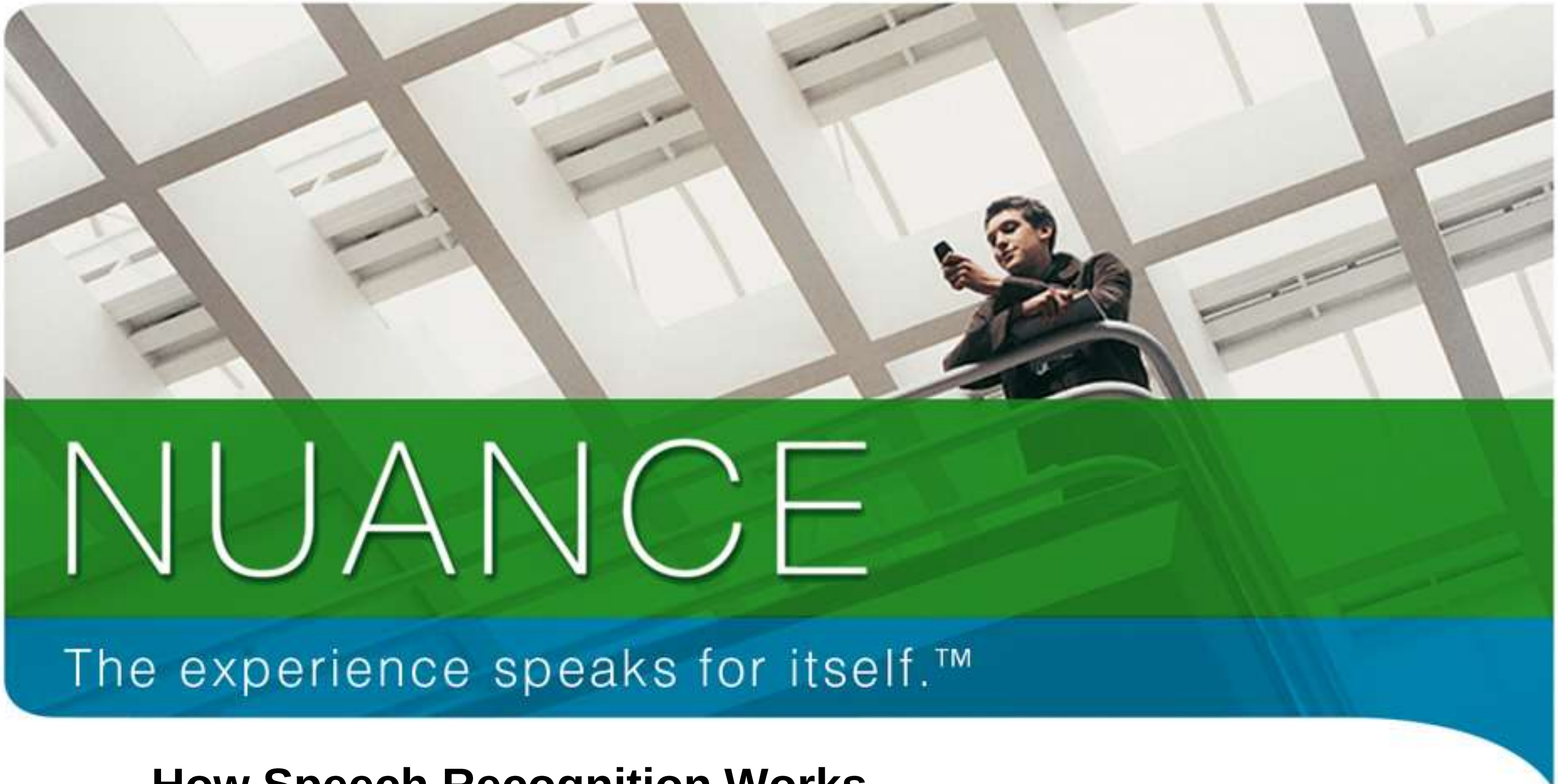


How Speech Recognition Works

Francis Ganong
Senior Technical Advisor

Nuance

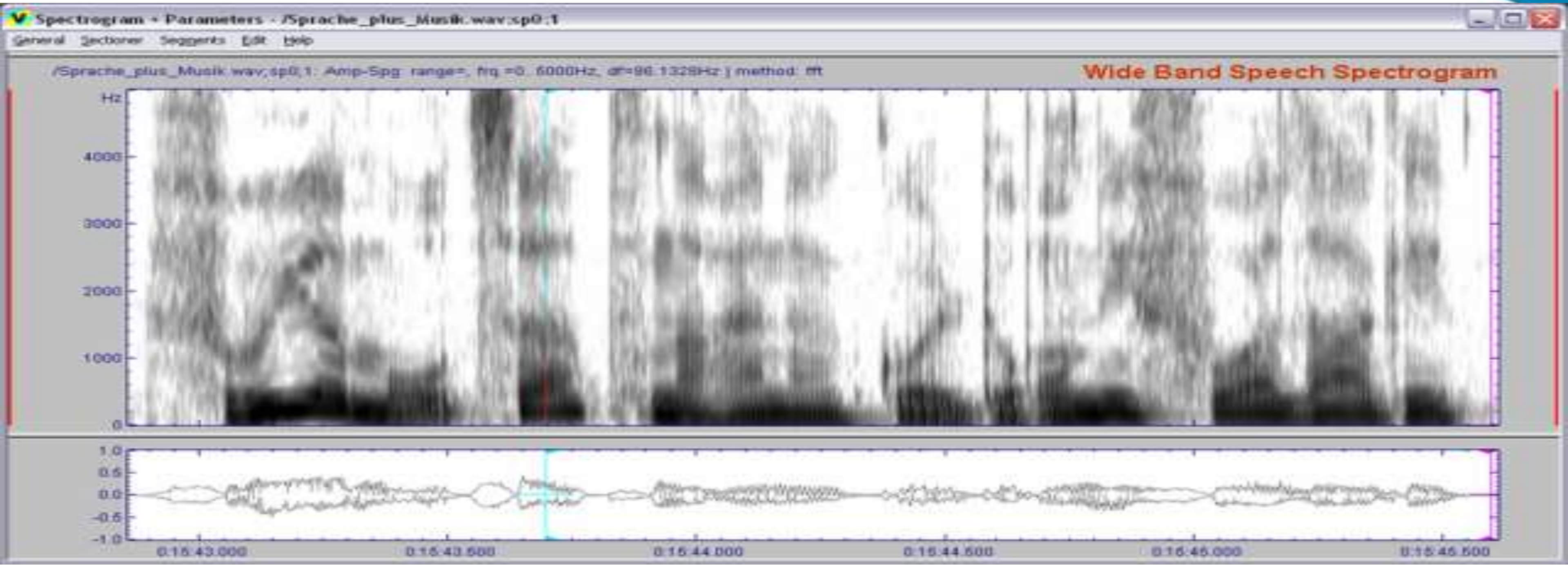
NUANCE®

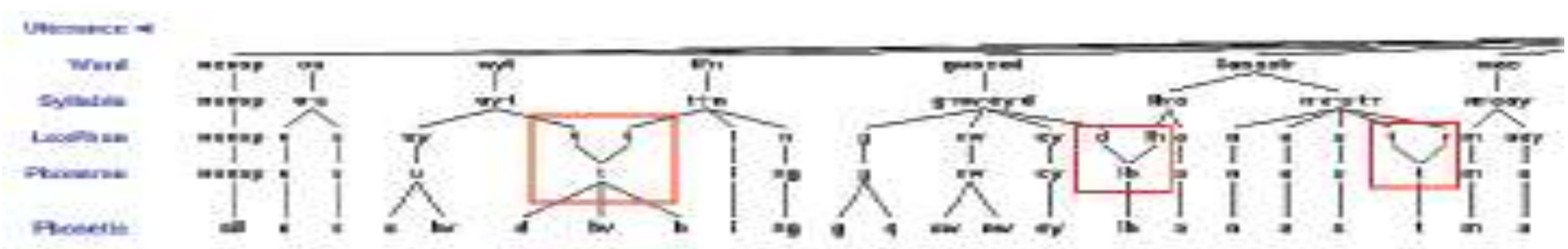
Setting

- Goal: Describe how speech recognition works, and what it's good for
- Speech Recognition
  - Fundamental equation of speech recognition
  - Modeling
  - Search
  - Are we there yet?
  - Beyond speech recognition: recognizing intent
