Challenges in Speech Recognition

Lawrence Rabiner

Rutgers University and the University of California at Santa Barbara

Automatic Speech Recognition

- ***Goal***: *Accurately* and *efficiently* convert a speech signal into a text message independent of the device, speaker or the environment.

- ***Applications***: Automation of complex operator-based tasks, e.g., customer care, dictation, form filling applications, provisioning of new services, customer help lines, e-commerce, etc.

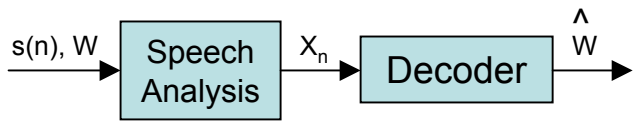
Basic ASR Formulation

The basic equation of Bayes rule-based speech recognition is

$$\hat{W} = \arg\max_{W} P(\mathbf{W} \mid \mathbf{X})$$

$$= \arg\max_{W} \frac{P(\mathbf{W})P(\mathbf{X} \mid \mathbf{W})}{P(\mathbf{X})}$$

$$= \arg\max_{W} P(\mathbf{W})P(\mathbf{X} \mid \mathbf{W})$$

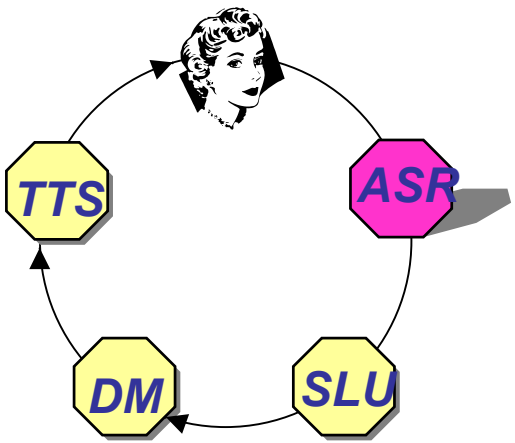


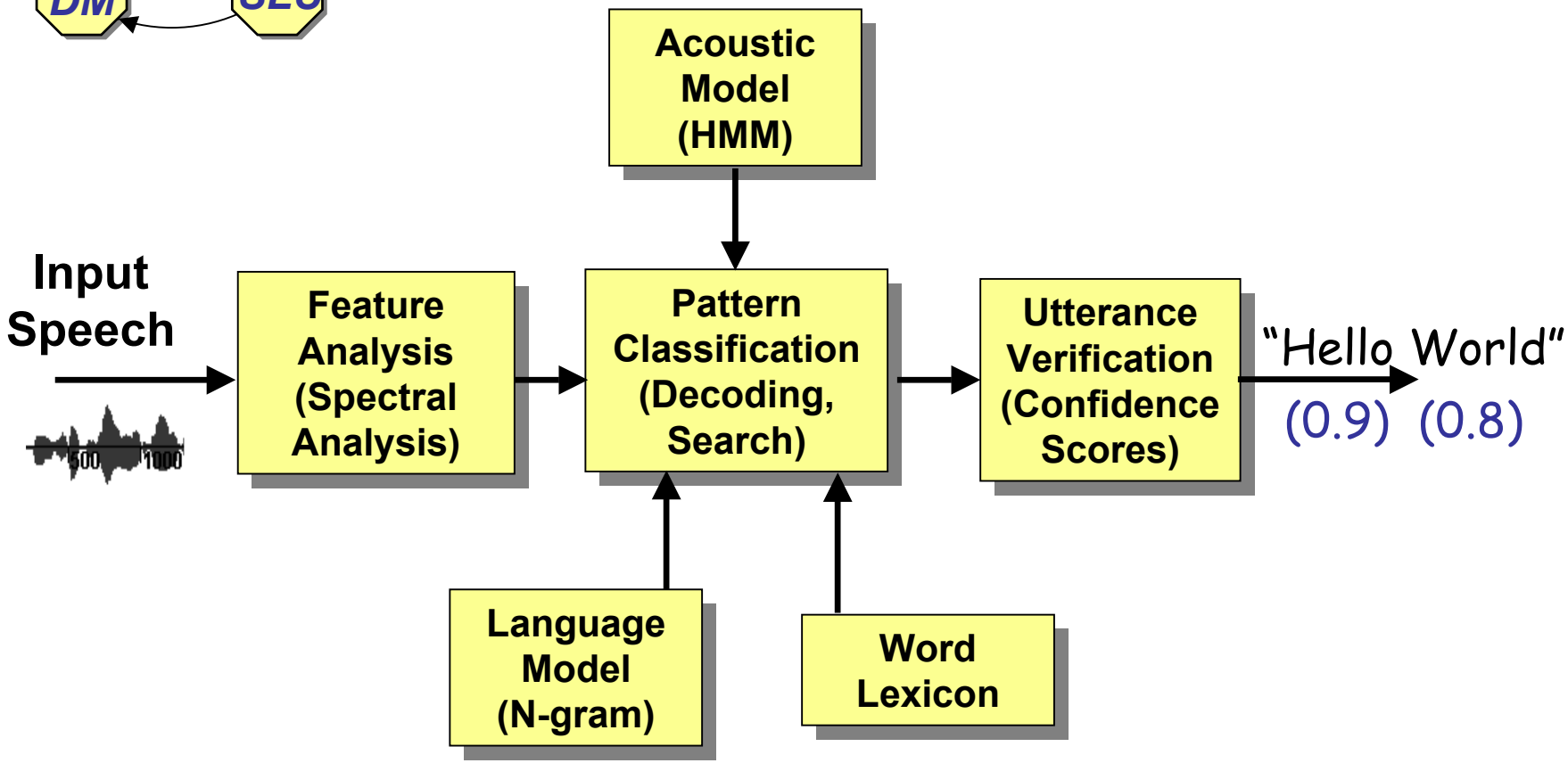
where $\mathbf{X}=X_1,X_2,\ldots,X_N$ is the acoustic observation (feature vector) sequence.

