SPEECH RECOGNITION TECHNOLOGY: COMMODITY OR LIABILITY

- **Joseph Picone**

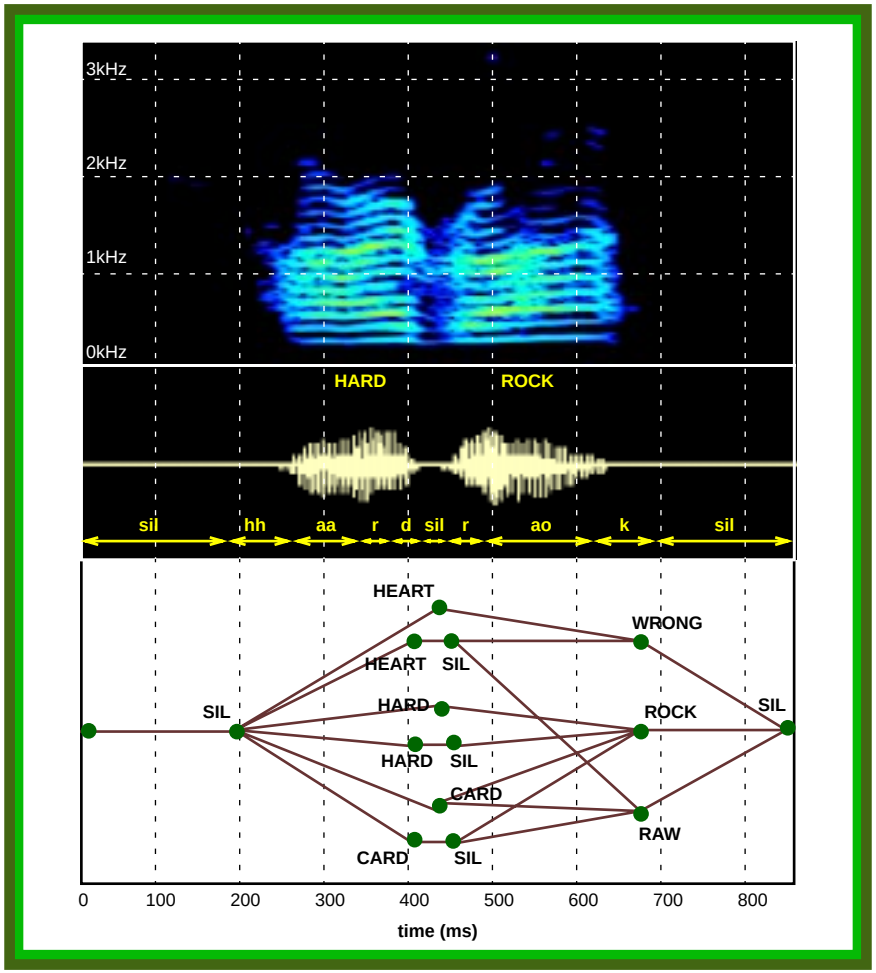
  Director, Institute for Signal and Information Processing
  Professor, Dept. Electrical and Computer Engineering
  Mississippi State University

- **Contact Information:**

  Box 9571
  Mississippi State University
  Mississippi State. Mississippi 39762
  Tel: 662-325-3149, Fax: 662-325-2298
  Email: picone@isip.msstate.edu

- **Visit our speech recognition web site at:**

  http://www.isip.msstate.edu/projects/speech



- **This talk is available at:**

  http://www.isip.msstate.edu/publications/seminars/external/2002/ivoice

INTRODUCTION

ABSTRACT

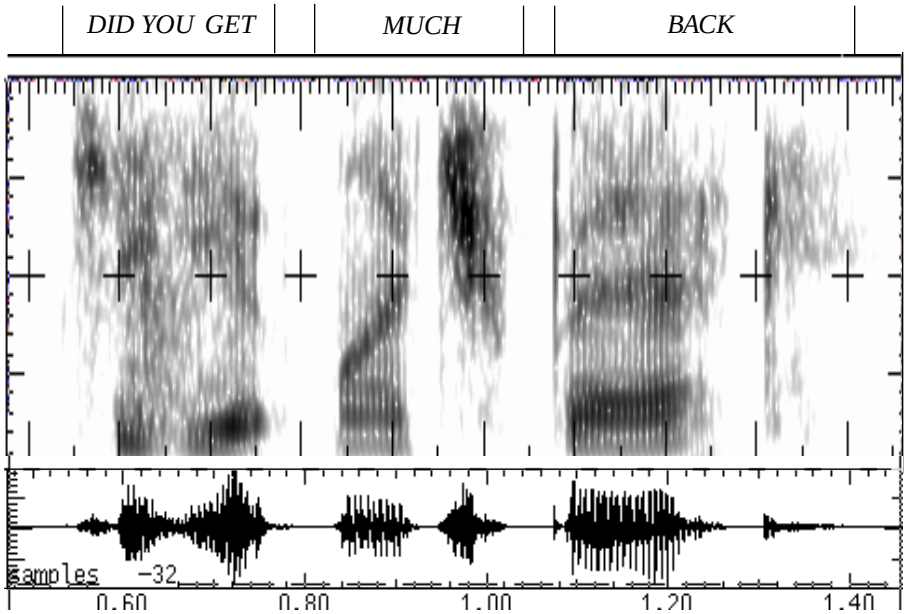
Language is a uniquely human tool by which people exchange ideas. Automatic speech recognition (ASR) is the conversion of a sound pressure wave representing these ideas to text. This signal is at best a noisy representation. Accurate conversion of this signal requires building machines that approach human intelligence. Modern ASR systems rely heavily on statistical methods and powerful computers to achieve this goal. In this talk, we will review the dominant approaches for achieving high performance speech recognition. On limited tasks, machines are approaching human performance. However, to provide flexible and intuitive voice interfaces, we must develop a more fundamental computational paradigm for representing language.

About the speaker:

Joseph Picone is currently a Professor and Hearin Eminent Scholar in the Department of Electrical and Computer Engineering at Mississippi State University, where he also directs the Institute for Signal and Information Processing.

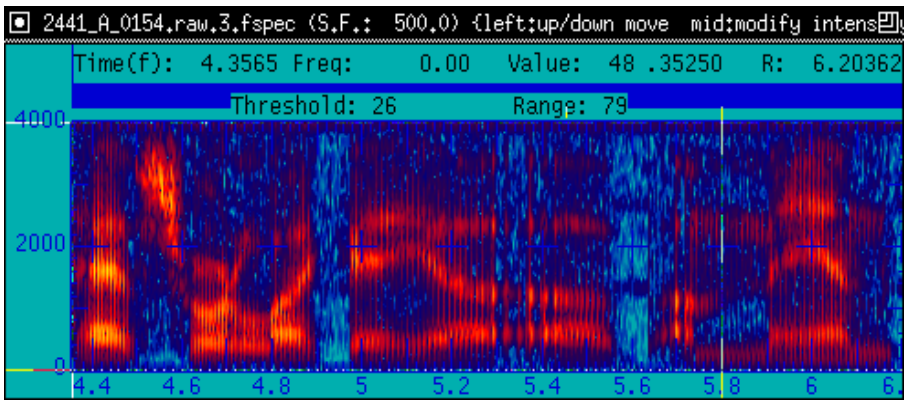
INTRODUCTION

FUNDAMENTAL CHALLENGE



- "Did you get" is reduced such that the resulting word is pronounced "jyuge."
- Phoneme deletion rate: ~12%
  Syllable deletion rate: ~1%
- Predicting pronunciations of words is crucial!

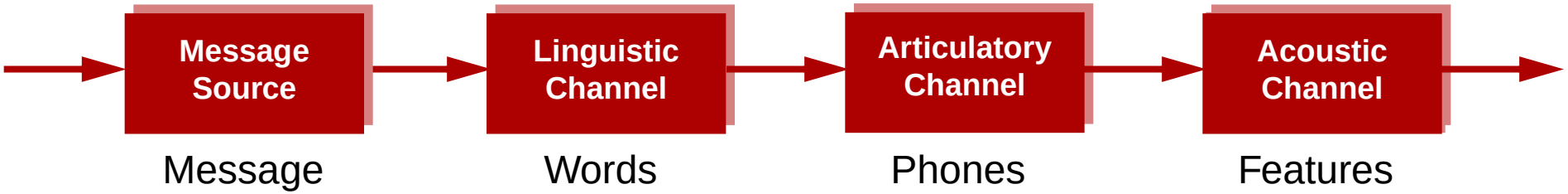
"have sort of like a a a manpower"



- Conversational speech defies conventional grammatical structure.
- Constrained interfaces have failed!

INTRODUCTION

NOISY COMMUNICATION CHANNEL MODEL



Bayesian formulation for speech recognition:

$$P(W|A) = P(A|W)P(W)/P(A)$$

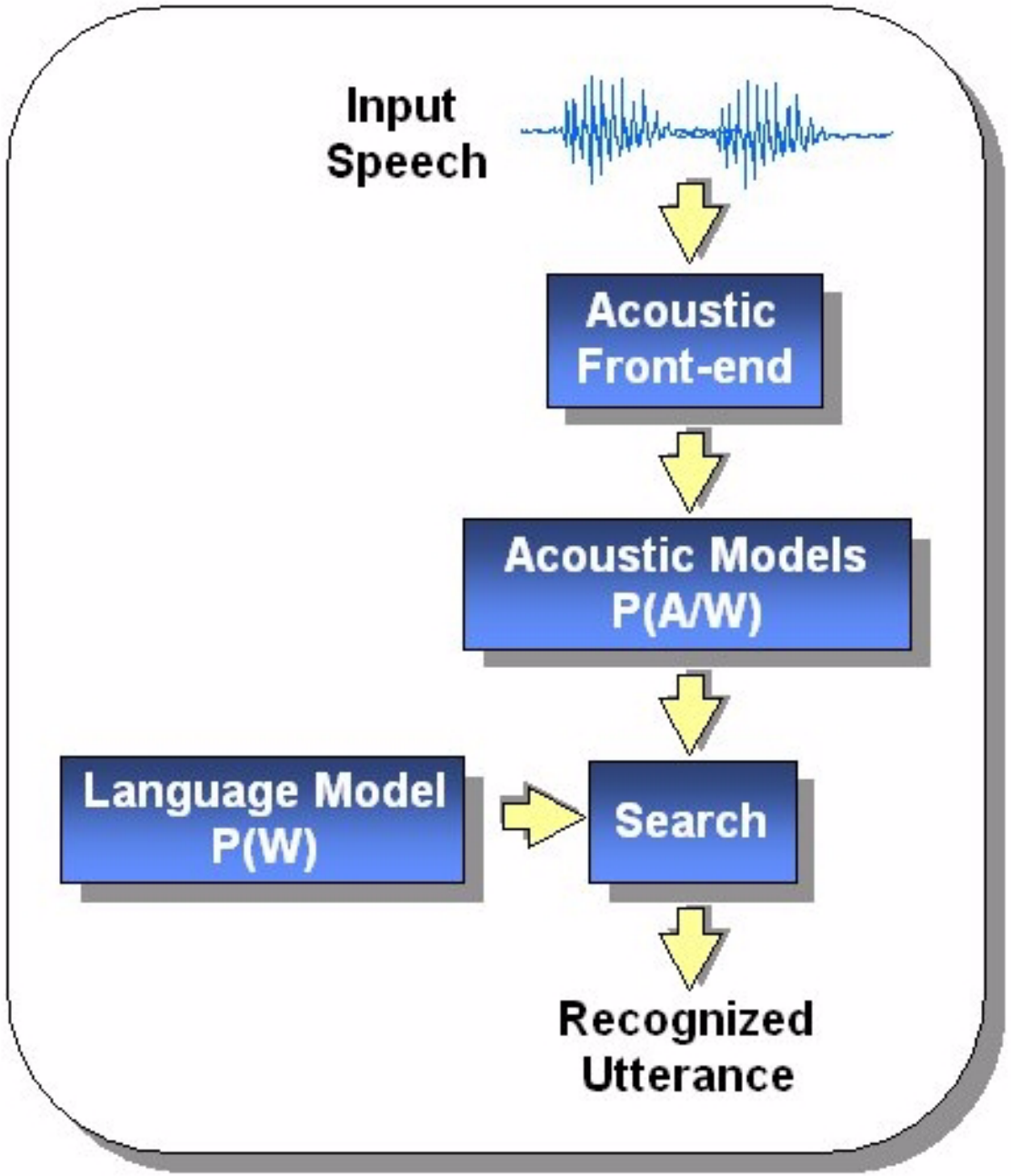
Objective: minimize the word error rate by maximizing $P(W|A)$

Approach: maximize $P(A|W)$ (training)

- $P(A|W)$: acoustic model (hidden Markov models, Gaussians)
- $P(W)$: language model (finite state machines, N-grams)
- $P(A)$: acoustics (ignore during maximization)

INTRODUCTION

SYSTEM COMPONENTS



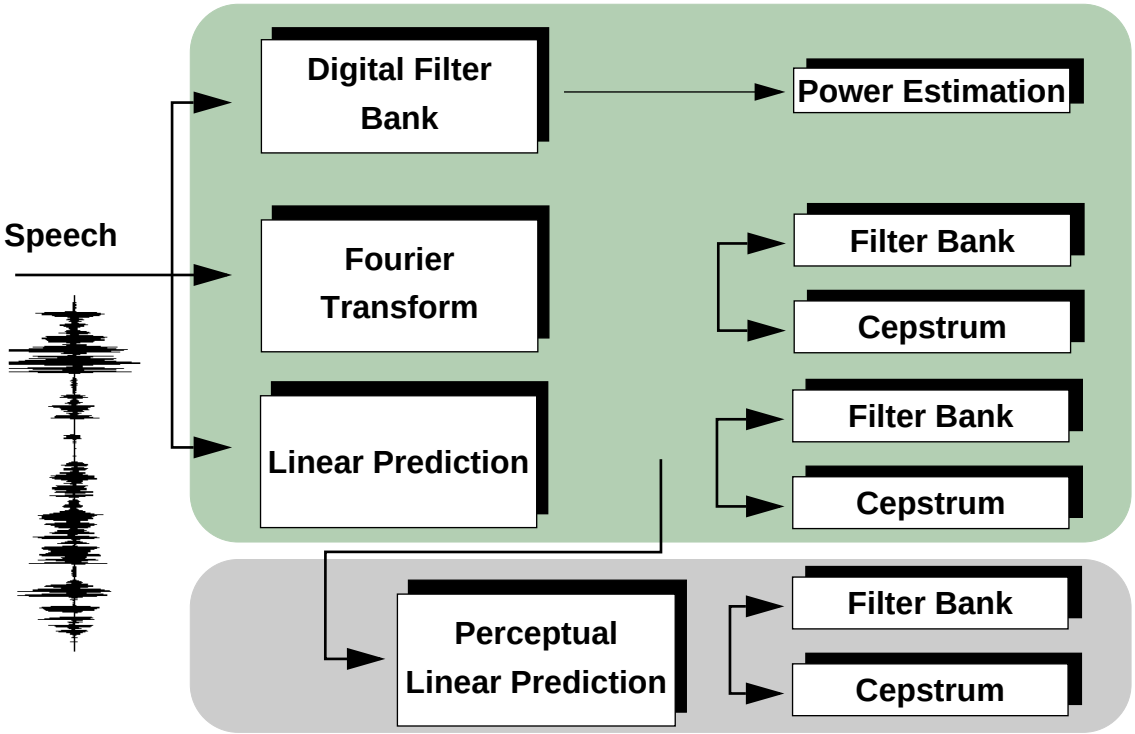
- The signal is converted to a sequence of feature vectors using spectral and temporal measurements.
- Acoustic models represent the sub-word units, such as phonemes, as a finite-state machine.
- The language model predicts the next set of possible words.
- Search is perhaps the most crucial component in the system.