- Industry
  - Nuance's products
- Summary, Conclusions, Questions
- Caveats:
  - Review of state of the art--Nothing said here about underlying technology necessarily applies to products of Nuance Communications, Inc.
  - This is my personal opinion and my not reflect the views of Nuance Communications, Inc.

NUANCE

Psycholinguistic Reality: Multi-Tiered Structure and Segments





A good prosodic description should go here!

Speech Recognition

- The problem in speech recognition:
  - Given an acoustic observation[1]:
    - What is the most likely sequence of words[2,3] to explain this input
      - Using
        - » Acoustic Model
        - » Language Model
  - Two problems:
    - How to score hypotheses (Modeling)
    - How to pick hypotheses to score (Search)

NUANCE

Fundamental Equations of Speech Recognition

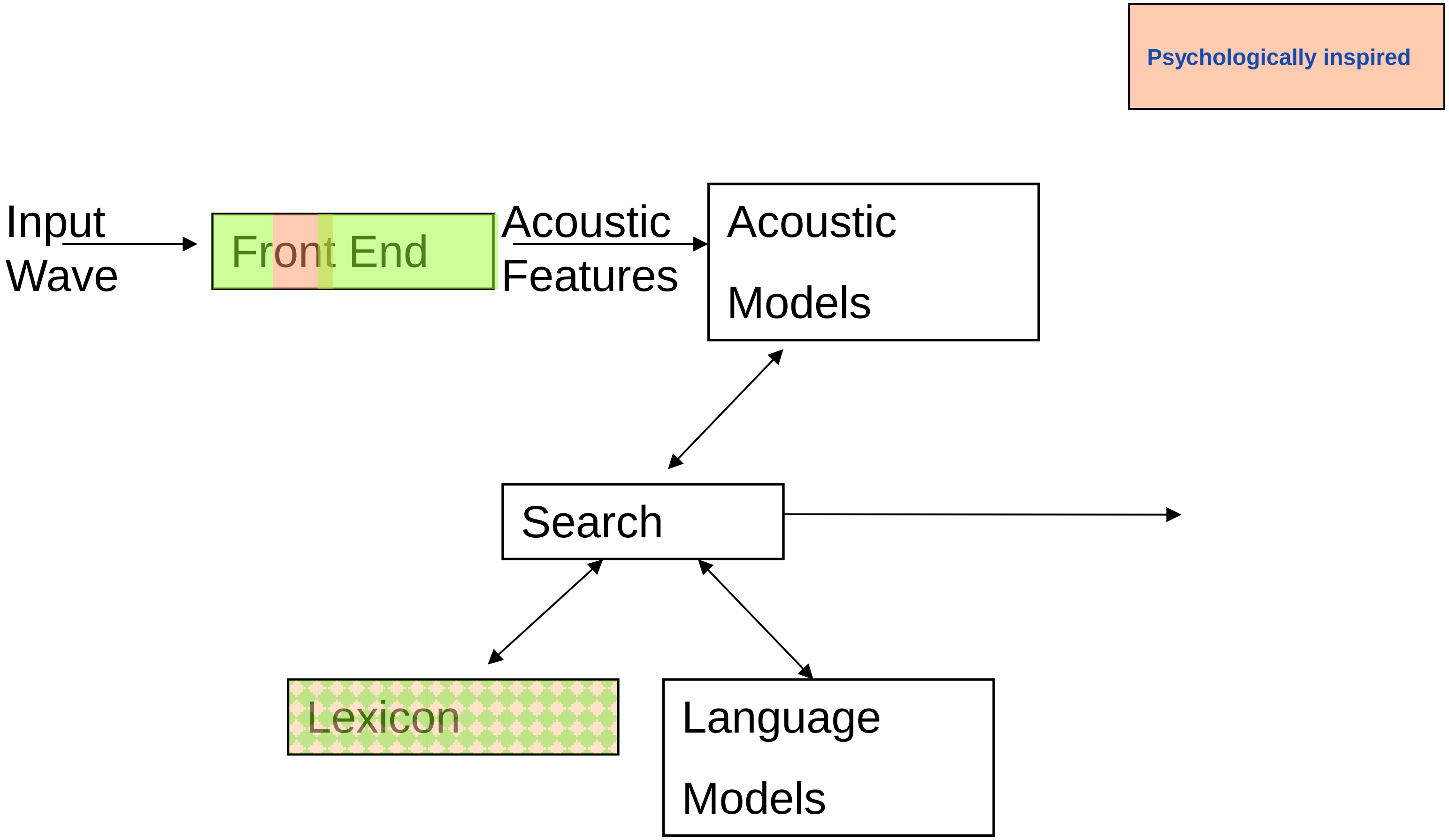
$$\tilde{W} = \arg\max_{W} (P(W \mid \bar{A}, LM))$$

