T-61.184
Automatic Speech Recognition: From Theory to Practice

http://www.cis.hut.fi/Opinnot/T-61.184/
September 14, 2004

Prof. Bryan Pellom
Department of Computer Science
Center for Spoken Language Research
University of Colorado

pellom@cslr.colorado.edu

Today's Outline

- **12.15 – 13.00**
  - ❑ Course Outline and Expectations
  - ❑ Historical Perspectives for the Speech Recognition Field

- **13.15 – 14.00**
  - ❑ Talk on "Virtual Humans"
  - ❑ Prof. Ronald Cole,
    Director of the Center for Spoken Language Research (CSLR)
    University of Colorado at Boulder

Who is this person?

- **Visiting Fulbright-Nokia Lecturer at the Helsinki University of Technology**
- **Center for Spoken Language Research, University of Colorado at Boulder (Boulder, Colorado, USA)**
- **My Research Areas**
  - ❑ Speech Recognition & Synthesis
  - ❑ Spoken Dialog Systems
  - ❑ Speech Enhancement
- **My Contact Information**
  - ❑ Office: C314
  - ❑ Email: pellom@james.hut.fi

Course Goals

- **To provide balance between fundamental <u>theory</u> and <u>practice</u> in automatic speech recognition**

- **To provide an enriched experience through hands-on projects and exercises**

- **To introduce several advanced topics:**
  - ❑ Search Architectures
  - ❑ Speaker Adaptation
  - ❑ Environmental Robustness

Course Pre-requisites

- **A Comprehensive Introductory Course**
- **Computer Science, Electrical Engineers and Linguistics students welcome**
- **No Pre-requisites, but the following are useful,**
  - ❑ Linear Algebra (basic Matrix operations)
  - ❑ Probability and Statistics
  - ❑ Signal Processing
  - ❑ Programming
    - ❑ C and or C++
    - ❑ Perl
    - ❑ Unix shell scripting languages
  - ❑ Familiarity with Unix (Linux and/or Sun is Fine)

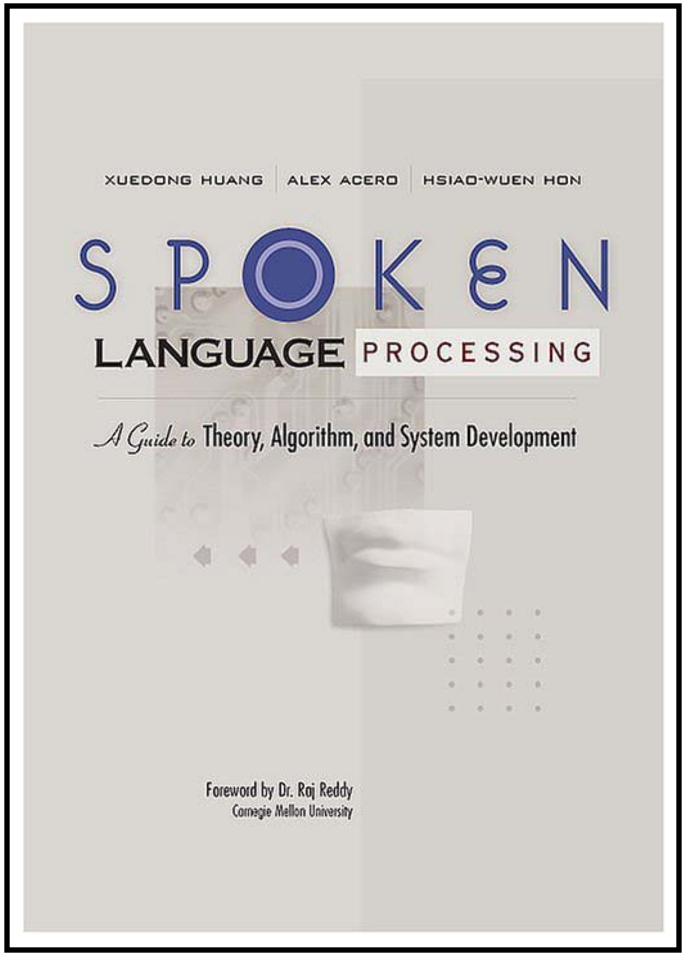
Course Outline

- Speech Recognition Problem Formulation
- Speech Production and Perception, Phonetics and Phonology
- Feature Extraction and Front-end Processing
- Introduction to Hidden Markov Models
- Acoustic & Language Modeling
- Search

Course Outline

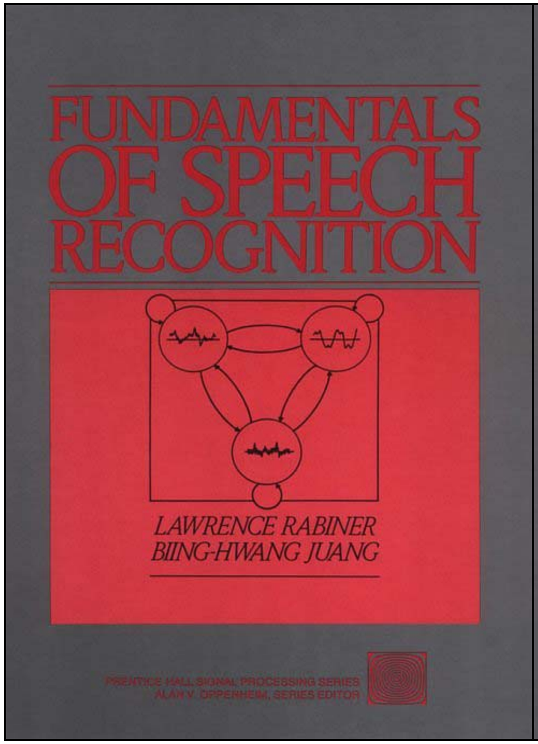
- Practical Issues in System Development and Tuning. Review of “tools of the trade”
- Comparison of existing speech recognition engines and architectures
- Speaker Adaptation
- Environmental Robustness