FEATURE EXTRACTION

A FAMILY OF FRONT ENDS

intervoice

Traditional Feature Extraction (1970's)



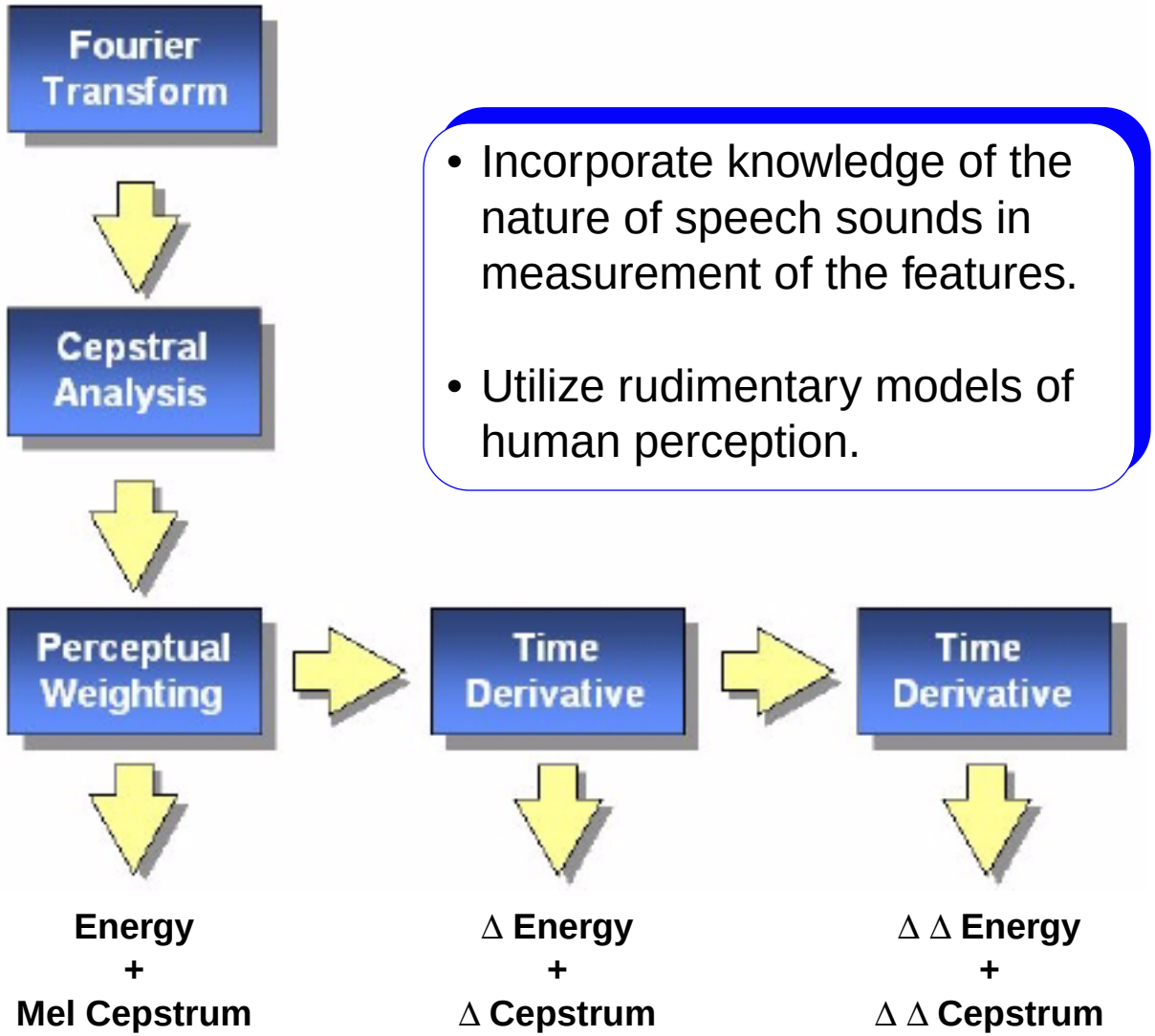
- The Fourier Transform is robust to noise.
- Use absolute measures of the spectrum such as filterbank energies.
- Add normalized temporal energy.

- Exotic spectral estimation techniques did not survive.
- Homomorphic processing (cepstrum) was shown to be an acceptable compromise between performance and complexity.

FEATURE EXTRACTION

MEL FREQUENCY CEPSTRUM COEFFS.

- Measure features 100 times per second.
- Use a 25 msec window for frequency domain analysis (40 Hz res.).
- Include absolute energy and 12 spectral measurements.
- Time derivatives model spectral change.

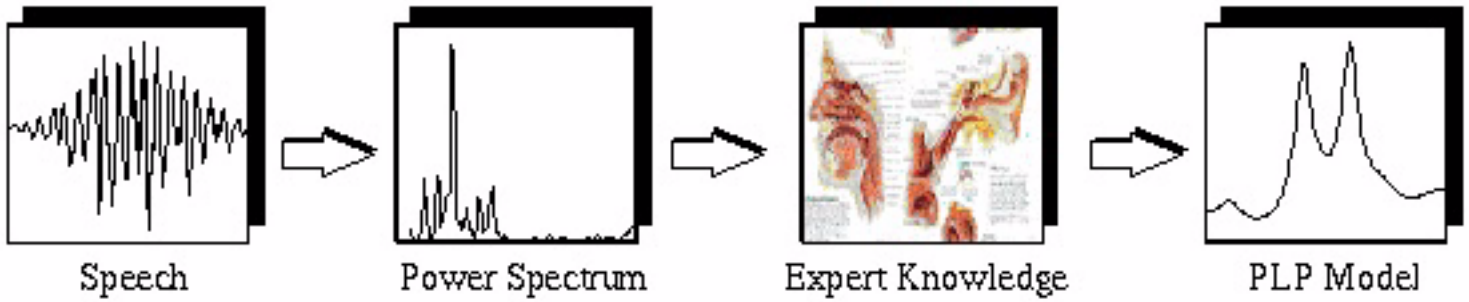


- Incorporate knowledge of the nature of speech sounds in measurement of the features.
- Utilize rudimentary models of human perception.

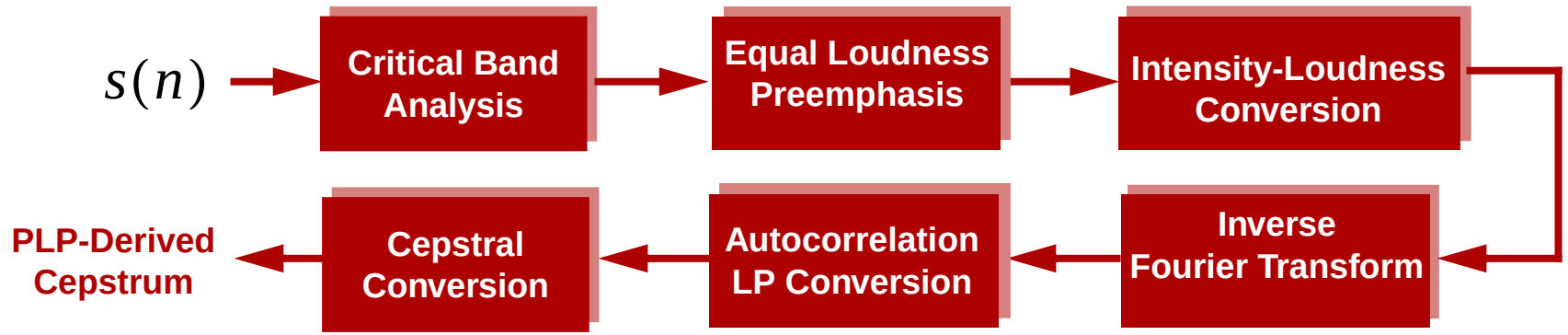
FEATURE EXTRACTION

PERCEPTUAL LINEAR PREDICTION

- Incorporate more knowledge about the physics of speech:



- Processing steps are similar to conventional analysis:

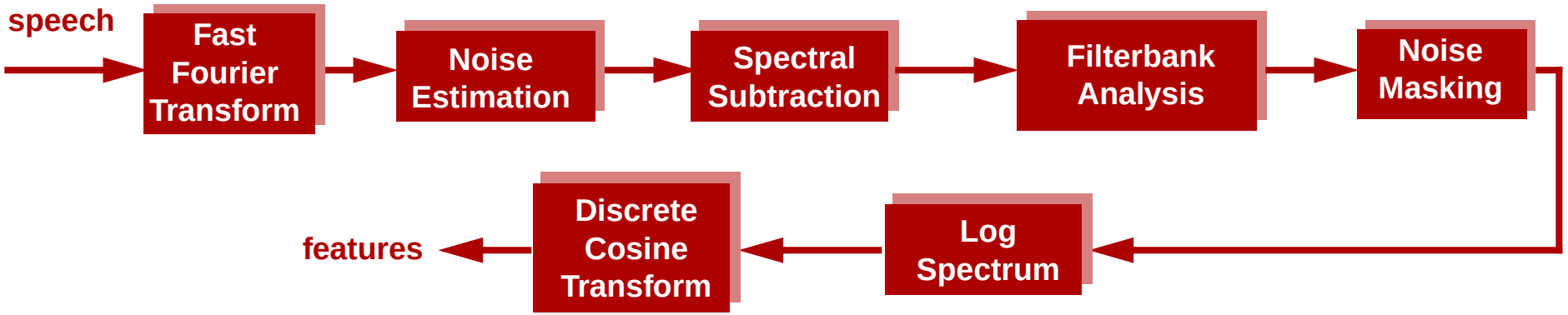


- Word error rate (WER) reduction is very small.

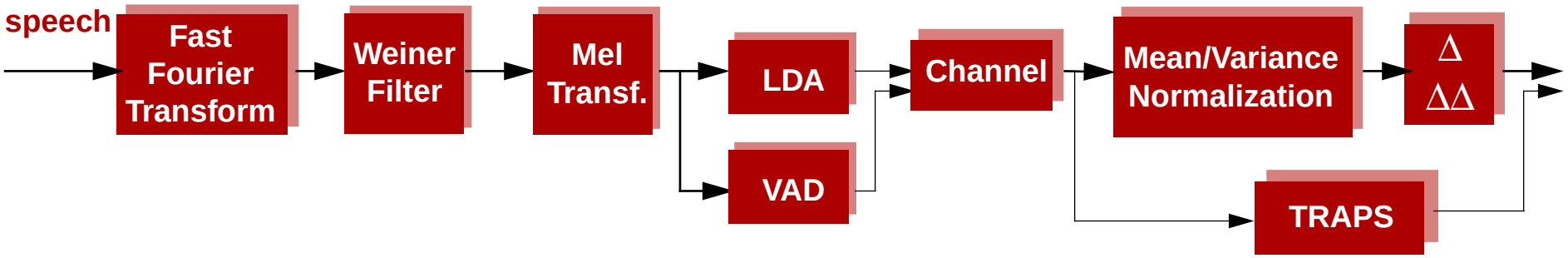
FEATURE EXTRACTION

NOISE COMPENSATION

- Most commercial front ends use adaptive noise compensation:



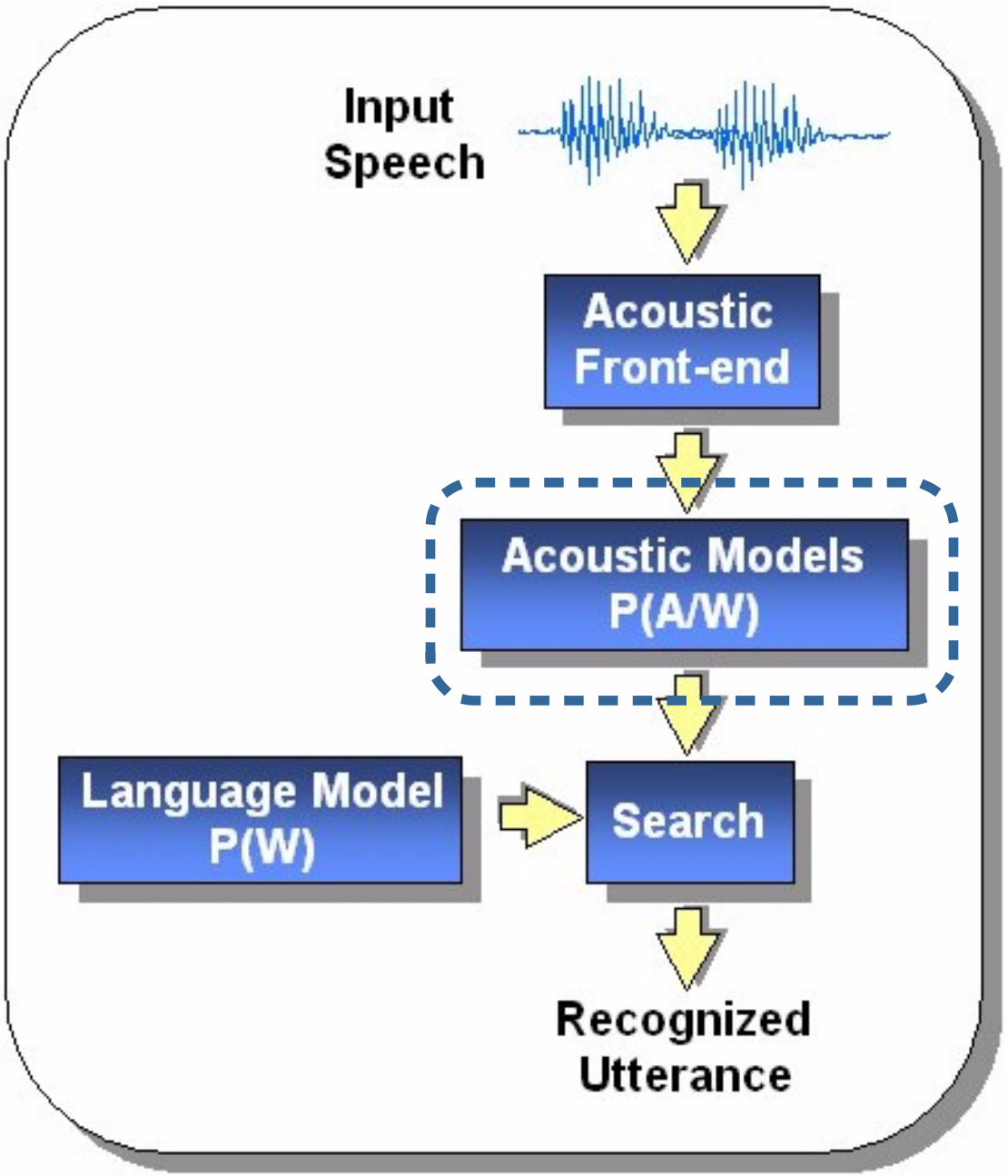
and use long-term spectral structure of speech to remove noise:



- Note that we haven't discussed normalization techniques such as Vocal Tract Length Normalization (VTLN) and adaptation techniques such as maximum likelihood linear regression (MLLR).

