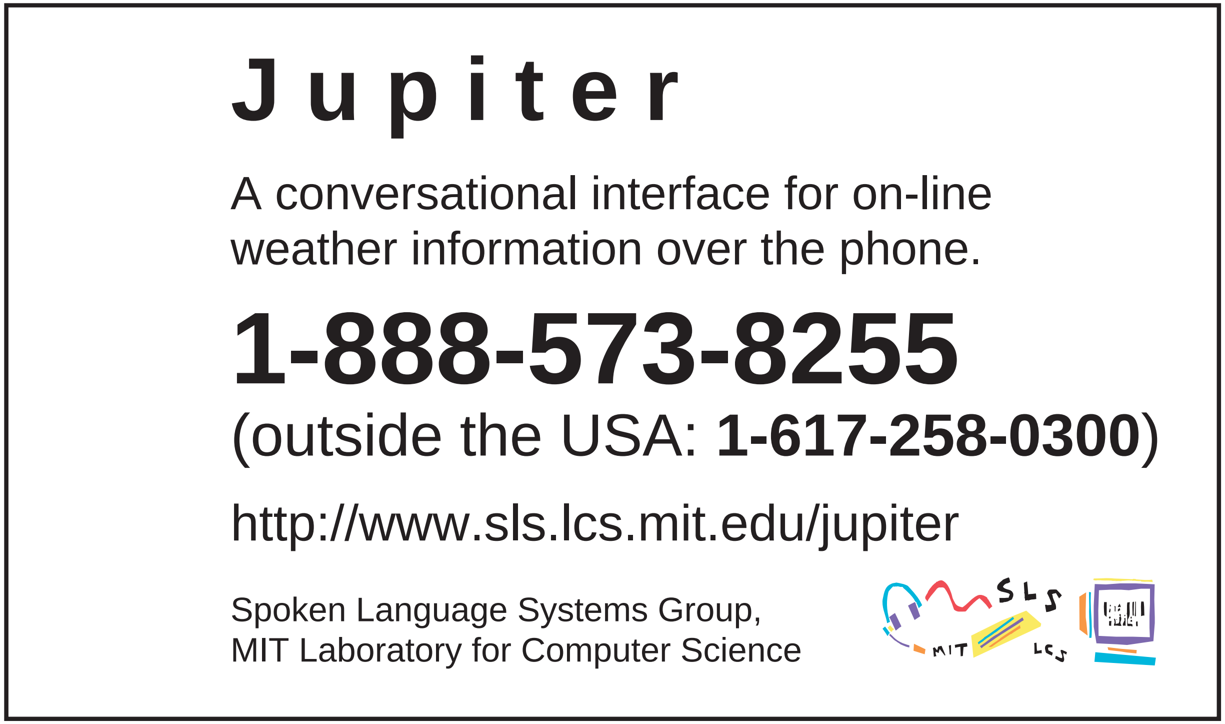
- **Enables us to *converse* with machines (in much the same way we communicate with one another) in order to create, access, and manage information and to solve problems**
- **Augments speech recognition technology with natural language technology in order to *understand* the verbal input**
- **Can engage in a *dialogue* with a user during the interaction**
- **Uses natural language to *speak* the desired response**
- **Is what Hollywood and every “futurist” says we should have!**

A Conversational System Architecture



Demo: Conversational Interface

- **Jupiter weather information system**
  - **Access through telephone**
  - **500 cities worldwide**
  - **Harvest weather information from the Web several times daily**



Jupiter

A conversational interface for on-line
weather information over the phone.