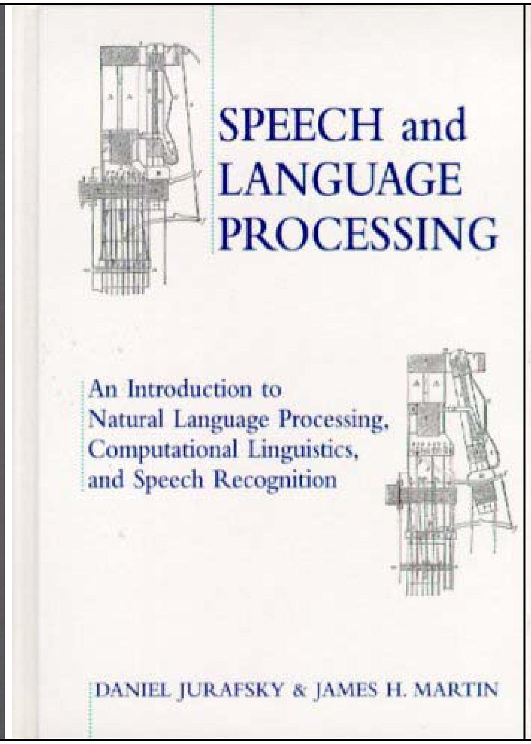
Primary Course Textbook

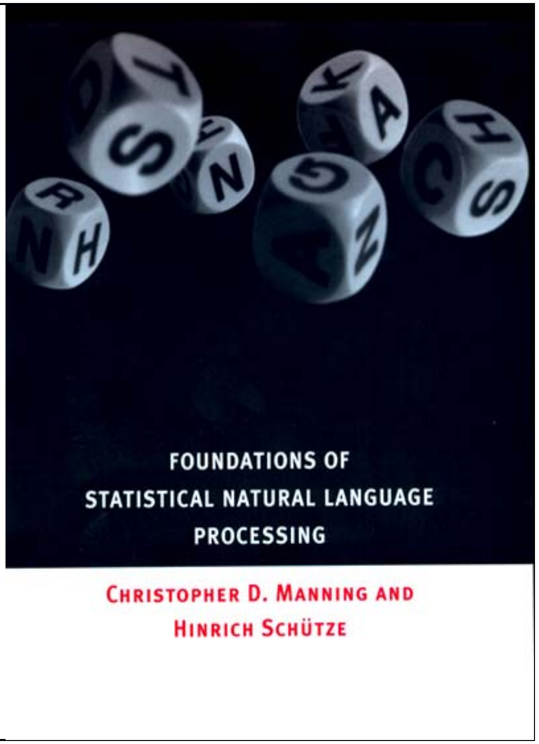

XUEDONG HUANG | ALEX ACERO | HSIAO-WUEN HON
SPOKEN
LANGUAGE PROCESSING
A Guide to Theory, Algorithm, and System Development
Foreword by Dr. Raj Reddy
Carnegie Mellon University

- *Spoken Language Processing*
- Prentice Hall, 2001. ISBN: 0-13-022616-5
- **Resource for Signal Processing, Speech Recognition and Synthesis**
- **Covers from ECE, CS and Linguistics perspective (somewhat ECE biased)**

Other Useful Textbooks


FUNDAMENTALS OF SPEECH RECOGNITION
LAWRENCE RABINER
BIING-HWANG JUANG
PRENTICE HALL SIGNAL PROCESSING SERIES
ALAN V. OPPENHEIM, SERIES EDITOR


SPEECH and LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
DANIEL JURAFSKY & JAMES H. MARTIN


FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING
CHRISTOPHER D. MANNING AND
HINRICH SCHÜTZE

Literature Resources

- **Conference Proceedings**
  - ❑ International Conference on Acoustics, Speech, and Signal Processing (ICASSP)
  - ❑ International Conference on Spoken Language Processing (ICSLP)
  - ❑ Eurospeech
- **Journal Publications**
  - ❑ Speech Communication
  - ❑ IEEE Transactions on Speech and Audio Processing

Software Resources

- **Snack Speech Toolkit**
  - ❑ http://speech.kth.se/snack/
- **OGI Speech Toolkit**
- **University of Colorado SONIC recognizer**
  - ❑ http://cslr.colorado.edu
- **Cambridge Hidden Markov Model Toolkit (HTK)**
- **CMU Sphinx-II Speech Recognizer**
- **NIST Speech Recognition Scoring Utilities**
- **SRI Language Model Toolkit**
- **CMU / Cambridge Language Model Toolkit**

Requirements to Pass the Course

- Attend 70% of the lectures
- Participate during class and read materials posted to course website
- Complete 80% of the exercises
- Complete a project in the area of speech recognition and/or language modeling.

Reading Assignments

- **Articles for assigned reading will be posted to the course website.**
- **Please be sure to check the website for reading materials and lecture slides**
- **Lecture slides will be posted shortly before each class**
- **`http://www.cis.hut.fi/Opinnot/T-61.184/`**

Exercises

- **A minimum of 5 assignments will be provided during the initial weeks of the course**
- **These assignments will require some degree of programming. You can solve the assignments using the language of your choice**
- **1 required assignment near mid-term:**
    - Project Description & Background Literature Survey
    - Must be written in English

Course Project

- A Hands-On project in the area of Computer Speech Recognition and/or Language Modeling
- Topic must be negotiated (you can email me a description of your proposed idea and I will approve)
- Expected: Background literature and manuscript (mid-term) and final project presentation with writeup (final).