AOUSTIC MODELING

HIDDEN MARKOV MODELS

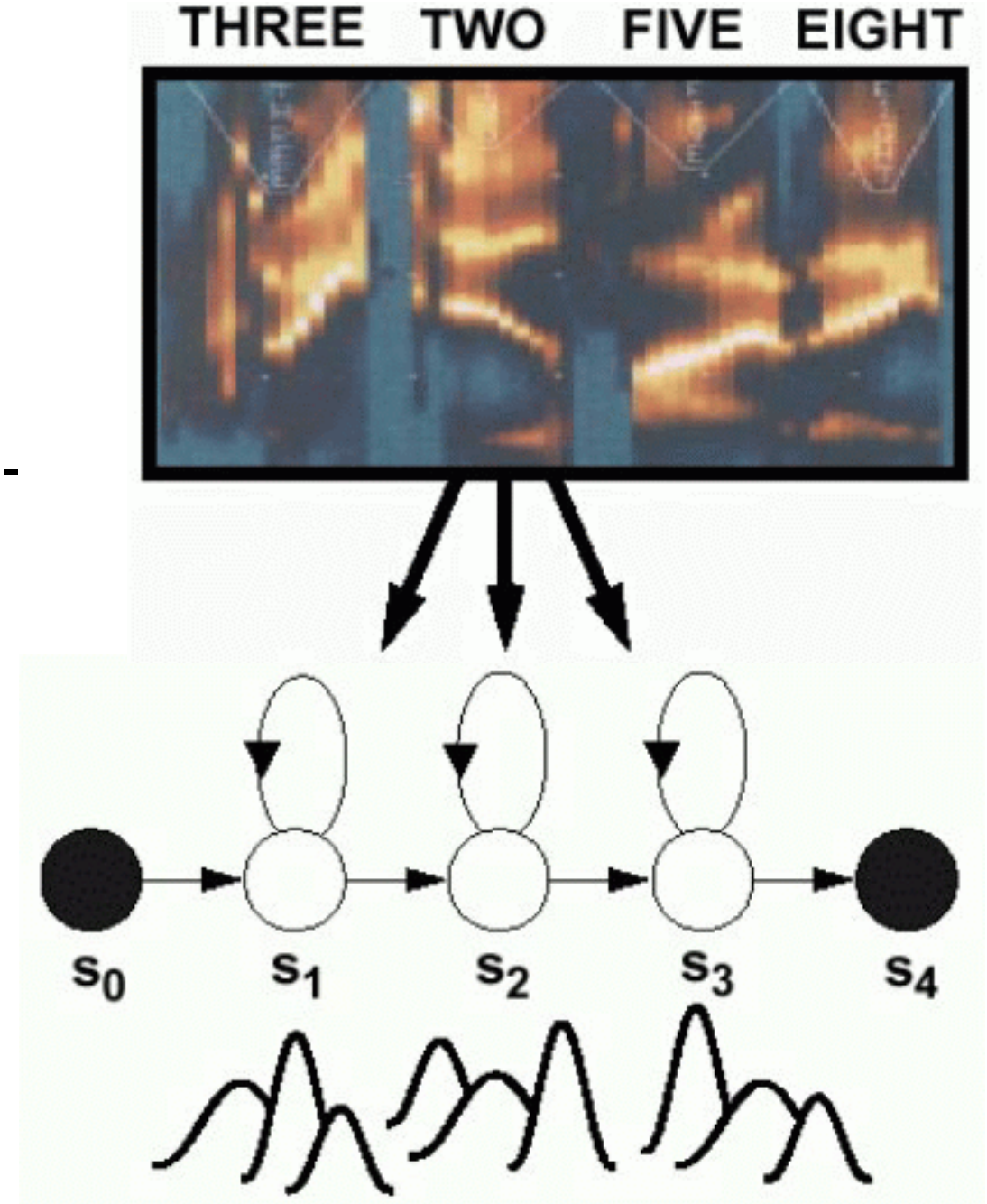


- Acoustic models encode the temporal evolution of the spectrum using a finite state machine consisting of statistical models and transition probabilities.
- We will examine some common acoustic modeling techniques for large vocabulary speech recognition systems.

ACOUSTIC MODELING

HIDDEN MARKOV MODELS

- Gaussian mixture distributions are used to account for variations in speaker, pronunciation, etc.
- Phonetic model topologies are simple three-state left-to-right structures.
- Model topologies can include skip states and multiple paths.
- Sharing model parameters is a common strategy to reduce complexity.

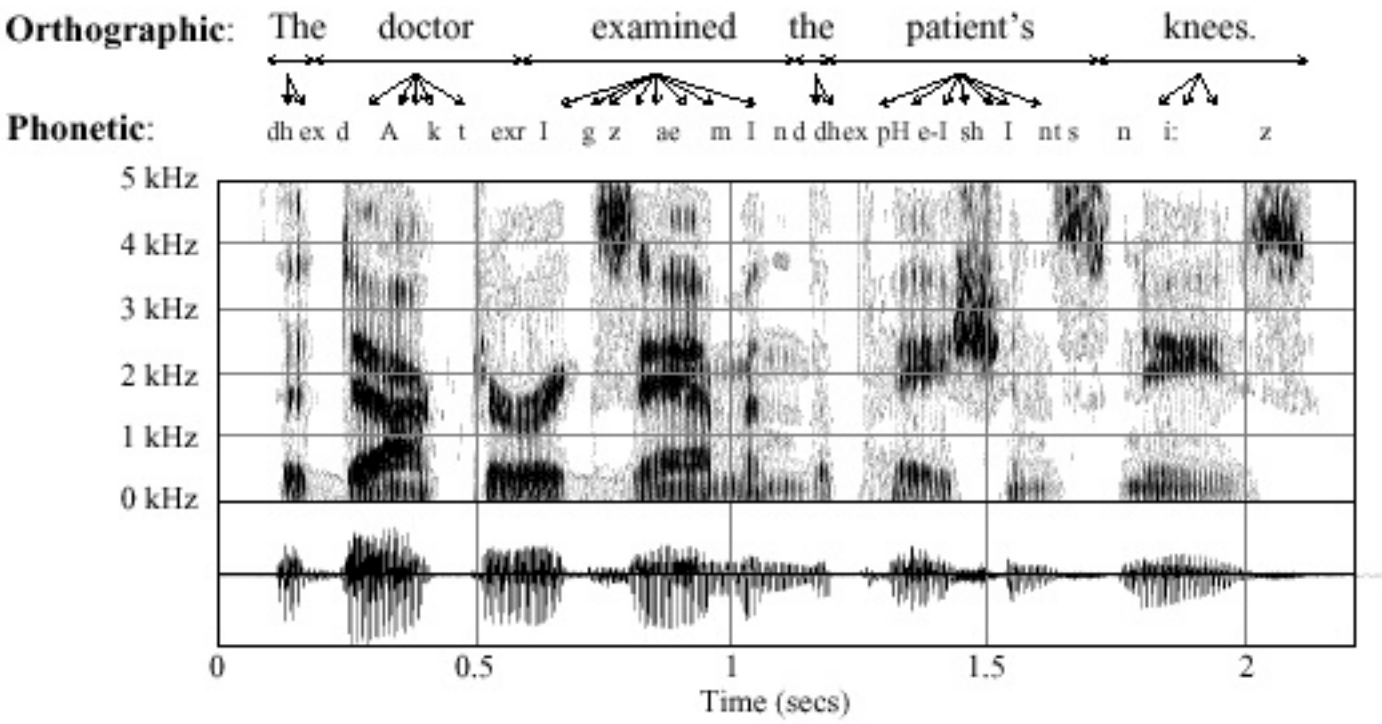


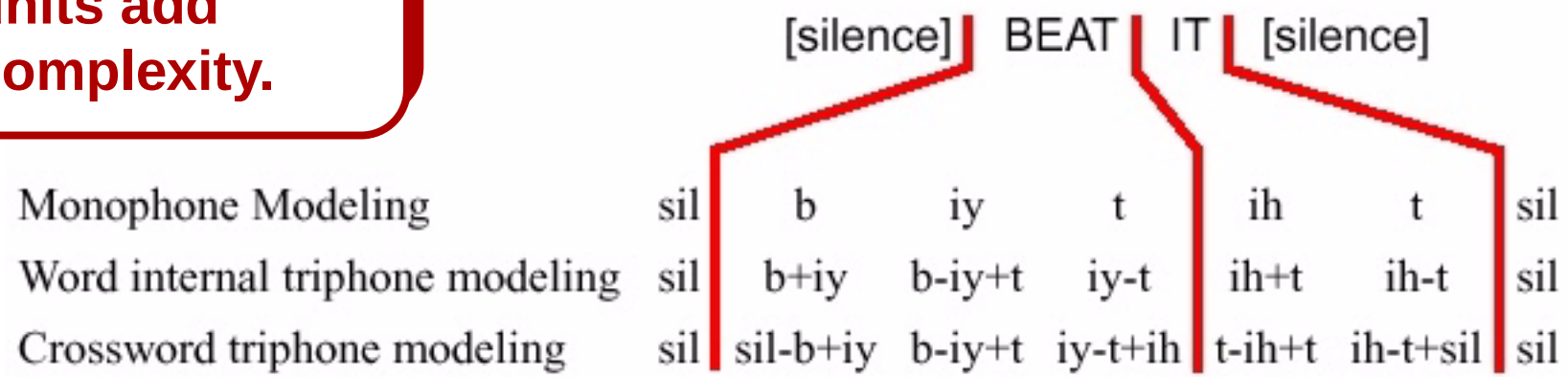
ACOUSTIC MODELING

CONTEXT-DEPENDENT UNITS

intervoice

- **Phonetic units are preferred.**
- **Training does not require phonetic tran-scriptions.**
- **Many types of phonetic units.**
- **Cross-word units add complexity.**

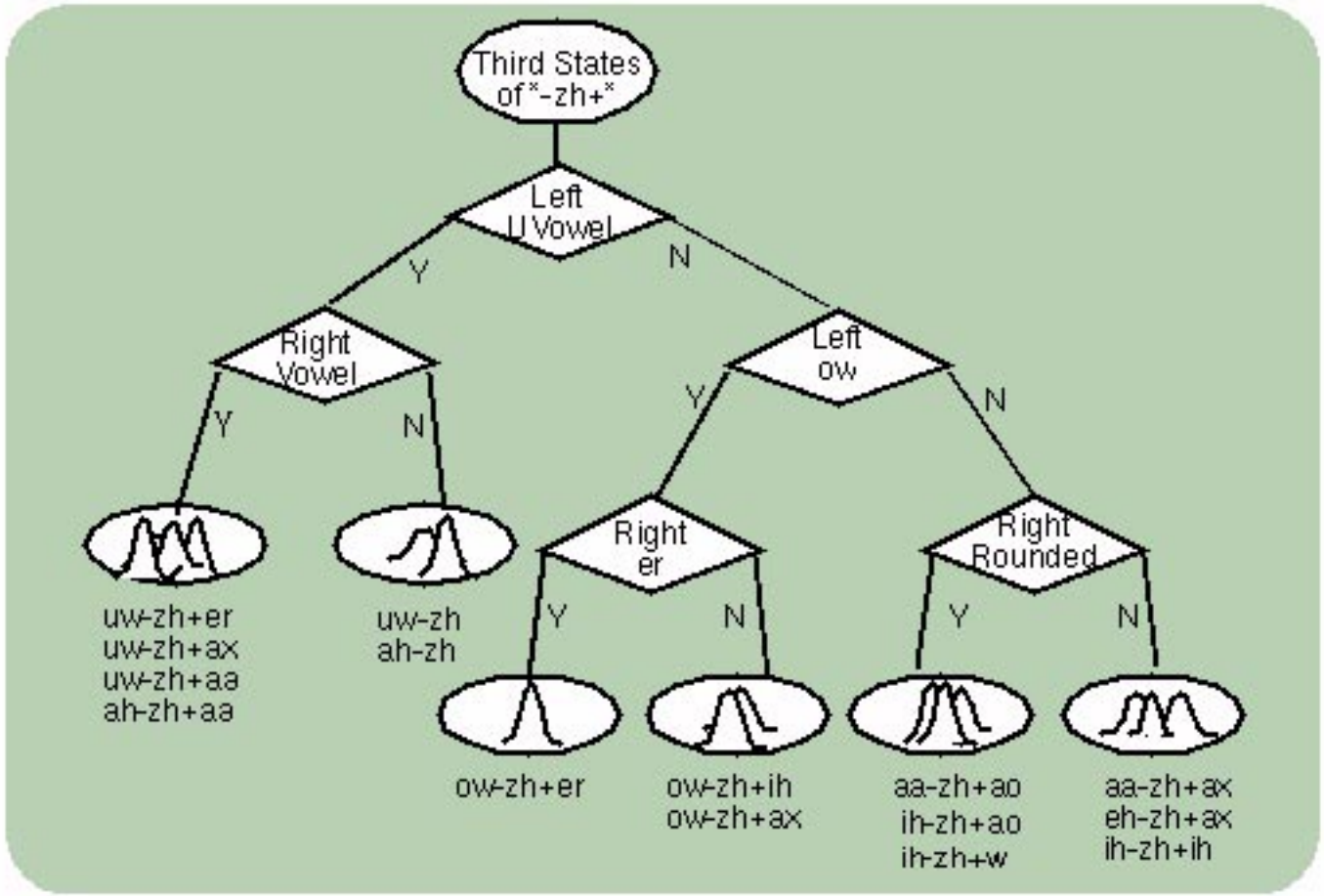




ACOUSTIC MODELING

PARAMETER TYING

- Decision trees are used to determine how to share parameters (e.g., states) between models (reduce complexity) based on linguistic considerations:



ACOUSTIC MODELING

PARAMETER ESTIMATION

intervoice

- Data-driven modeling supervised only from a word-level transcription.
- The EM algorithm is used to improve our estimates:

$$\log P(Data|\bar{\lambda}) \geq \log P(Data|\lambda)$$

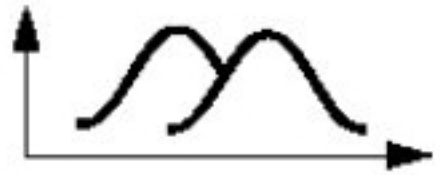
using an MLE approach.