$$\hat{\mathbf{W}} = w_1 w_2 \ldots w_M$$

is the corresponding word sequence, $P(\mathbf{X}|\mathbf{W})$ is the acoustic model and $P(\mathbf{W})$ is the language model

Speech Recognition Process





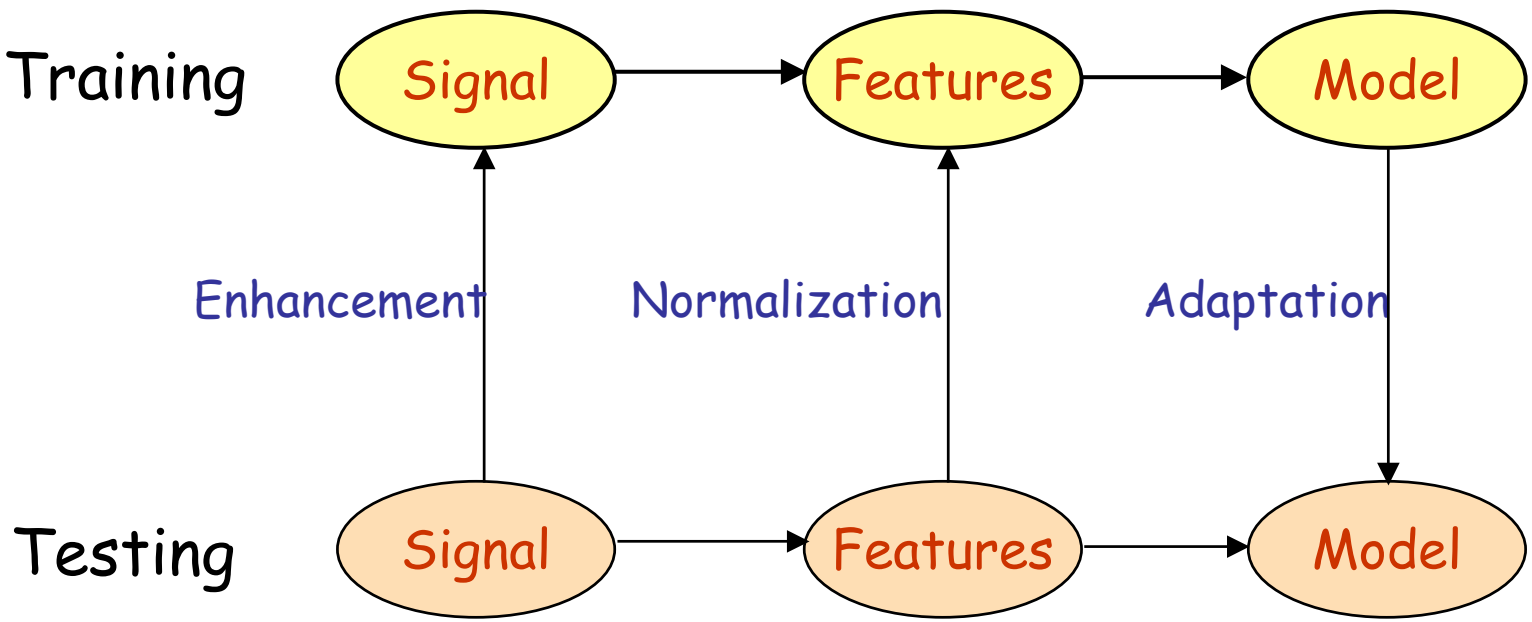
Speech Recognition Processes

- Choose task => sounds, word vocabulary, task syntax (grammar), task semantics
  - Text training data set => word lexicon, word grammar (language model), task grammar
  - Speech training data set => acoustic models
- Training algorithm => build models from training set of text and speech
- Evaluate performance—testing algorithm
  - Speech testing data set