Past Course Projects

- Music Recognition and Classification
- SVM Approach to Speaker Verification
- Wireless iPaq Speech Recognition
- Speech Enhancement and Segmentation
- Assessment of Slurred Speech
- Turkish speech recognition
- Speech Recognition in Noisy Environments
- Spanish Accented Speech Recognition
- “Query by Humming”

Two Suggested Course Projects

- **Finnish Adult Speech Recognition (Small Group)**
  - ❑ Collect a Finnish speech database
    - ❑ Minimum of 50 adult speakers (25 male / 25 female)
  - ❑ Read Speech from Newspapers + Continuous Digits
  - ❑ Train and evaluate your recognition system using digits (Univ. of Colorado SONIC recognizer provided)

- **Finnish Child Speech Recognition (Small Group)**
  - ❑ Finnish children's speech database
  - ❑ Collect minimum of 50 children reading stories in Finnish
  - ❑ Ages 7-10 preferred
  - ❑ Train and evaluate your recognition system

Signing Up for the Course

- A sign-up sheet is being passed around
- Please sign up for the course today
- Please clearly write your first and last name and also provide your contact email address

Regular Meeting Time and Location

- **Monday's: 14.00 – 16.00**
  - ❑ 45 minute lecture
  - ❑ 15 minute break
  - ❑ 45 minute lecture
- **Lecture Hall T4**
  Computer Science Building,
  Konemiehentie 2,
  Otaniemi, Espoo

Office Hours

- **Janne Pylkkönen will help assist as needed with the course { `jpylkkon@james.hut.fi` }**

- **My Office Hours**
  - ❑ Monday's 12.15 – 13.15 -or- by appointment
  - ❑ Room C314
  - ❑ pellom@james.hut.fi

Introduction to the Field: History and Challenges

What is Speech Recognition?

- Ultimate Goal: To accurately convert an acoustic signal X into a word sequence W independent of speaker and environment.

- Reality: Several types of recognizers
  - ❑ Isolated Word Recognizers
  - ❑ Word Spotters
  - ❑ Continuous Speech Recognizers

Speech Recognizer Components

- **Acoustic Model**
  - ❑ Knowledge of acoustics, phonetics,
  - ❑ Microphone and Environment
  - ❑ Speaker Differences

- **Lexicon (Pronunciation Dictionary)**
  - ❑ How words are formed from their constituent sounds

- **Language Model**
  - ❑ What constitutes a word,
  - ❑ What words are likely to occur and in what sequence,

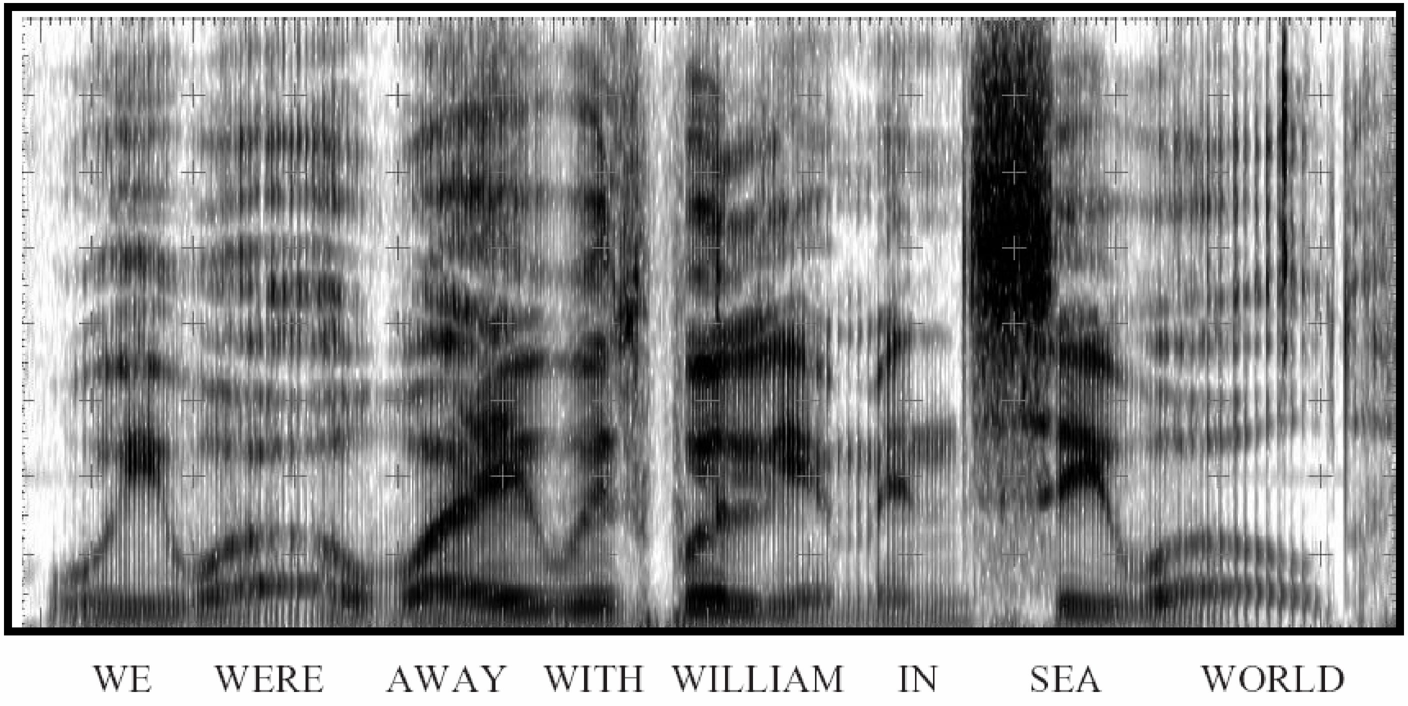
Speech Recognition Modalities

- **Isolated Word Recognition**
  - ❑ Each word is assumed to be surrounded by silence
  - ❑ “this…is…isolated…word…recognition”
- **Connected-Word Recognition**
  - ❑ Word sequences constrained by a fixed grammar (e.g., telephone numbers)
- **Word Spotting**
  - ❑ Detect word in presence of surrounding words
  - ❑ “this is word spotting”
- **Continuous Speech Recognition**
  - ❑ Fluent, uninterrupted speech

Why is Continuous Speech Difficult?