- Computationally efficient training algorithms have been crucial.
- Training is an iterative process.
- Batch mode parameter updates are typically preferred.

- Initialization

- Single Gaussian Estimation

- 2-Way Split



- Mixture Distribution Reestimation

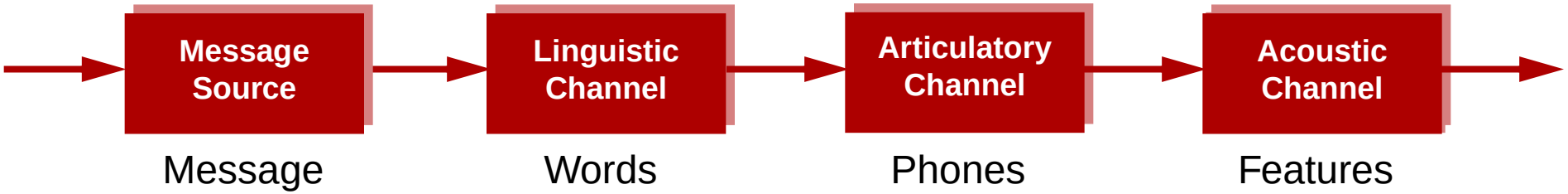
- 4-Way Split

- Reestimation

...

LANGUAGE MODELING

NOISY COMMUNICATION CHANNEL MODEL



Bayesian formulation for speech recognition:

$$P(W|A) = P(A|W)\boldsymbol{P(W)}/P(A)$$

Objective: minimize the word error rate by maximizing $P(W|A)$

- A language model typically predicts a small set of next words based on knowledge of a finite number of previous words (N-grams) — leads to search space reduction.

- There are many ways to estimate or approximate $\boldsymbol{P(W)}$; smoothing of these estimates is also important.

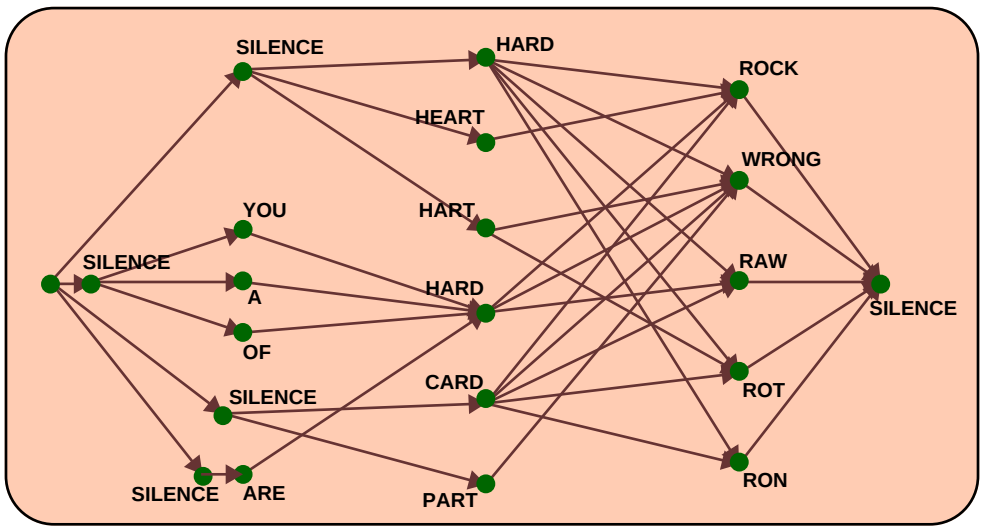
LANGUAGE MODELING

WORD PREDICTION — HUMANS DO IT!

LANGUAGE MODELING

FINITE STATE AUTOMATA

- The search space for large vocabularies is unmanageable if we allow any word to follow any other word (e.g., loop grammar).
- Only a small subset of the vocabulary can follow a given word hypothesis, but this subset is “context-sensitive”.
- In real applications, a user-interface design results in a specification of a language or collection of sentence patterns that are permissible.
- A simple way to express and manipulate this information in a dynamic programming framework is a via a state machine, shown to the right.
- Such networks are often called finite state grammars (FSG), automata (FSA), or transducers (FST).



LANGUAGE MODELING

FORMAL LANGUAGES

intervoice

Finite state machines are one of many types of grammar formalisms that can be used to process language. We categorize these formalisms by their generative capacity (the Chomsky hierarchy):

Type of Grammar	Constraints	Automata
Phrase Structure	A -> B	Turing Machine (Unrestricted)
Context Sensitive	aAB -> aBb	Linear Bounded Automata (N-grams, Unification)
Context Free	A -> w A -> BC	Push down automata (CFG, BNF, JSGF, RTN)
Regular	A -> w A -> wB	Finite state automata (Network Decoding)

• CFGs offer a good compromise between parsing efficiency and representational power, and provide a natural bridge between speech recognition and natural language processing.

LANGUAGE MODELING

N-GRAM LANGUAGE MODELS

intervoice

$$P(\boldsymbol{W}) = P(w_1 w_2 w_3 \ldots w_n)$$

$$= \prod_{i=1}^{n} P(w_i | w_1, w_2, \ldots, w_{i-1})$$

$$= \prod_{i=1}^{n} P(w_i | \Phi(w_1, w_2, \ldots, w_{i-1}))$$

Three important simplifications:

- Unigram: $\Phi(w_1, w_2, \ldots, w_{i-1}) = \phi$
- Bigram: $\Phi(w_1, w_2, \ldots, w_{i-1}) = w_{i-1}$
- Trigram: $\Phi(w_1, w_2, \ldots, w_{i-1}) = w_{i-1}, w_{i-2}$

N-grams: approx. $P(\boldsymbol{W})$ as a product of conditional probabilities, referred to as histories.

- Histories can be merged; negligible loss in performance (equivalence classes).
- Many real-time systems use bigrams for computational efficiency reasons.
- Trigram models require statistical smoothing techniques for reliable estimation of probabilities.
- Performance improvements for trigrams are modest (less than 10% relative).

LANGUAGE MODELING

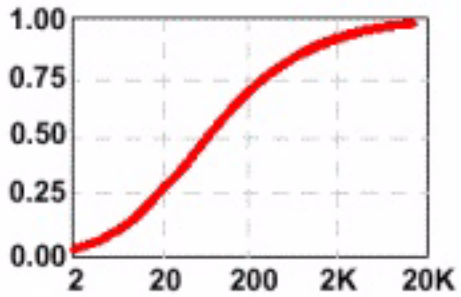
ESTIMATING N-GRAM LANGUAGE MODELS

intervoice

- N-gram models are a popular alternative because they can be implemented efficiently and provide a CSG capability:

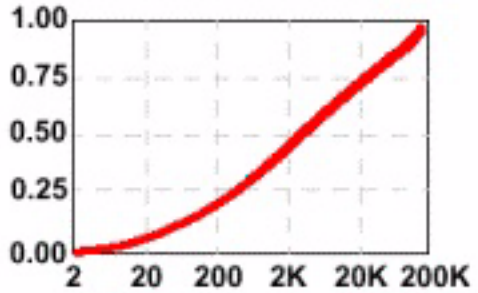
Unigrams (SWB):

- Most Common: I, and, the , you, a
- Rank-100: she, an, going
- Least Common: Abraham, Alastair, Acura



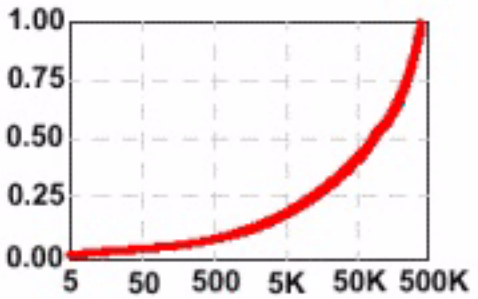
Bigrams (SWB):

- Most Common: "you know", "yeah S!", "!S um-hum", "I think"
- Rank -100: "do it", "that we", "don't think"
- Least Common: "raw fish", "moisture content, "Reagan Bush"



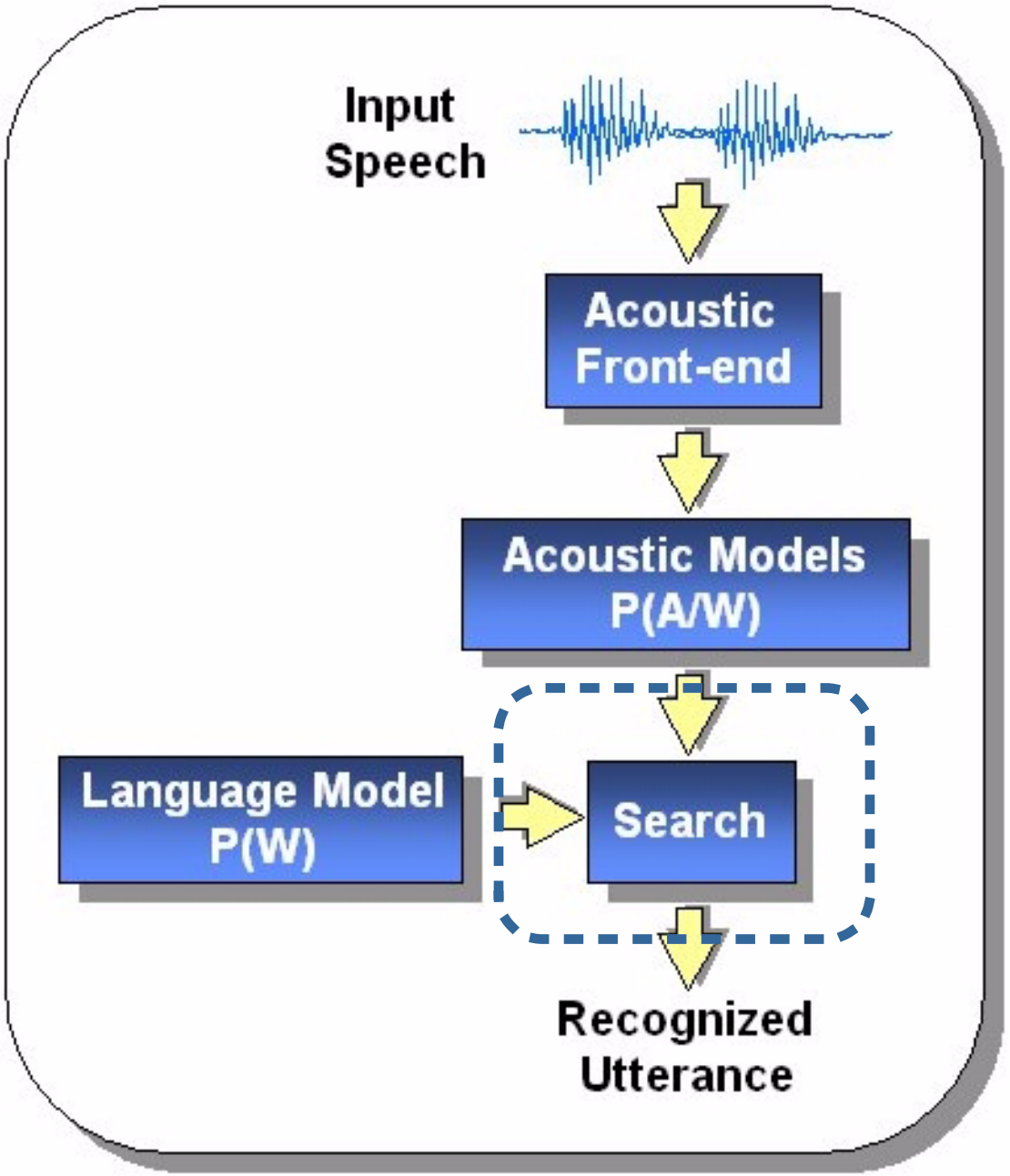
Trigrams (SWB):

- Most Common: "!S um-hum S!", "a lot of", "I don't know"
- Rank-100: "it was a", "you know that"
- Least Common: "you have parents", "you seen Brooklyn"