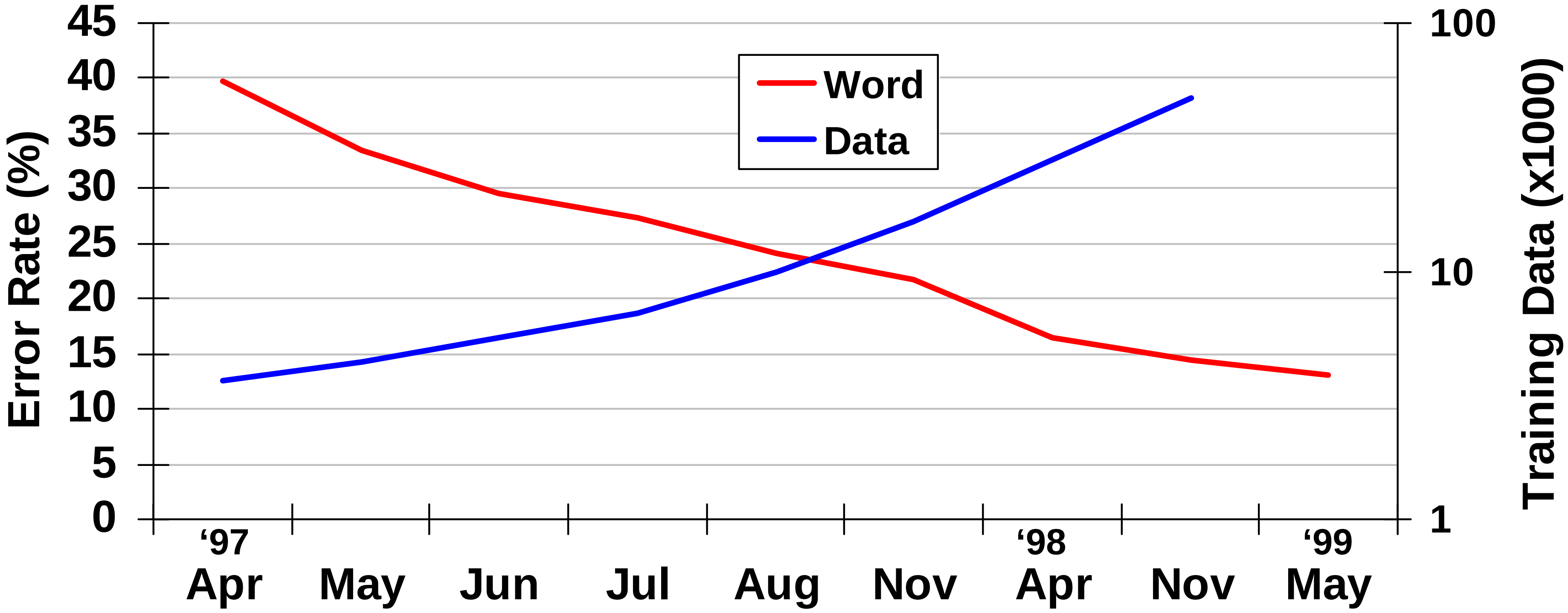
1-888-573-8255

(outside the USA: **1-617-258-0300**)

http://www.sls.lcs.mit.edu/jupiter

Spoken Language Systems Group,
MIT Laboratory for Computer Science

(Real) Data Improves Performance (Weather Domain)