- **Word Boundaries are unclear → complex search**
- **Continuous speech is less clearly articulated**
- **Co-articulation and phonetic context impacts speech both <u>within words</u> and <u>across word boundaries</u>**

Context Variability of / w /



- **Realization of "were" and "world" similar, the rest are different**

Other Issues

- **Speaker-Independent vs. Speaker-Dependent**
- **Inter and Intra Speaker Variability**
  - ❑ Stress,
  - ❑ Emotion,
  - ❑ Speaking Rate
- **Environment Variability**
  - ❑ Stationary vs. Non-Stationary Noise
  - ❑ Microphone vs. Telephone vs. Cell Phone Speech

My Point:

Tough, “Grand Challenge” problem,

Long History (1920’s onward),

Riddled with failure and frustration,

No clear solution so far!

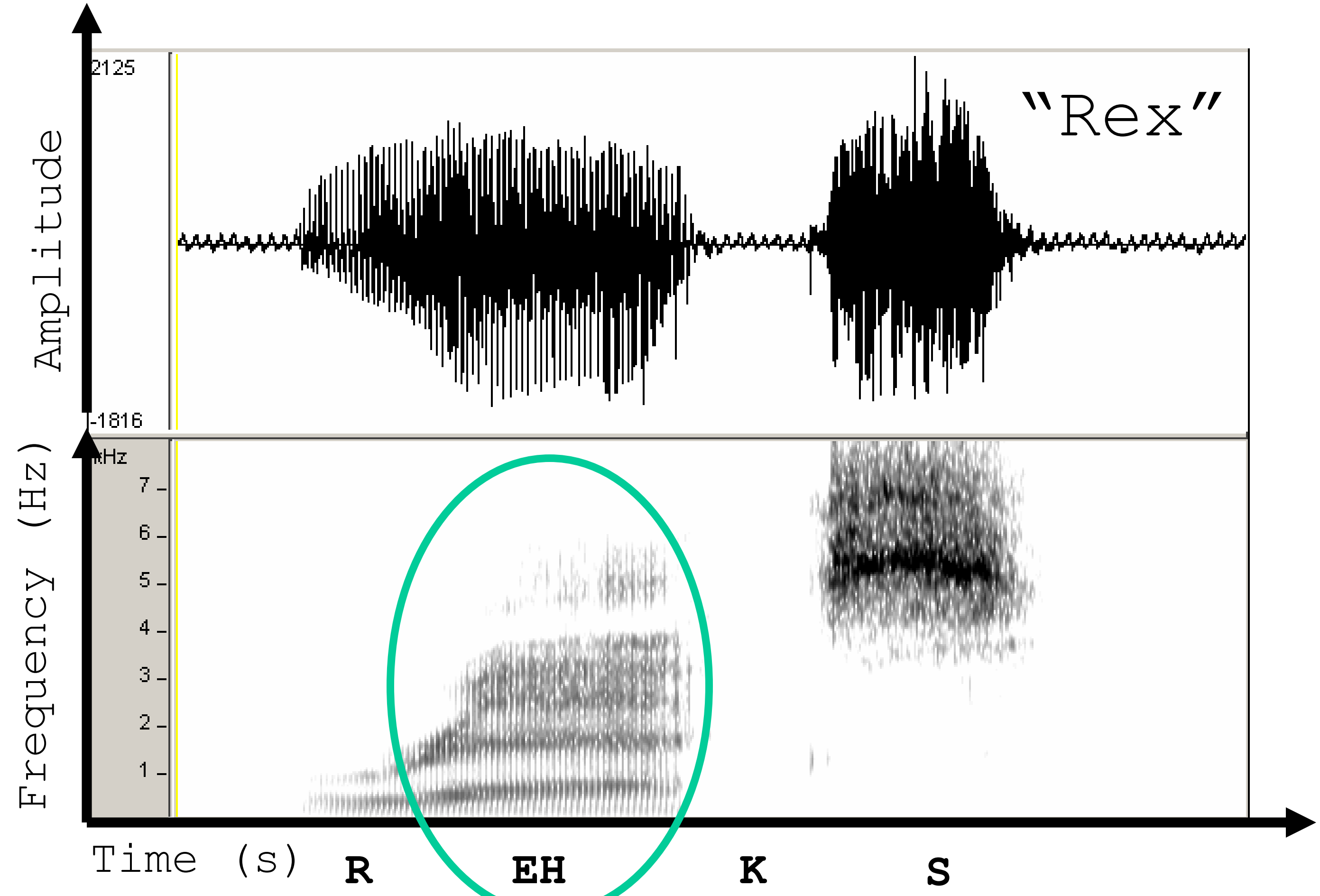
1920’s : Radio Rex

- Celluloid toy dog
- Developed by Walker Balke
- National Company, Inc.
- Attached to the turntable of a phonograph
- Controlled by resonant reed
- Would jump out of it’s kennel (dog house) when certain note played on record

Description of Radio Rex

"It consisted of a celluloid dog with an iron base held within its house by an electromagnet against the force of a spring. Current energizing the magnet flowed through a metal bar which was arranged to form a bridge with 2 supporting members. This bridge was sensitive to 500 cps acoustic energy which vibrated it, interrupting the current and releasing the dog. The energy around 500 cps contained in the vowel of the word Rex was sufficient to trigger the device when the dog's name was called."