SEARCH TECHNIQUES

PRINCIPLES OF SEARCH

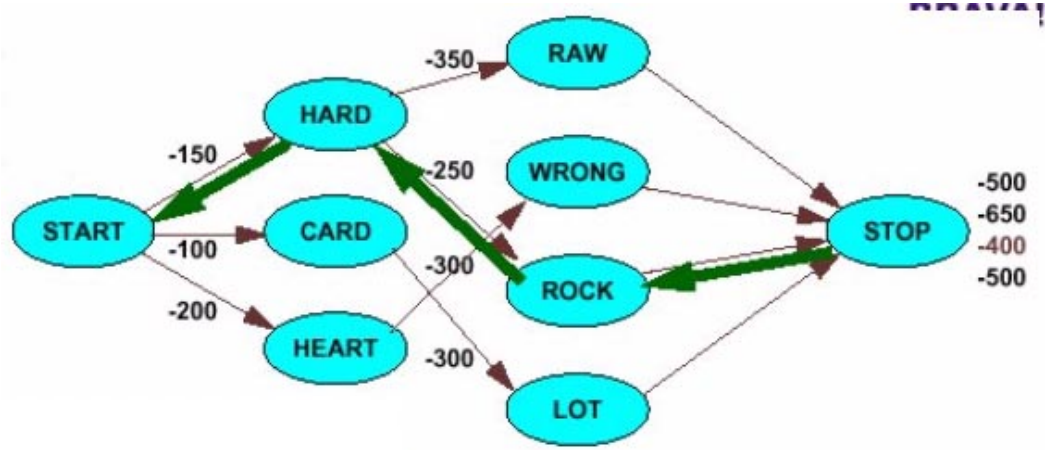


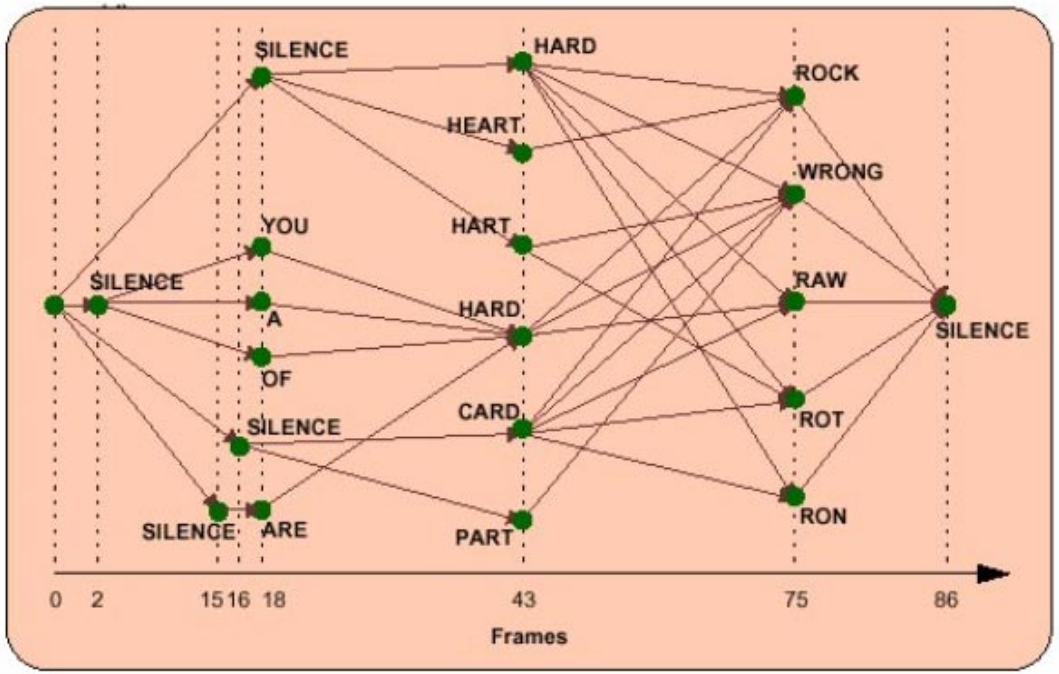
- Search algorithms are based on principles of dynamic programming (Viterbi decoding)
- Finding globally optimal solutions can be very expensive
- Suboptimal solutions work well in practice
- Search complexity must be linear w.r.t. the length of the utterance to be practical
- Most research systems use multiple passes and invoke several search algorithms
- Lookahead and pruning are essential parts of search

SEARCH TECHNIQUES

DYNAMIC PROGRAMMING

- Dynamic programming is used to find the most probably path through the network.
- Beam Search: paths with low probabilities are discarded early in the search process.



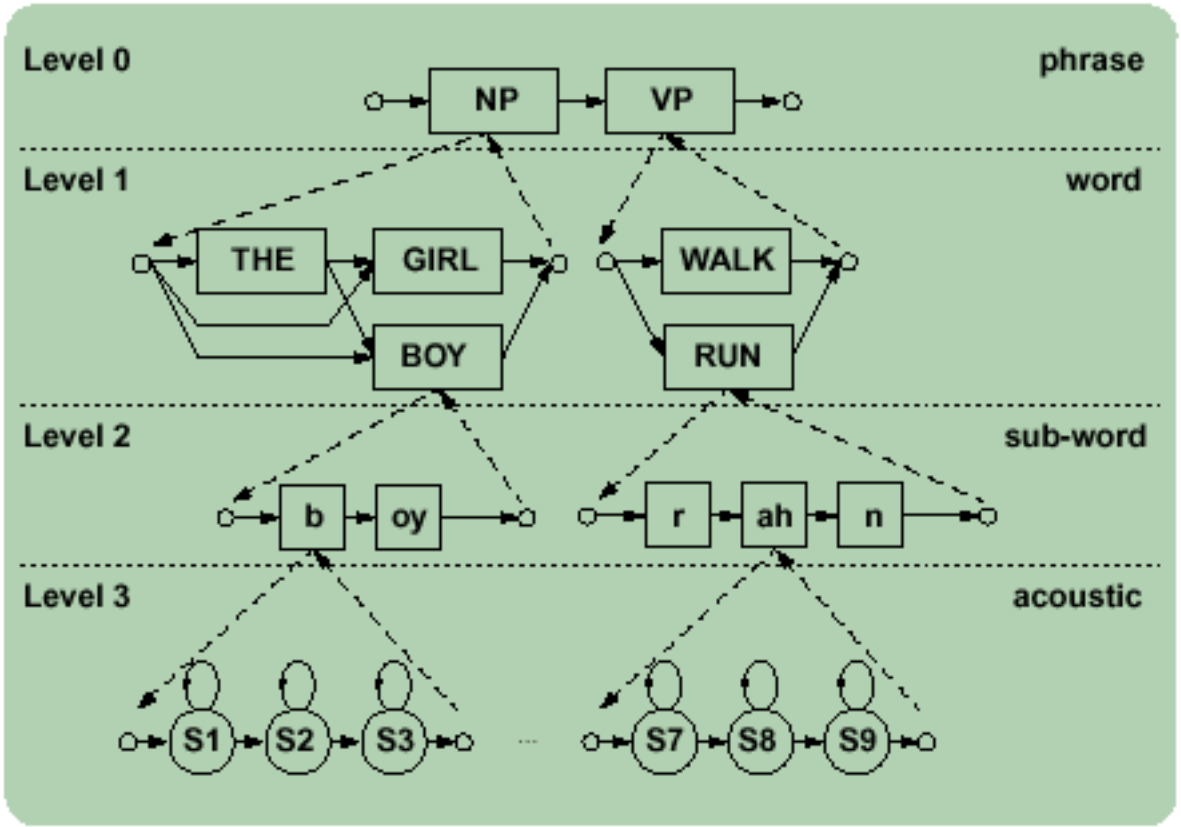


- Search is time synchronous and left-to-right.
- Arbitrary amounts of silence must be permitted between each word.
- Words are hypothesized many times with different start/stop times, which significantly increases search complexity.

SEARCH TECHNIQUES

HIERARCHICAL SEARCH

- In practice, a system might utilize many knowledge sources (e.g., part of speech, word, phone, and acoustic model.
- Breadth-first time-synchronous hierarchical search is very convenient for integrating linguistic constraints.
- Efficient Viterbi search of a hierarchical network is a much more complicated problem because of ambiguity in the network (e.g., the same word sequence can appear multiple places in the network.
- Special care must be taken to synchronize all hypotheses so each acoustic model is evaluated as few times as possible.

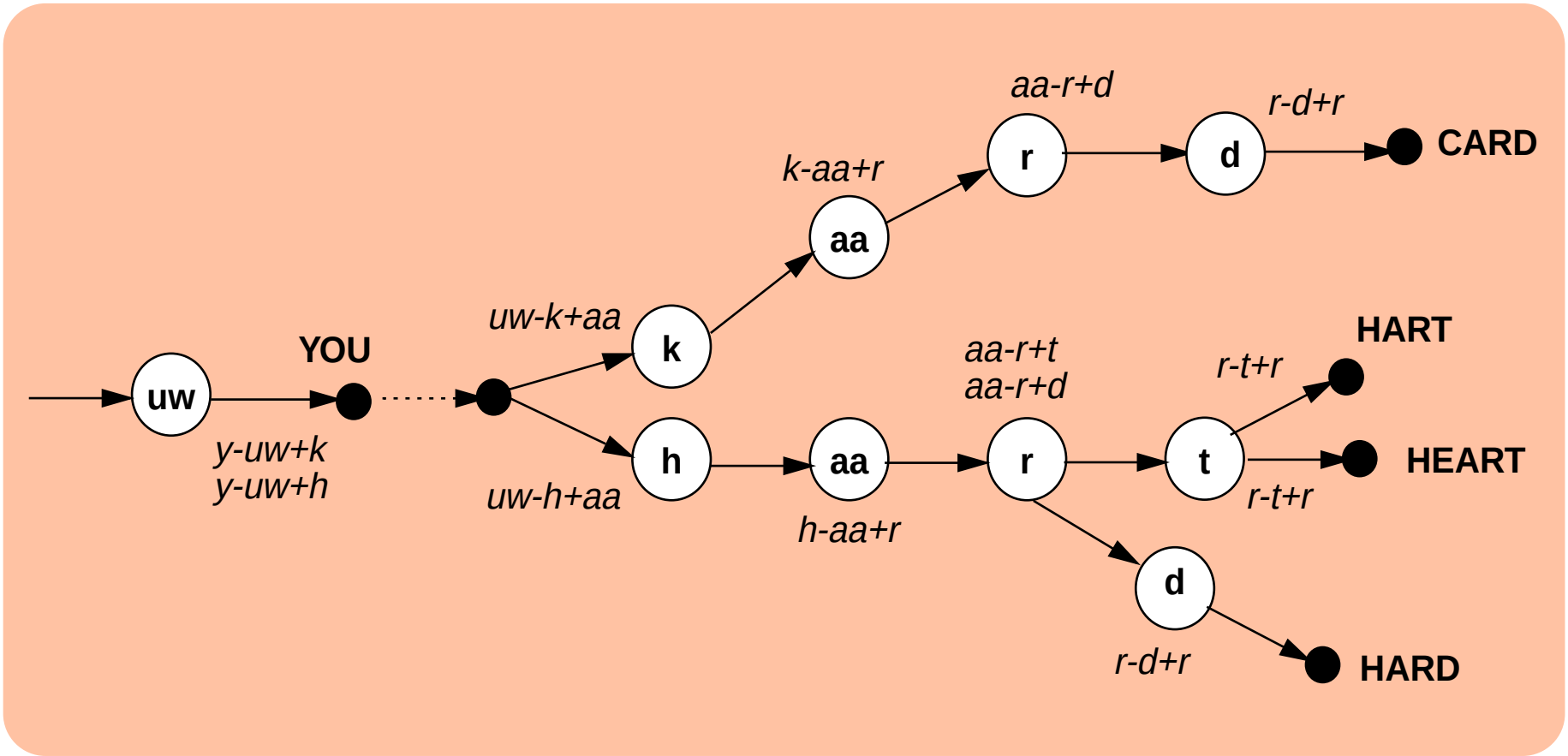


- Since many hypothesis might need the same phone at the same time, coordinating this search becomes a nontrivial problem.

SEARCH TECHNIQUES

CROSS-WORD DECODING

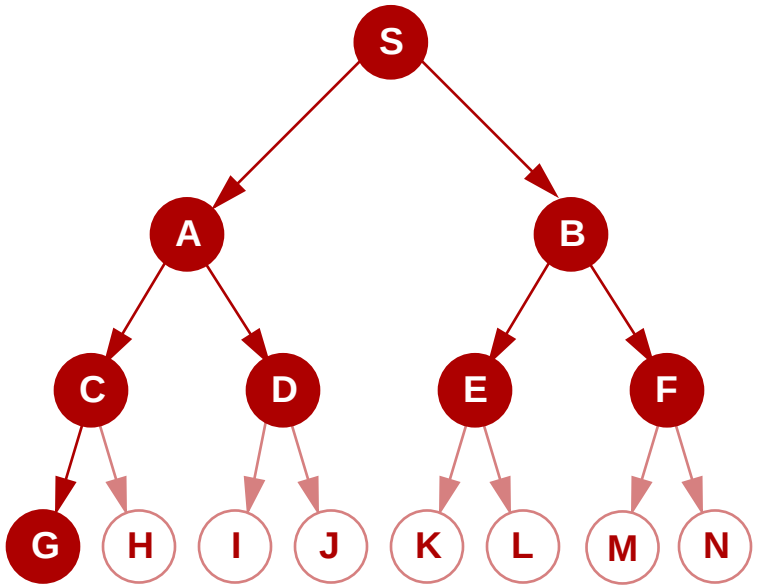
- Cross-word decoding: since word boundaries don't occur in spontaneous speech, we must allow for sequences of sounds that span word boundaries.
- Cross-word decoding significantly increases memory requirements.
- The lexicon can be converted to a tree structure (lexical trees) to improve efficiency.



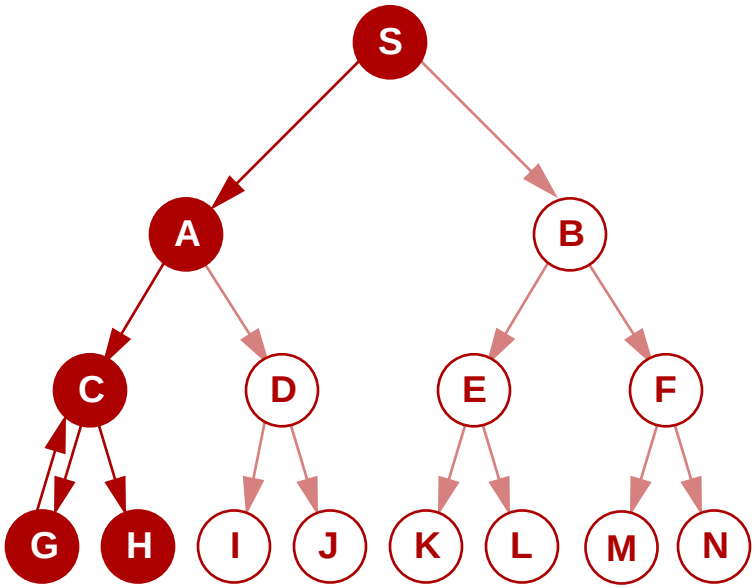
SEARCH TECHNIQUES

FAMILY OF SEARCH ALGORITHMS

- Time-synchronous Viterbi search is one of many types of search algorithms.
- It belongs to a class of search algorithms known as breadth-first.
- Beam search (suboptimal search) is typically easier to implement for breadth-first search algorithms.