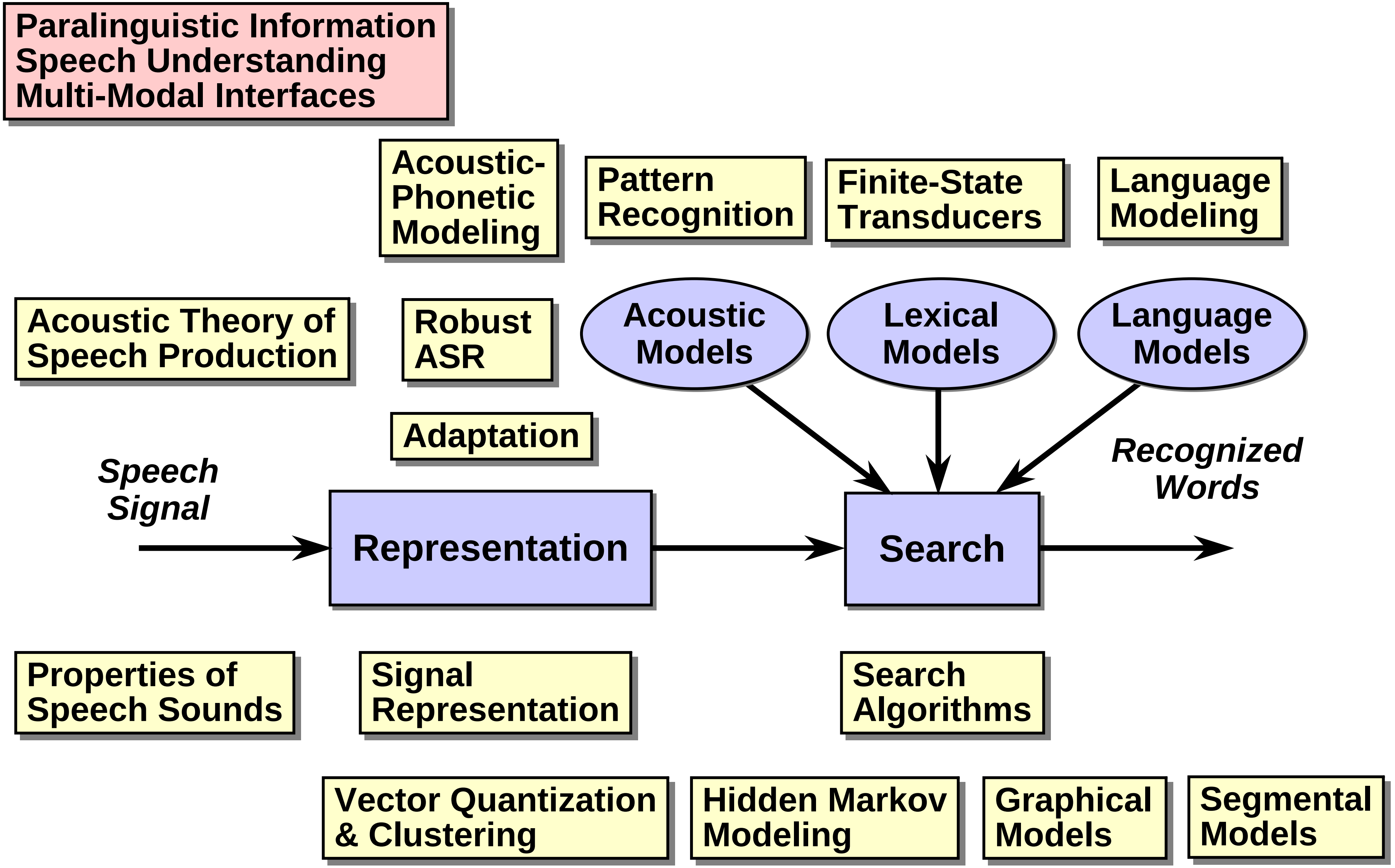


- Longitudinal evaluations show improvements
- Collecting real data improves performance:
  - Enables increased complexity and improved robustness for acoustic and language models
  - Better match than laboratory recording conditions
- Users come in all kinds

But We Are Far from Done!

Corpus	Speech Type	Lexicon Size	Word Error Rate (%)	Human Error Rate (%)
Digit Strings (phone)	spontaneous	10	0.3	0.009
Resource Management	read	1000	3.6	0.1
ATIS	spontaneous	2000	2	--
Wall Street Journal	read	64000	6.6	1
Radio News	mixed	64000	13.5	--
Switchboard (phone)	conversation	10000	19.3	4
Call Home (phone)	conversation	10000	30	--

Course Outline



Course Logistics

- Lectures:Two sessions/week, 1.5 hours/session
- Labs: All week during school hours

Grading

9 Assignments	45%
2 Quizzes	30%
Term Project (about 4 weeks)	25%

Assignments

- There will be 9 weekly assignments
  - Problems that expand on the lecture material
  - Lab assignments to reinforce the lecture material
  - Assignments are due the following week on Wednesday
- Lab work will be done in the computer lab
- Lab sign-up (on the course web page) is necessary
- Solutions will be provided

Term Project

- Investigate a contrasting condition in an ASR experiment
- We will provide different recognizers and domains for you to select from, and will work with you to select a topic
- You choose:
  - Evaluation condition: e.g., phonetic classification, word recognition)
  - Database (e.g., TIMIT, RM, Jupiter, Aurora, …)
  - Recognizer (e.g., Sphinx, Summit, GMTK, …)
  - Contrasting condition (e.g., signal representation, acoustic model, language model)
- Requirements:
  - Proposal
  - Experiments (the bulk of the work)
  - Write-up
  - Presentation on extended last day of class

References (on reserve at Barker)

- **Huang, Acero, & Hon, *Spoken Language Processing*, Prentice-Hall, 2001.**
- **Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, 1997.**
- **Rabiner & Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1983.**
- **Duda, Hart, & Stork, *Pattern Classification*, Wiley & Sons, 2001.**
- **Stevens, *Acoustic Phonetics*, MIT Press, 1998.**