(1947) Spectrogram



T-61.184

1952 Bell Labs Digit Recognizer

- Davis, Biddulph, Balashek (1952), "Automatic Recognition of Spoken Digits," The Journal of the Acoustical Society of America, 24(6), 637-642.
- Digits 0→9 as spoken over the telephone
- Based on analysis of the spectrum divided into 2 frequency bands (above and below 900 Hz).
- Identified vowel sounds with 93% accuracy
- < 2% digit error <u>if the user didn't move his head</u>!

1952 Bell Labs Digit Recognizer

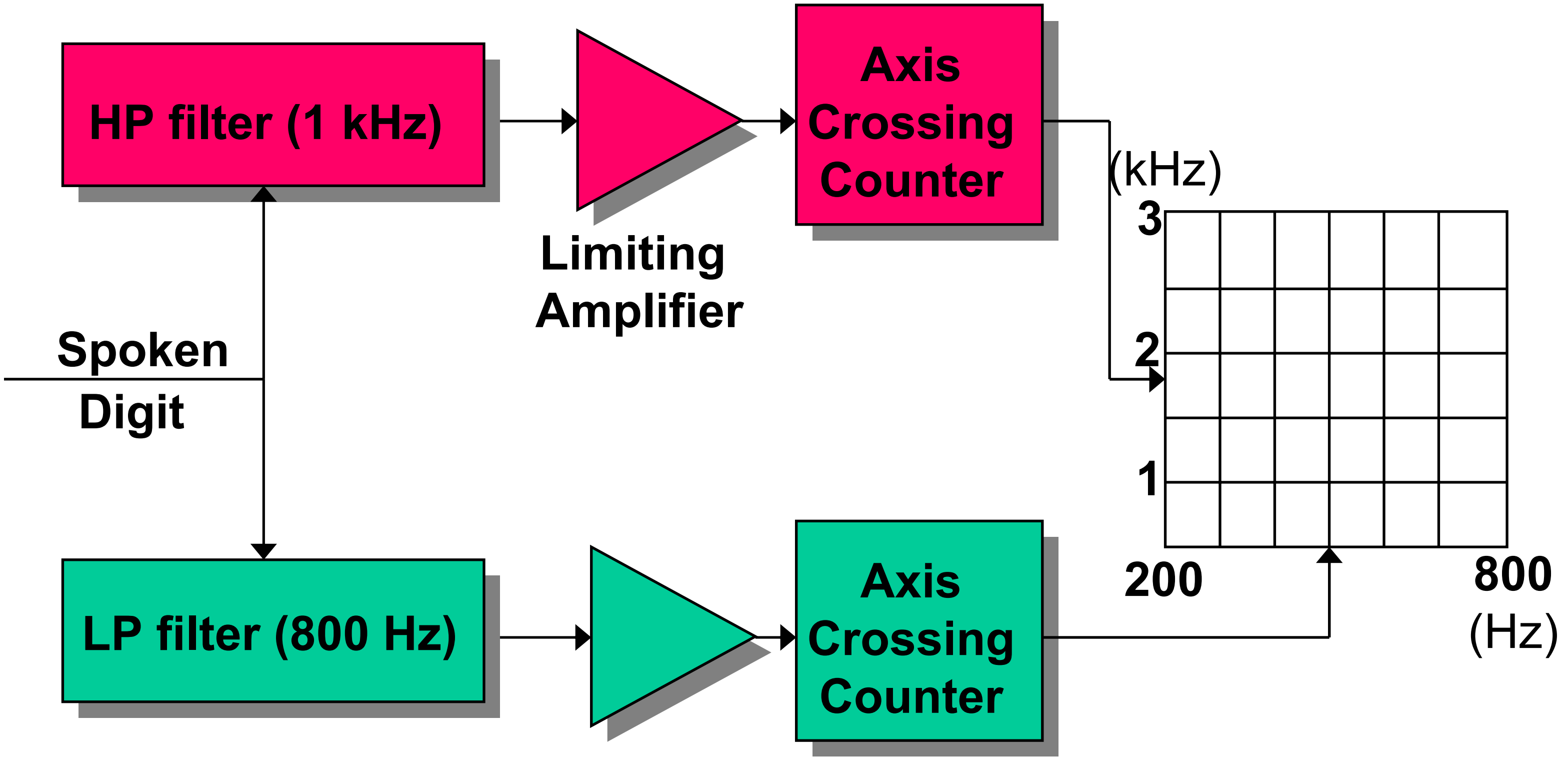


Illustration from ICSCI-Berkeley Course Notes

1960’s – 1970’s

- Fast Fourier Transform (FFT)
- Hidden Markov Model Theory (1966-1972)
- Dynamic Time Warping (1970)
- ARPA Speech Understanding Project (1971-76)
  - ❑ The shift from isolated to connected word recognition
  - ❑ Modest vocabulary size
  - ❑ Ambitious, well-funded project