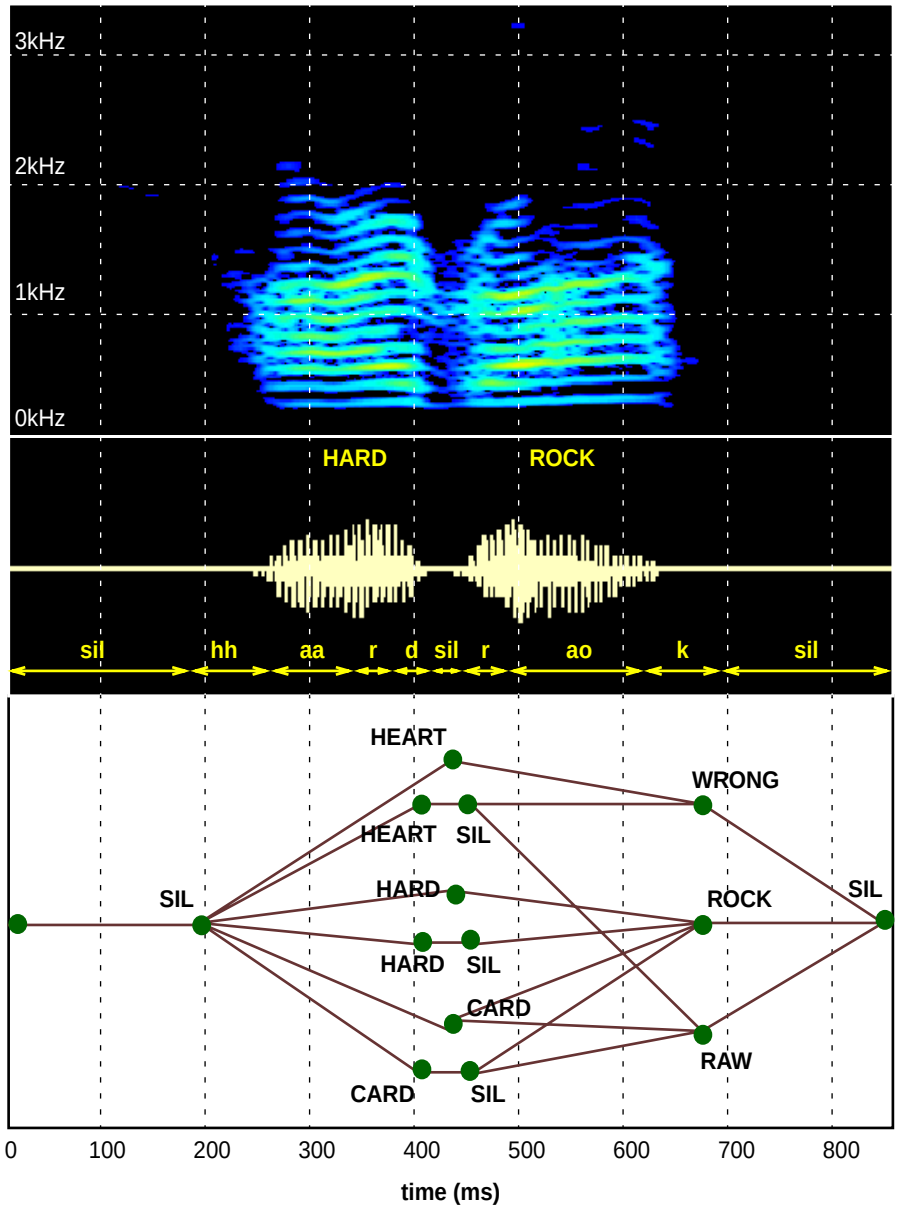
- Other popular search algorithms are based on depth-first search.
- Stack decoding (IBM) and N-best list generation are two examples of this search approach.
- Stack decoding can be very fast, and use minimal resources, if accurate heuristics are available.



SEARCH TECHNIQUES

WORD GRAPH GENERATION

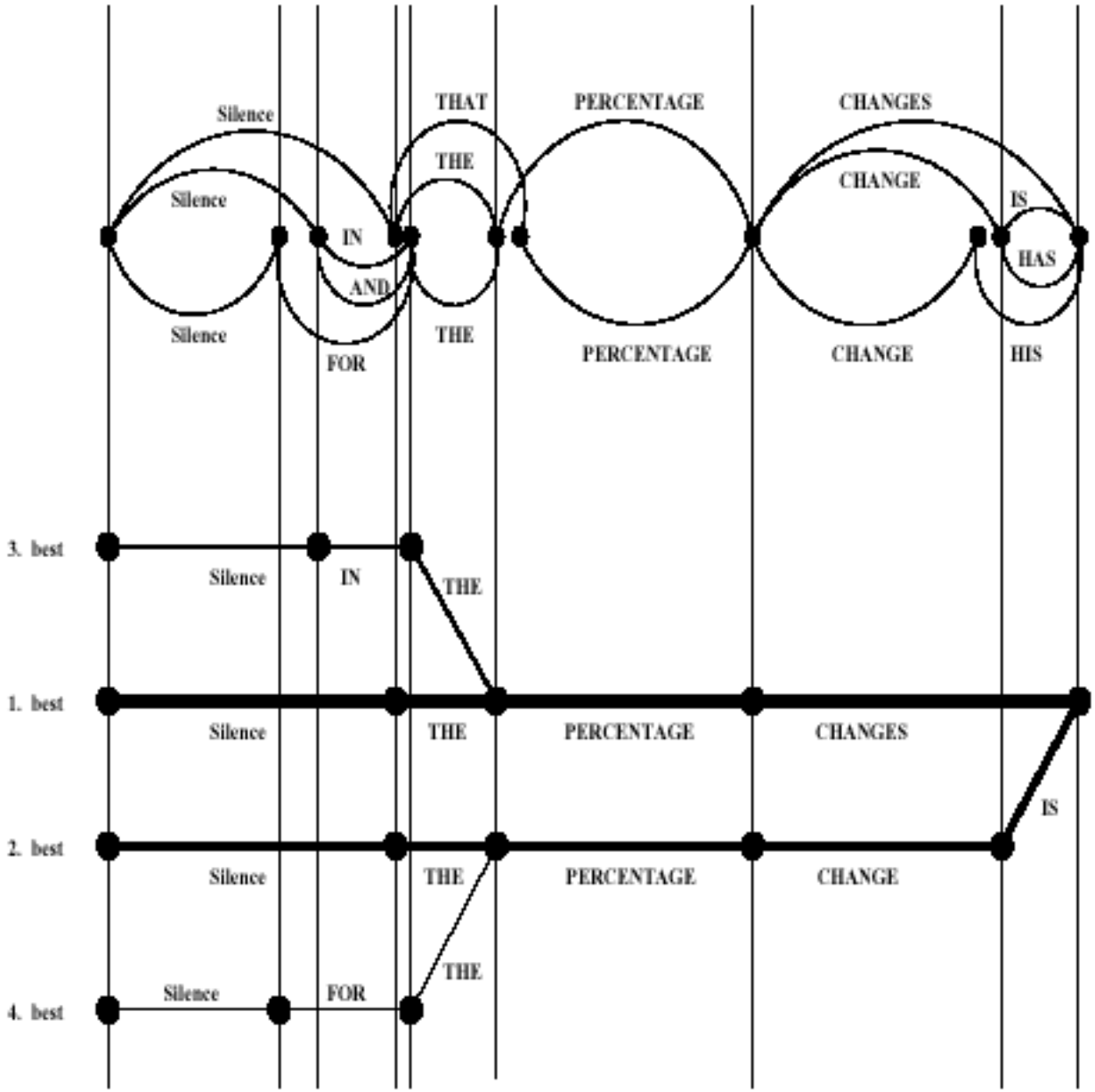
- Direct searching of trigram language models is very expensive.
- Application of higher order language models (quadgrams) and acoustic models (pentaphones) is difficult in a single-pass search.
- Rescoring of word graphs is a practical alternative.
- Word graph generation is expensive, and performed using an expanded Viterbi-style search.
- An important figure of merit is the word graph error rate.
- Word graph compaction and postprocessing is a popular area of research (e.g., sausages).



SEARCH TECHNIQUES

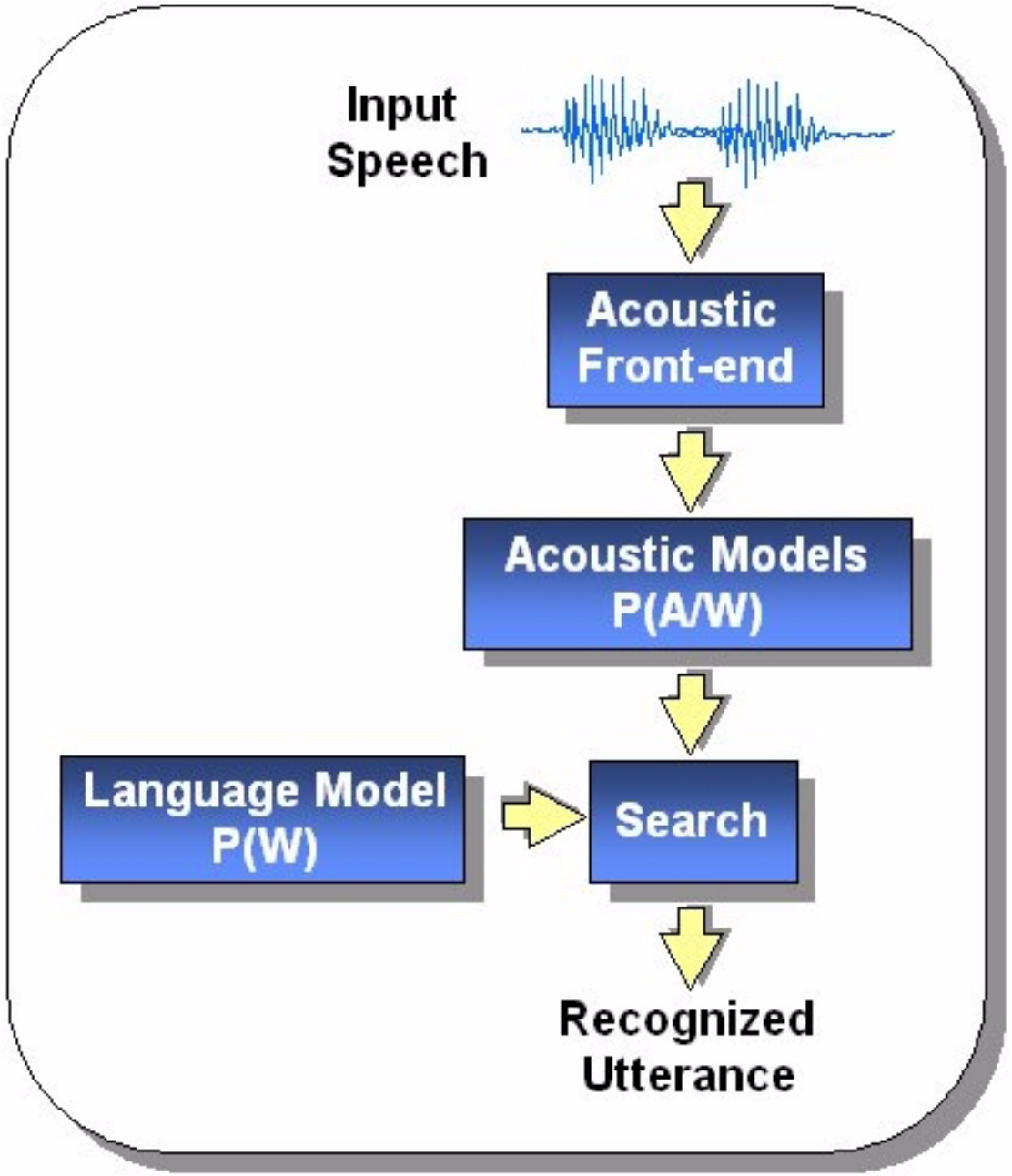
N-BEST LISTS

- N-best lists are a popular alternative to word graphs (smaller, faster).
- Useful as input to natural language postprocessors
- Word error rate asymptotically approaches zero as N increases.
- The top N hypotheses can be rescored using more complex acoustic models, language models, and linguistic constraints.
- A more compact format than word graphs.



STATE OF THE ART

COMMODITY TECHNOLOGY?

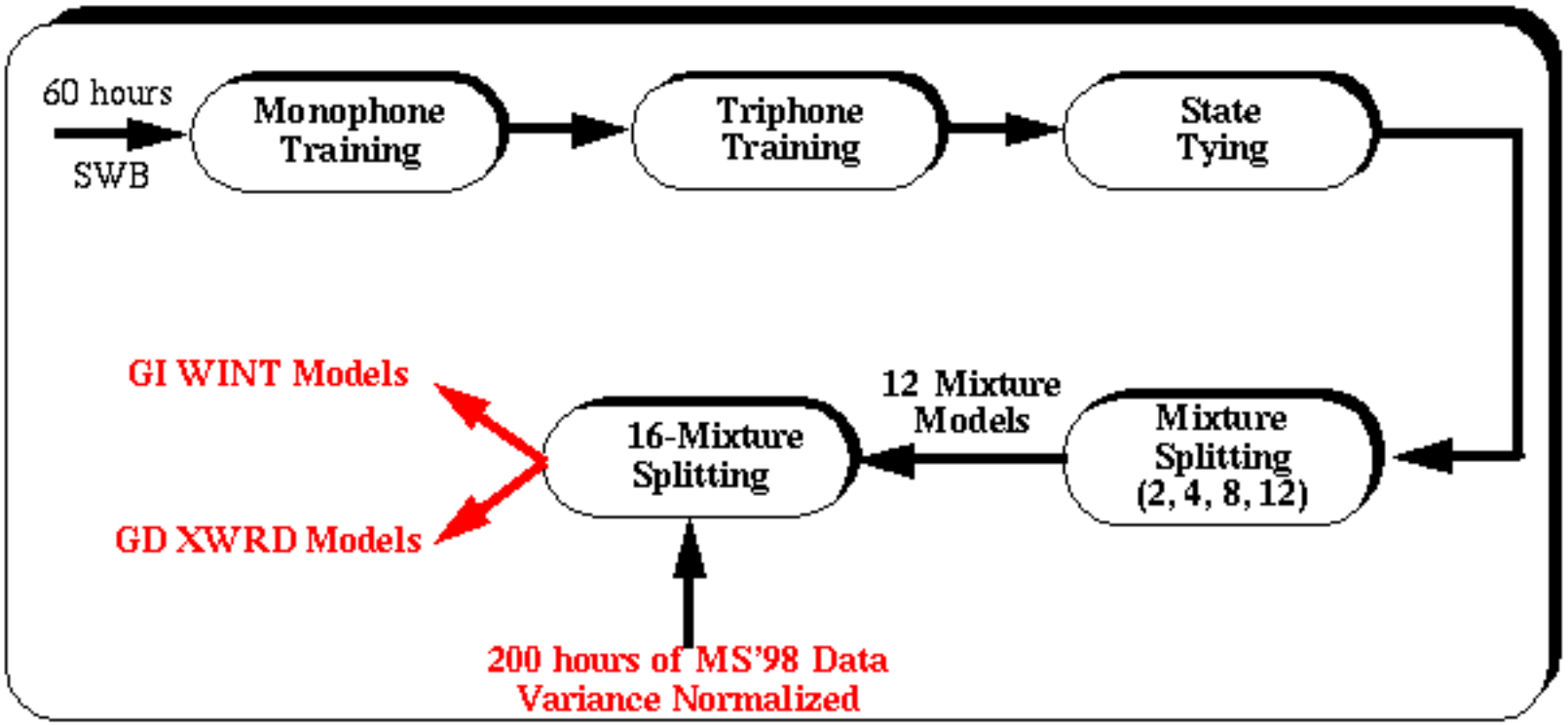


- Software might be a commodity, but the experience required to build a complex system is great.
- Speech recognition systems are far too complex for the performance they deliver.
- We will examine some common characteristics of state of the art systems.