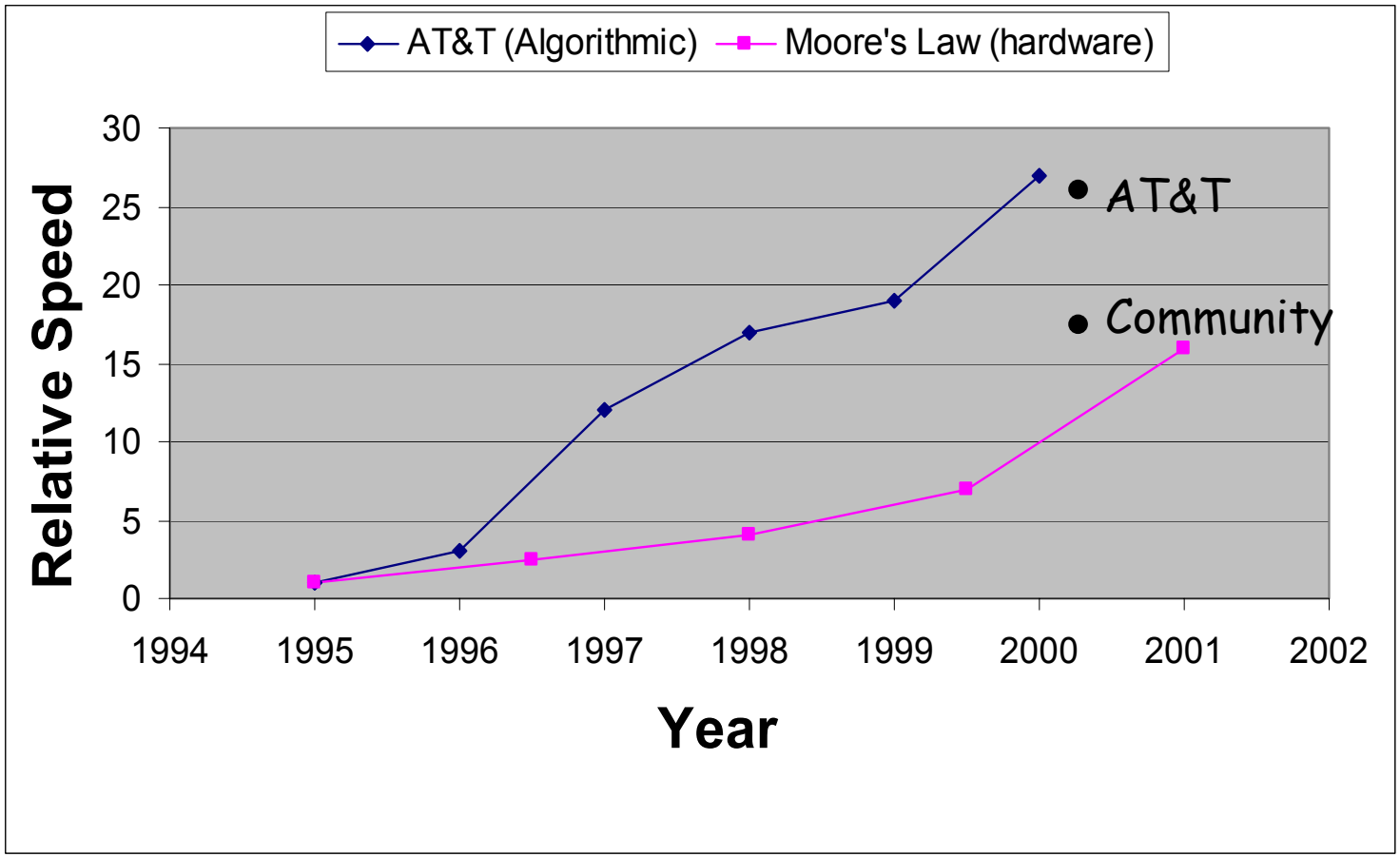
Challenges for ASR and NLU

1. Robustness
2. Robustness
3. Robustness
4. Is HMM the end of the line
5. Automatic generation of word lexicons
6. Automatic generation of language models for new tasks
7. Finding the theoretical limit for FSM implementations of ASR/NLU systems