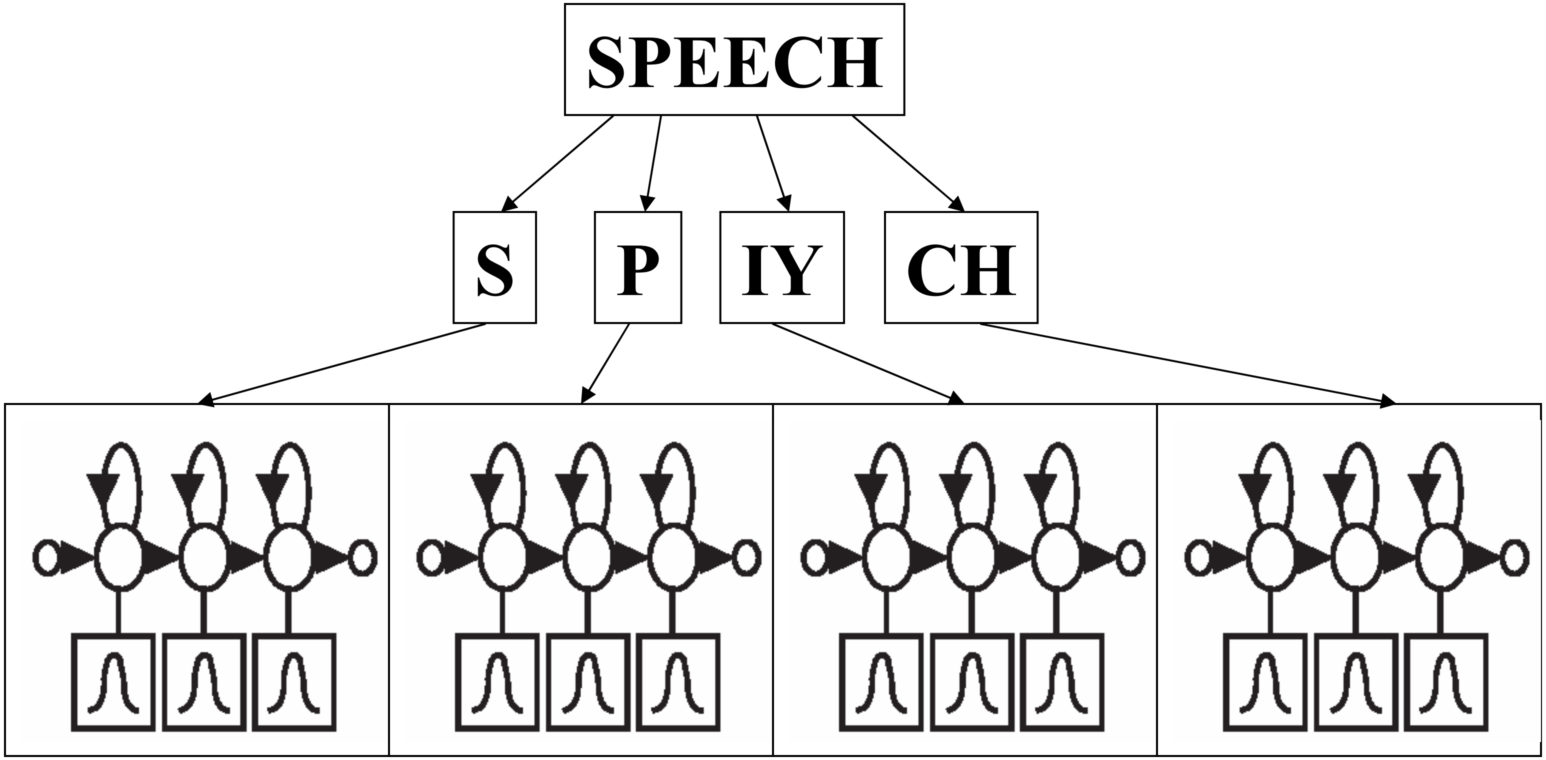
ARPA Speech Understanding Project

- **1971 – 1976 ($15M Funding)**
- **Goals: 1000-word vocabulary, connected speech, constrained grammar**
- **CMU, Systems Development Corporation, BBN**
- **Outcomes:**
  - ❑ CMU Harpy System: Best system, met program goals
  - ❑ 5% sentence error
  - ❑ Restricted word-order for improved recognition, centisecond processing, phoneme-based templates, use of linear prediction based features

1970's –

- **Viterbi Algorithm (1973)**
- **Hidden Markov Models for Speech Recognition (1974-)**
  - ❑ Jim Baker in 1974
  - ❑ IBM (Baker, Jalinek, Bahl, Mercer) 1976 –
- **Linear Predictive Coding (1975)**

HMM Representation of Words



1980's –

- **Front-End Features Developed and Standardized**
  - ❑ Mel-Cepstrum (Davis & Mermelstein, 1980)
  - ❑ PLP (Hermansky)
  - ❑ Delta-Cepstrum (Furui)
- **Larger Speech Corpora shared and evaluated (1984 >)**
  - ❑ TIMIT, Resource Management, ATIS
- **2nd DARPA Project (1988)**
  - ❑ Common Evaluation Paradigm and Metrics
  - ❑ Techniques began to converge
  - ❑ NIST and LDC Involvement
- **Shift from template based to statistical approaches**
- **Hidden Markov Model (HMM) becomes mainstream**