STATE OF THE ART

TYPICAL TRAINING RECIPES

- State of the art systems use several million free variables.
- Training requires almost 40 passes over the data, and several hundred hours of data to achieve high performance.
- Models are often “bootstrapped” from a previous stage of training, or even a previous application development.



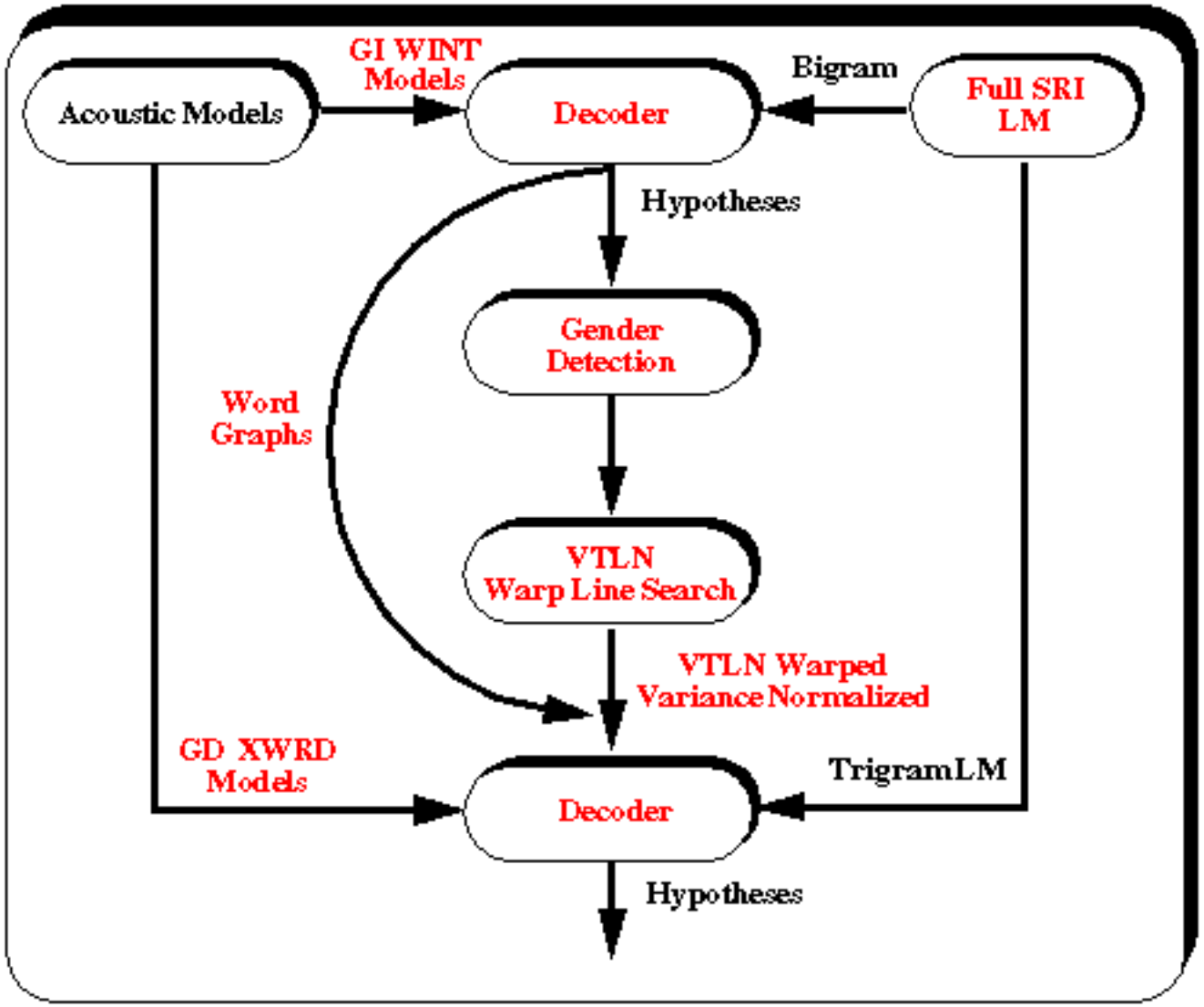
STATE OF THE ART

TYPICAL MULTIPASS DECODING

Decoding typically involves two steps:

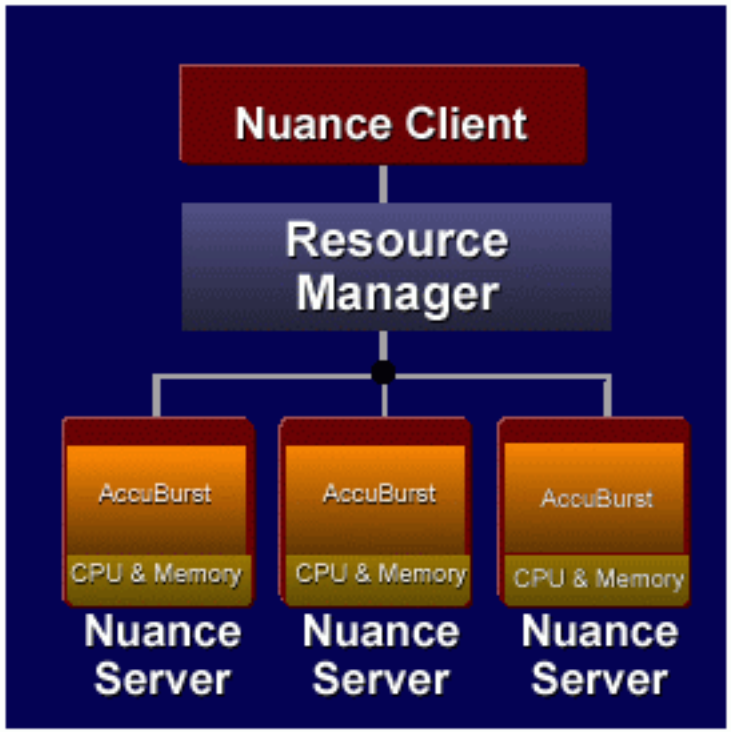
(1) generation of a word graph using a bigram language model;

(2) recognition using cross-word triphones and a trigram LM.



STATE OF THE ART

ENTERPRISE SOLUTIONS



Strengths:

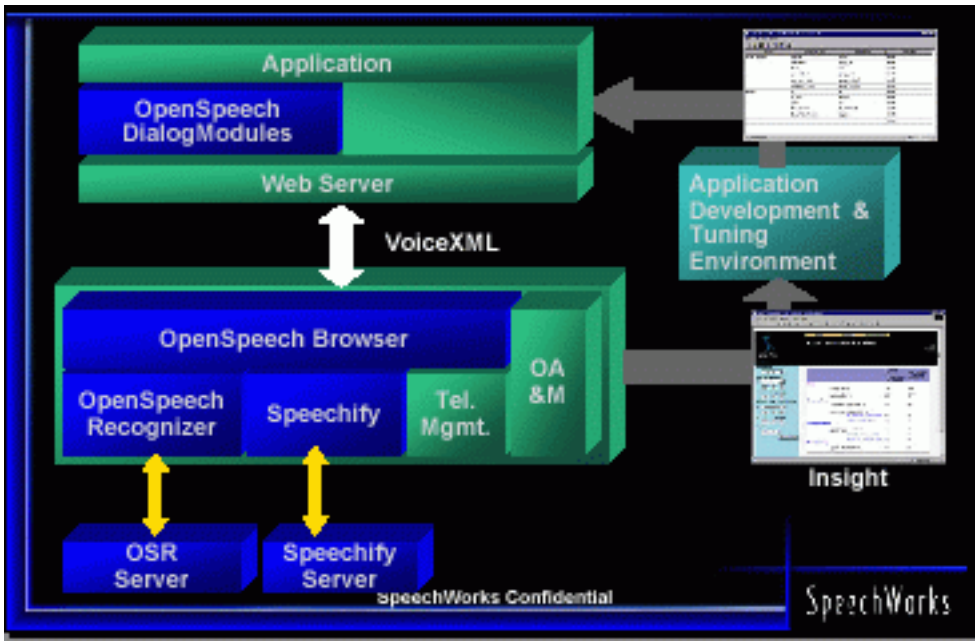
- Resource efficiency
- Multilingual support
- Robustness/Adaptation

Nuance v8.0 Features:

- Based on SRI's DECIPHER system
- 27-dimensional mel-frequency cepstral coefficients
- 3-state triphone hidden Markov models (with mixture-tying)
- N-gram and network language models
- Barge-in (echo cancellation); voice activity detection
- Dynamic language detection
- MLLR and MAP adaptation
- Noise robustness (acoustic models for land lines, cellular, and automotive)

STATE OF THE ART

RAPID APPLICATION DEVELOPMENT



Strengths:

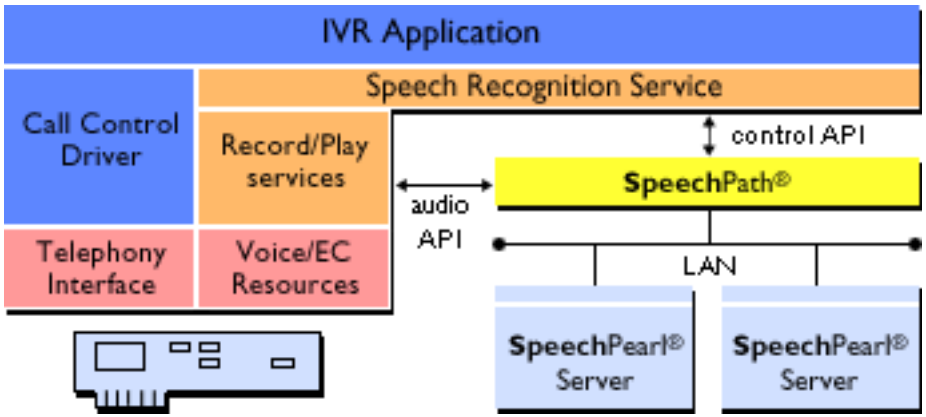
- Flexible configuration and run-time efficiency through finite state transducer technology
- Early adopter of VoiceXML and open architectures

SpeechWorks (OpenSpeech):

- Originally based on MIT's segment-based recognizer.
- Transitioned to AT&T's finite State Transducer technology.
- Segmental statistical models used for phone classification.
- Unsupervised, automatic adaptation.
- Parallel grammars; dynamic grammar compilation; grammar caching; grammar and lexicon updates.

STATE OF THE ART

DIVERSE APPLICATION SUPPORT



Strengths:

- Core search engine
- Support dictation, telephony, and mobile computing applications
- Multilingual support
- Natural language support

Philips SpeechPearl:

- Leverages years of internal speech recognition research
- Open and closed grammars
- Natural language interpretation
- Mixed acoustic models (whole word and triphone models)
- Confidence measures and out-of-vocabulary rejection
- Dynamic grammar and lexicon switching
- Optimization for tonal languages

STATE OF THE ART

OPEN SOURCE TECHNOLOGY

UNIVERSITY OF CAMBRIDGE

htk3

- Hidden Markov Model Toolkit (HTK)
- Research-only license
- Known for high performance research systems and solid engineering
- Released software lags published research results

Speech at CMU

- Sphinx / Hephaestus
- Research-only license
- Known for impressive demonstrations of integrated technology (e.g., speech to speech translation)
- Developing Sphinx 4 in Java

Colorado

CSLR

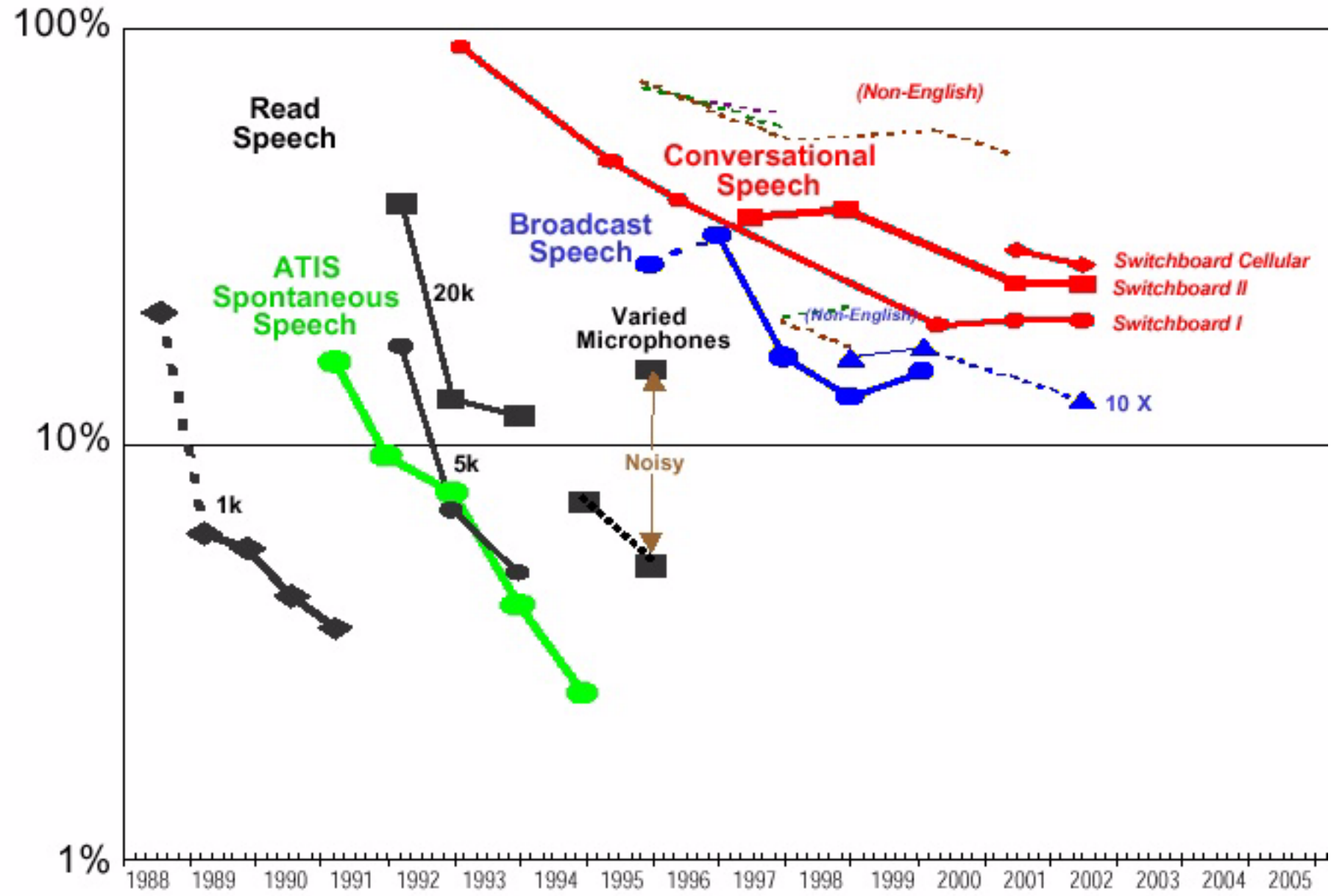
- CU Communicator
- Research-only license; consortium fee
- Known for dialog systems and application development
- Released a DARPA Communicator application for travel