8. Optimal utterance verification-rejection algorithms
9. Achieving or surpassing human performance on ASR tasks, NLU tasks

Methods for Robust Speech Recognition



Algorithmic Speed-up for Speech Recognition



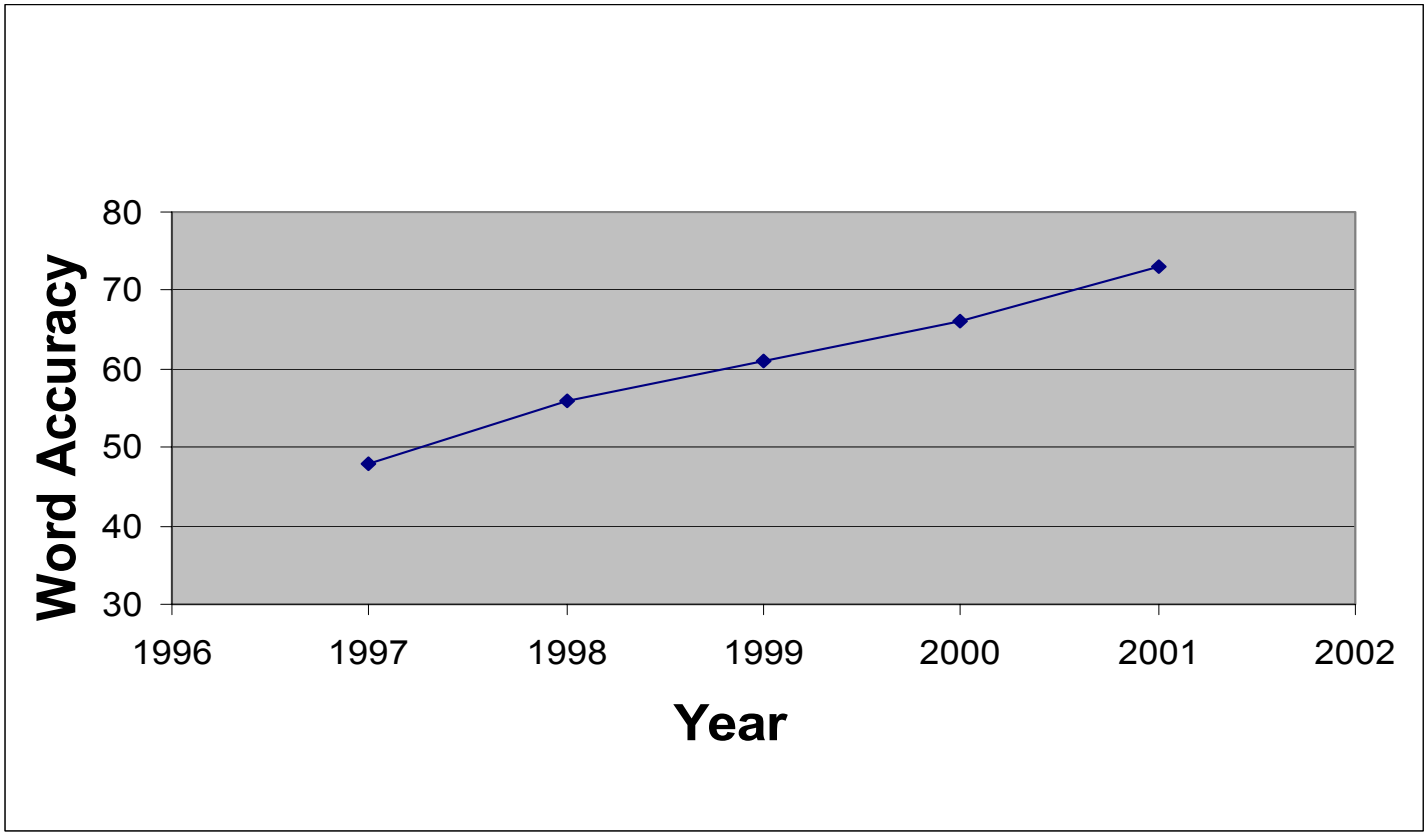
North American Business
vocabulary: 40,000 words
branching factor: 85