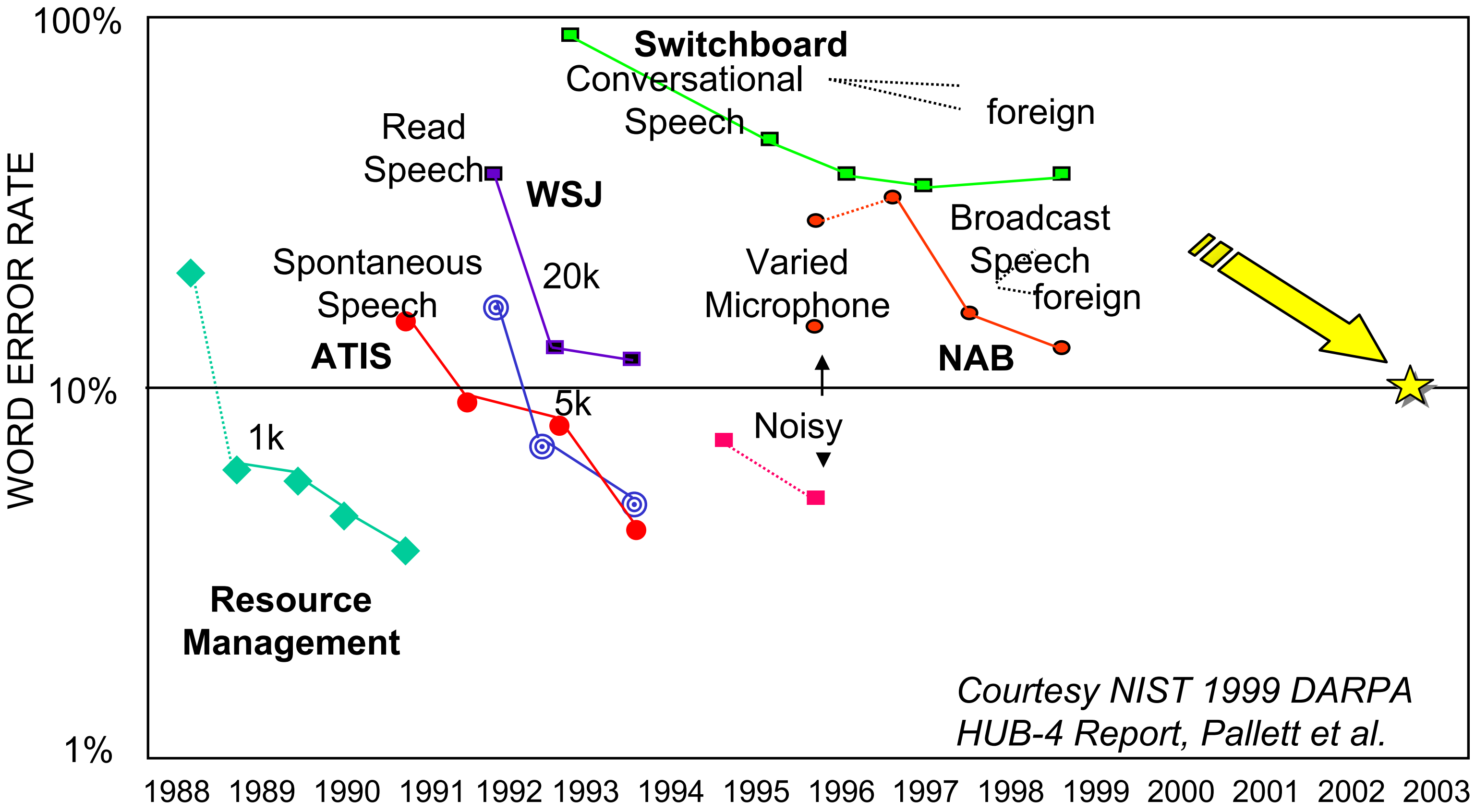
1990's –

- **Refinements to HMM Approaches**
- **Noise Robustness**
- **HMM Training & Modeling Methods**
  - ❑ Decision Tree State Clustering
- **Speaker Adaptation**
  - ❑ Maximum Likelihood Linear Regression (MLLR)
  - ❑ Vocal Tract Length Normalization (VTLN)
- **Larger tasks, higher complexity, more training data**

Increasing Speech Task Complexities

1988 – 1991:	**Resource Management (RM)** 1k word vocabulary
1991 – 1994:	**Air Travel Information System (ATIS)** 5k word vocabulary
1992 – 1994:	**Wall Street Journal (WSJ)** 5k – 20k word vocabulary
1996 –	**Broadcast News & Switchboard** 45k - 64k word vocabulary **Meeting Transcription**

DARPA Benchmark Evaluations



Courtesy NIST 1999 DARPA HUB-4 Report, Pallett et al.

T-61.184

DARPA EARS Program (2001-)

- ***Effective, Affordable, Reusable Speech-to-Text***
- **Speech-to-Text (English, Arabic, Mandarin)**
  - □ Broadcast News System: XX% WER
  - □ Conversational Telephone System: XX% WER
- **Rich Transcription**
  - □ Detection of Sentence Boundaries
  - □ Detection of Disfluencies
  - □ Speaker ID and Speaker Change Detection
- **10x and 1x real-time evaluation systems**

Human vs. Machine Performance

Tasks	Vocabulary Size	Humans (word error %)	Machines (word error %)
Connected Digits	10	0.009%	0.4-0.7%
Alphabet Letters	26	1%	3.0-5.0%
Spontaneous Telephone Speech	40,000	3.8%	20-25%
WSJ* with clean speech	5,000	0.9%	3.0-4.0%
WSJ with noisy speech	5,000	1.1%	7.0-10%

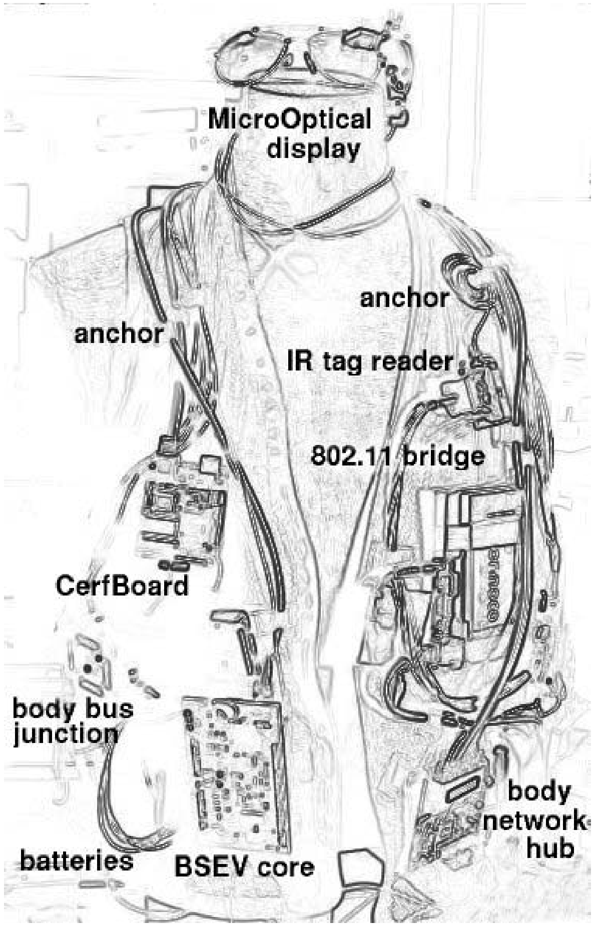
*WSJ = Wall Street Journal, read newspaper text