Mississippi State UNIVERSITY

- First and only state of the art public domain speech recognition system
- Designed to accelerate progress in research and to increase participation
- Known for software engineering, ease of use, and comprehensive toolkits

STATE OF THE ART

COMMON EVALUATIONS

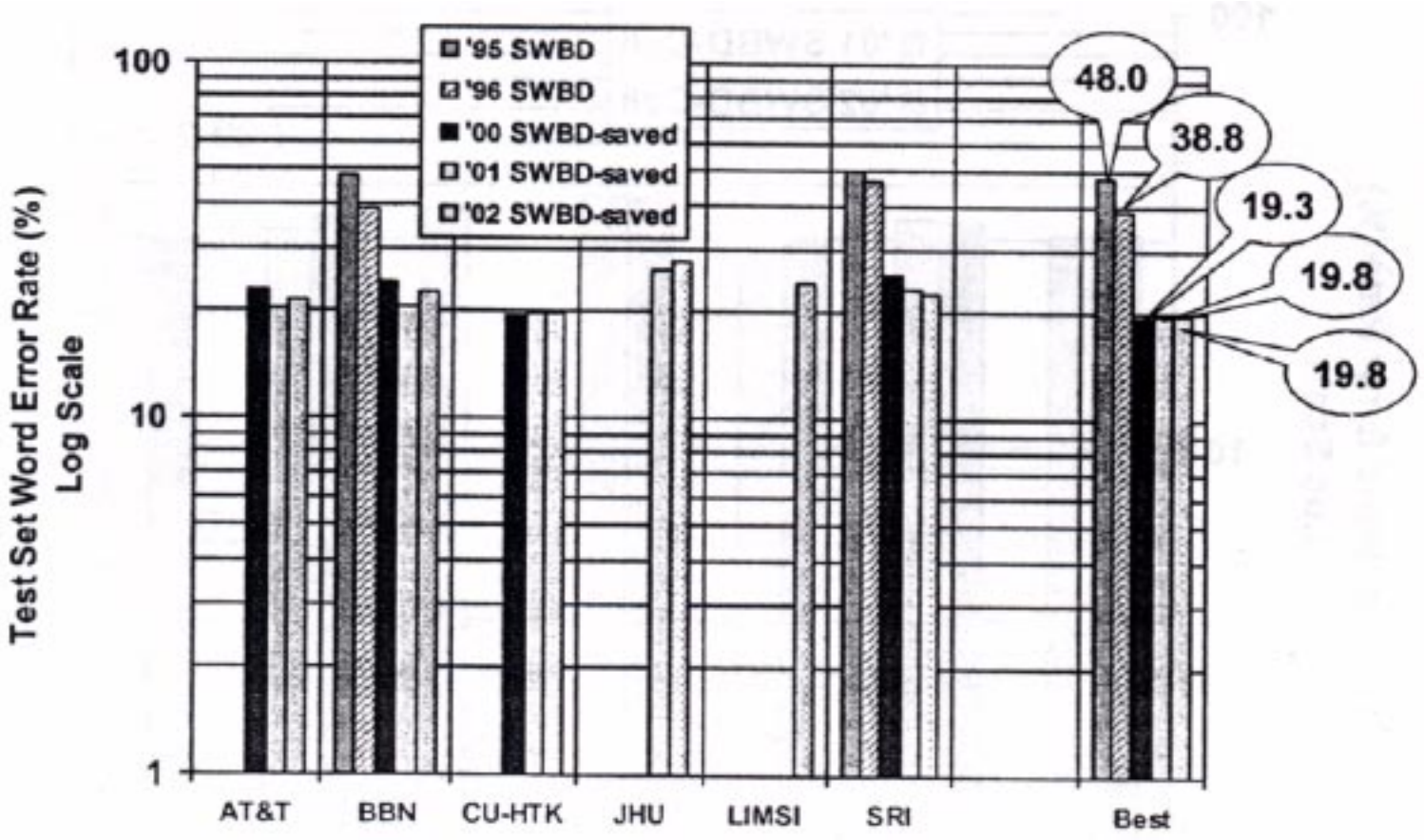


STATE OF THE ART

PERFORMANCE HISTORY

intervoice

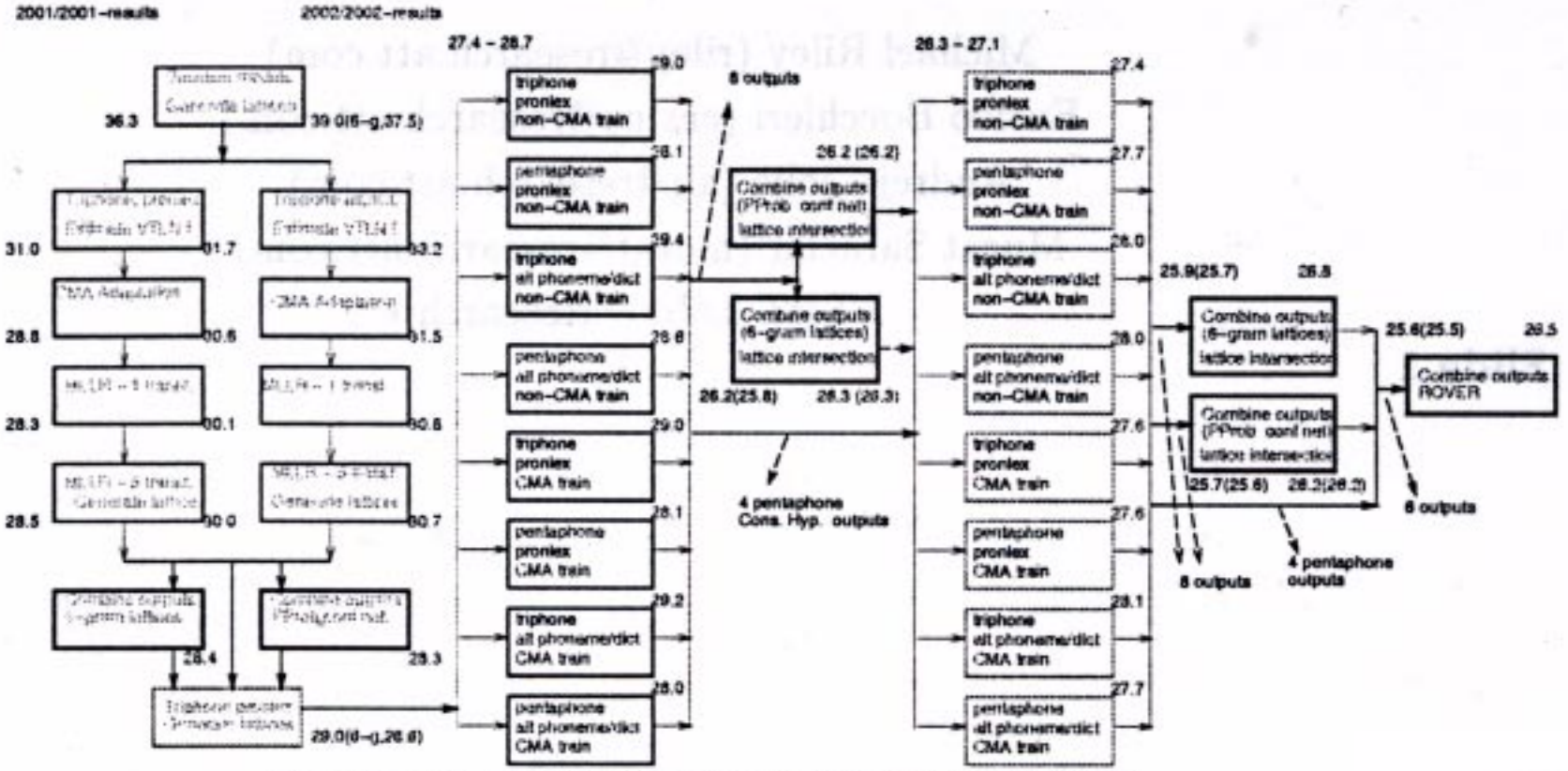
- Error rates on research systems have dropped 50% in 7 years:



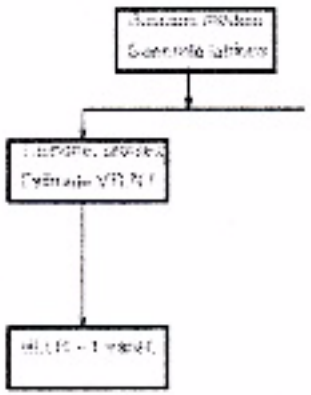
- The performance of real-time (xRT) systems is about 50% higher.
- The performance of 10xRT systems is about 25% higher.

STATE OF THE ART

ARE RESEARCH SYTEMS PRACTICAL?



- The 2001 Hub 5E AT&T system is shown above.
- A real-time version of the same system, developed for the “How may I help you” application, is shown to the right.
- Moral: Research results don’t often translate to real systems.



STATE OF THE ART

PERFORMANCE VS. TASK

intervoice

Performance as a function of task:

Corpus	Vocabulary Size	Perplexity	WER
TI Digits (TIDigits)	11	11	~0%
OGI Alphadigits (AD)	36	36	8%
Resource Management (RM)	1,000	60	4%
Air Travel Information Service (ATIS)	1,800	12	4%
Wall Street Journal (WSJ)	20,000	200 - 250	15%
Broadcast News (BN, Hub 4)	> 80,000	200 - 250	18%
Conversational Speech (SWB, Hub 5)	> 50,000	100 - 150	20%

- WER is proportional to perplexity:

$$WER \approx -12.37 + 6.48 \bullet \log_2(Perplexity)$$

- Acoustic confusability of highly probable and interchangeable words most often dominates performance.

CONCLUSIONS

COMMODITY OR LIABILITY?

intervoice

- Commercial speech recognition systems are based on hidden Markov Model technology and include context-dependent cross-word phonetic models and bigram/trigram language models.
- Research systems use a multipass decoding strategy and often combine outputs from each of these passes (e.g., ROVER). Many of these system perform 50 to 100 passes on the data before the final recognition result is achieved.
- Such systems typically run 100 to 500 times real-time on a 1 GHz processor and use at least 0.5 Gbytes of memory.
- Real-time systems often deliver performance close to these research systems (no more than 25% higher word error rate).
- Robustness to noise is becoming increasingly important, particularly in the automotive and cellular telephony markets.

CONCLUSIONS

REFERENCES AND RESOURCES

On-line Speech Recognition Resources:

[1] "Internet-Accessible Speech Recognition Technology," *http://www.isip.msstate.edu/projects/speech/index.html,* Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, January 2002.

[2] "Speech Recognition System Training Workshop," *http://www.isip.msstate.edu/conferences/srstw/current/index.html,* Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, May 2002.

[3] "Fundamentals of Speech Recognition — A Tutorial Based on a Public Domain C++ Toolkit," *http://www.isip.msstate.edu/projects/speech/software/tutorials/production/fundamentals/current/*, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, May 2002.

[4] "Speech and Signal Processing Demonstrations," *http://www.isip.msstate.edu/projects/speech/software/demonstrations/index.html,* Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, January 2002.

[5] "Fundamentals of Speech Recognition," *http://www.isip.msstate.edu/publications/courses/ece_8463/,* Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, January 2001.

[6] "About our Software," *http://www.isip.msstate.edu/projects/speech/software/*," Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, November 2002.

Reading Material:

[7] X. Huang, A. Acero, and H.W. Hon, *Spoken Language Processing - A Guide to Theory, Algorithm, and System Development*, Prentice Hall, Upper Saddle River, New Jersey, USA, ISBN: 0-13-022616-5, 2001.

[8] J. Deller, et. al., *Discrete-Time Processing of Speech Signals*, MacMillan Publishing Co., ISBN: 0-7803-5386-2, 2000.

[9] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Boston, Massachusetts, USA, ISBN: 0-262-10066-5, 1998.

[10] J. Picone, "Signal Modeling Techniques in Speech Recognition," *IEEE Proceedings*, vol. 81, no. 9, pp. 1215-1247, September 1993.

[11] N. Deshmukh, A. Ganapathiraju and J. Picone, "Hierarchical Search for Large Vocabulary Conversational Speech Recognition," *IEEE Signal Processing Magazine*, vol. 16, no. 5, pp. 84-107, September 1999.

[12] S. Young, "Talking to Machines (Statistically Speaking)," *International Conference on Spoken Language Processing*, pp. 9-16, Denver, Colorado, USA, September 2002.

[13] "Benchmark Tests," *http://www.nist.gov/speech/tests/index.htm*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, November 2002.

[14] "Publications," *http://www.nist.gov/speech/publications/index.htm*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, November 2002.

CONCLUSIONS

BIOGRAPHY

intervoice

Dr. Picone is currently a Professor and Eminent Scholar in the Department of Electrical and Computer Engineering at Mississippi State University. He also serves as the Director of the Institute for Signal and Information Processing (ISIP). He founded ISIP in 1994 with a vision to develop and disseminate public domain speech recognition software. ISIP is now known worldwide as a leading provider of speech recognition software and training.

Dr. Picone received his Ph.D. from Illinois Institute of Technology in 1983. He has published over 120 papers on speech processing. He holds 8 patents, is a Senior Member of the IEEE, and is very active in the IEEE and related professional organizations He has previously worked at Texas Instruments where he was involved in research on speech recognition and compression for military and commercial applications, as well as AT&T Bell Laboratories, where he was involved in similar areas of research.