Word Error Rates

CORPUS	TYPE	VOCABULARY SIZE	WORD ERROR RATE
Connected Digit Strings--TI Database	Spontaneous	11 (zero-nine, oh)	0.3%
Connected Digit Strings--Mall Recordings	Spontaneous	11 (zero-nine, oh)	2.0%
Connected Digits Strings--HMIHY	Conversational	11 (zero-nine, oh)	5.0%
RM (Resource Management)	Read Speech	1000	2.0%
ATIS(Airline Travel Information System)	Spontaneous	2500	2.5%
NAB (North American Business)	Read Text	64,000	6.6%
Broadcast News	News Show	210,000	13-17%
Switchboard	Conversational Telephone	45,000	25-29%
Call Home	Conversational Telephone	28,000	40%

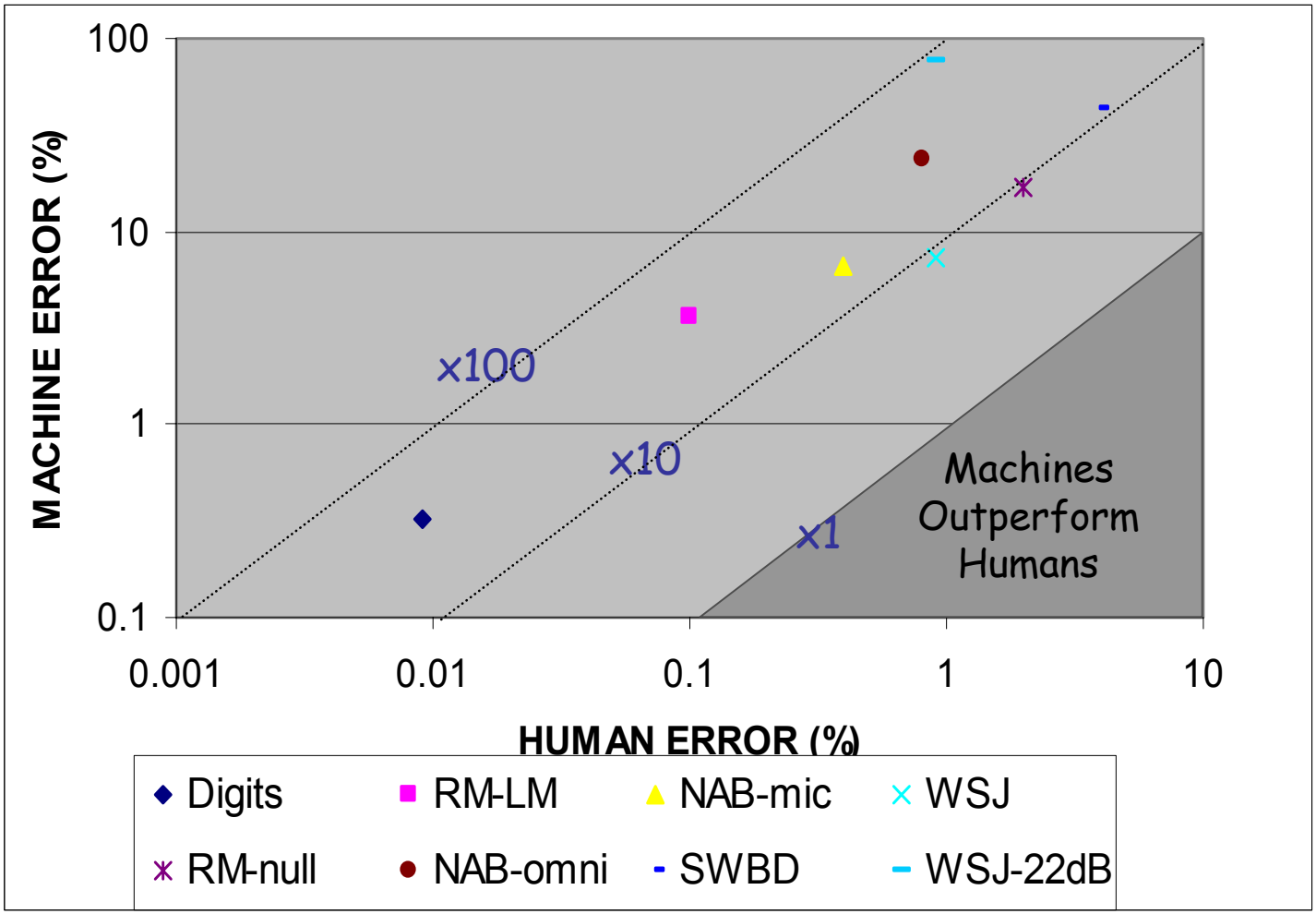
Factor of 17 increase in digit error rate

Algorithmic Accuracy for Speech Recognition



Switchboard/Call Home
Vocabulary: 40,000 words
Perplexity: 85

Human Speech Recognition vs ASR



Milestones in Speech and Multimodal Technology Research

Small Vocabulary, Acoustic Phonetics-based → **Medium Vocabulary, Template-based** → **Large Vocabulary, Statistical-based** → **Large Vocabulary; Syntax, Semantics,** → **Very Large Vocabulary; Semantics, Multimodal Dialog, TTS**

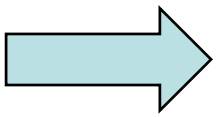
Isolated Words	Isolated Words; Connected Digits; Continuous Speech	Connected Words; Continuous Speech	Continuous Speech; Speech Understanding	Spoken dialog; Multiple modalities
Filter-bank analysis; Time-normalization; Dynamic programming	Pattern recognition; LPC analysis; Clustering algorithms; Level building;	Hidden Markov models; Stochastic Language modeling;	Stochastic language understanding; Finite-state machines; Statistical learning;	Concatenative synthesis; Machine learning; Mixed-initiative dialog;

1962 1967 1972 1977 1982 1987 1992 1997 2002

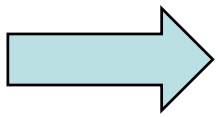
Year

Future of Speech Recognition Technologies

Very Large Vocabulary, Limited Tasks, *Controlled* Environment



Very Large Vocabulary, Limited Tasks, *Arbitrary* Environment



Unlimited Vocabulary, *Unlimited* Tasks, Many Languages

Dialog Systems

Robust Systems

Multilingual Systems; Multimodal Speech Enabled Devices

2002 2005 2008 2011

Year