Current Research Trends

- **Discriminative System Training Methods**
  - ❑ Acoustic Models
  - ❑ Language Models
- **Recognizer Hypothesis Combination**
- **Bottom-Up “Event Detection” Based ASR**
  - ❑ Shift the paradigm a little (back to “science”)
  - ❑ Speech event lattice (“voicing”, “nasality”, “frication”, etc.)
  - ❑ Integrate more knowledge sources into the solution

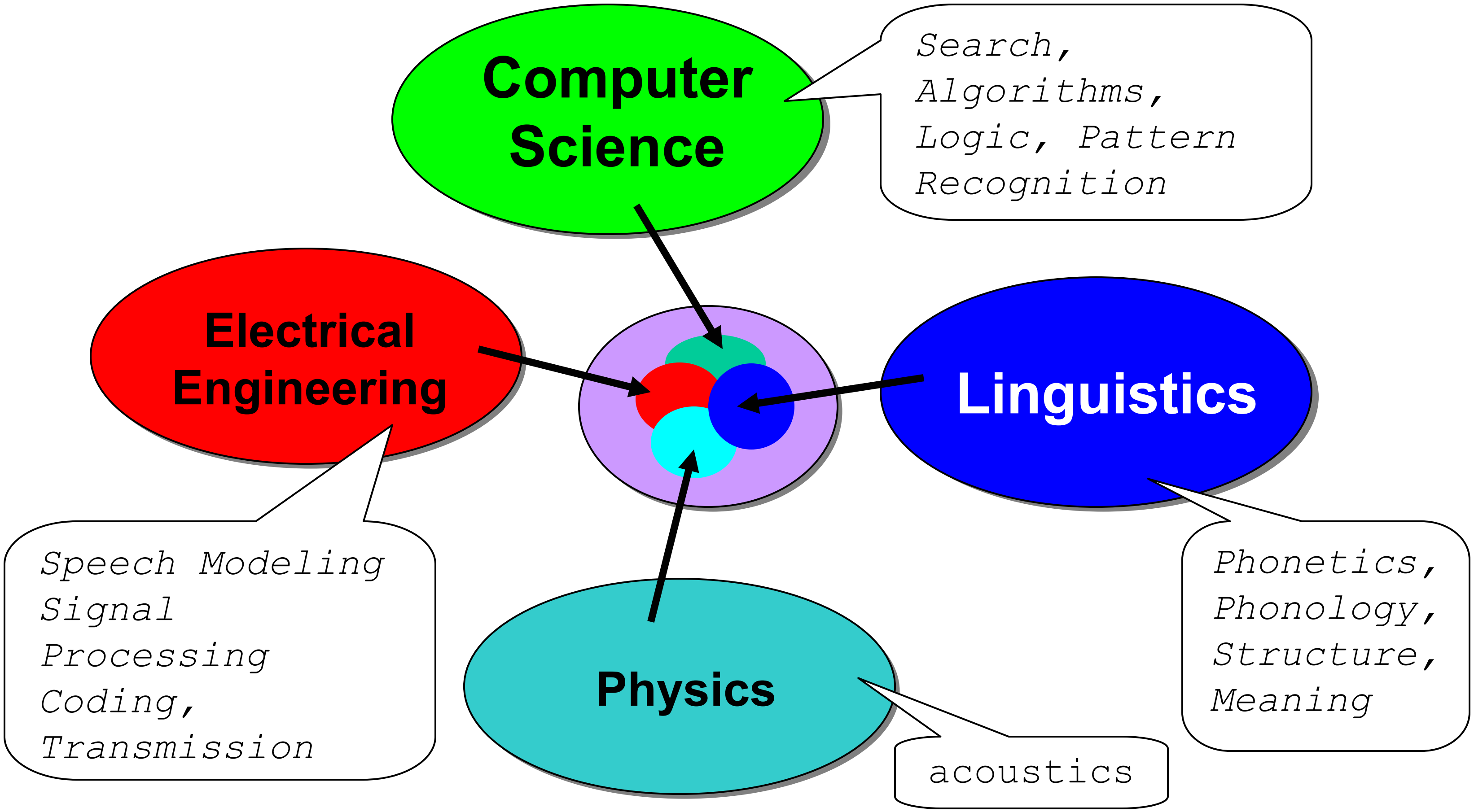
The Future

- **Ubiquitous Computing**
  - ❑ Wearable Computers
  - ❑ ASR & Speech Understanding anytime-anywhere
  - ❑ “Mobile” and “On-the-Move”
  - ❑ Wireless
- **Rugged Environment ASR**
- **Multiple Languages**
- **Speech-to-Speech Translation**
- **Animated Agents**



T-61.184

Multidisciplinary Contributions



Computer Science

Search, Algorithms, Logic, Pattern Recognition

Electrical Engineering

Speech Modeling Signal Processing Coding, Transmission

Linguistics

Phonetics, Phonology, Structure, Meaning

Physics

acoustics

T-61.184

Reminder: Next Meeting

- **Monday, September 20th, 14.00 – 16.00,**
- **Lecture Hall T4**

- **Topics**
  - ❑ The Speech Recognition Problem Formulation
  - ❑ Speech Production, Perception, and Phonetics
  - ❑ Exercise #1 will be assigned