$$P(W \mid \bar{A}, LM) = \frac{P(\bar{A}, \mid W) P_{LM}(W)}{P(\bar{A})}$$

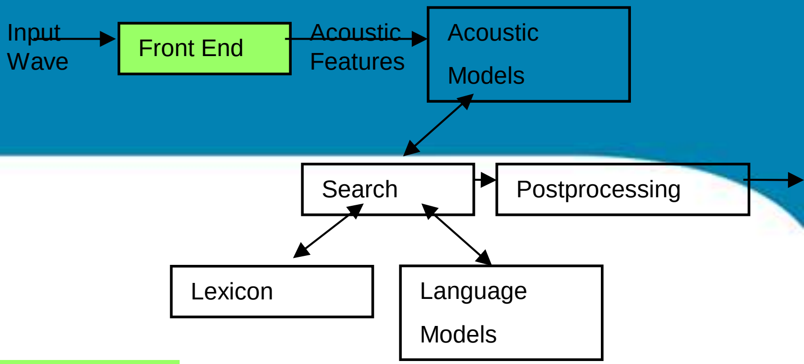
$$P(W) = \prod_{i} P(w_i \mid w_{i-1}, w_{i-2}, w_{i-3}, \ldots, w_2, w_1)$$

NUANCE

Basic Continuous Speech Recognition



Front End: MFCC



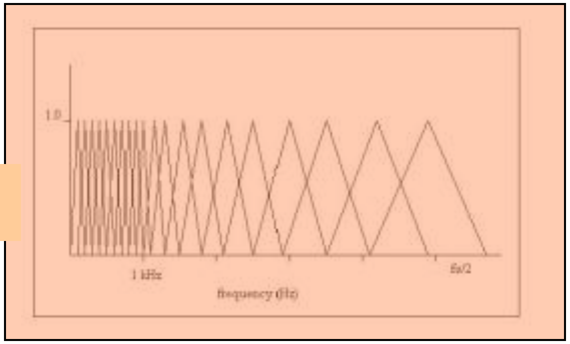
Input Wave → Sampling, Windowing

↓

FastFourierTransform

Mel Filter Bank:

cosine transform
first 8-12 coefficients



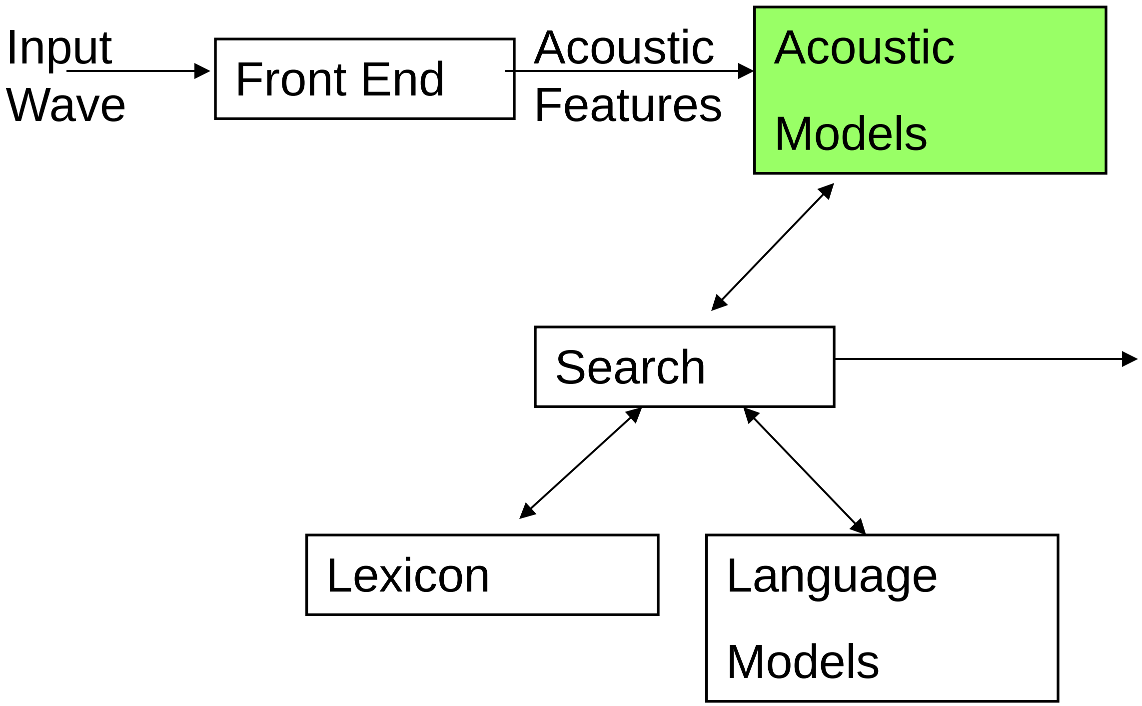
↓

Stacking, computation of deltas:Normalizations: filtering, etc

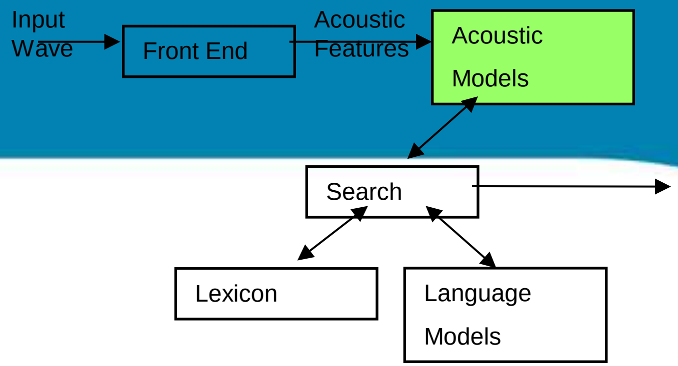
↓

Linear Transformations:dimensionality reduction → Acoustic Features

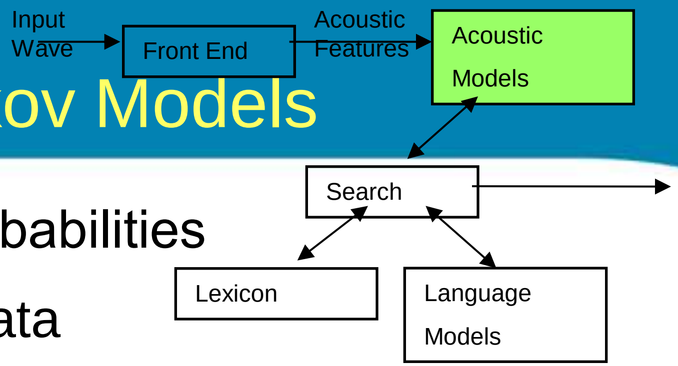
Basic Acoustic Modeling



Basic Acoustic Modeling: Hidden Markov Models

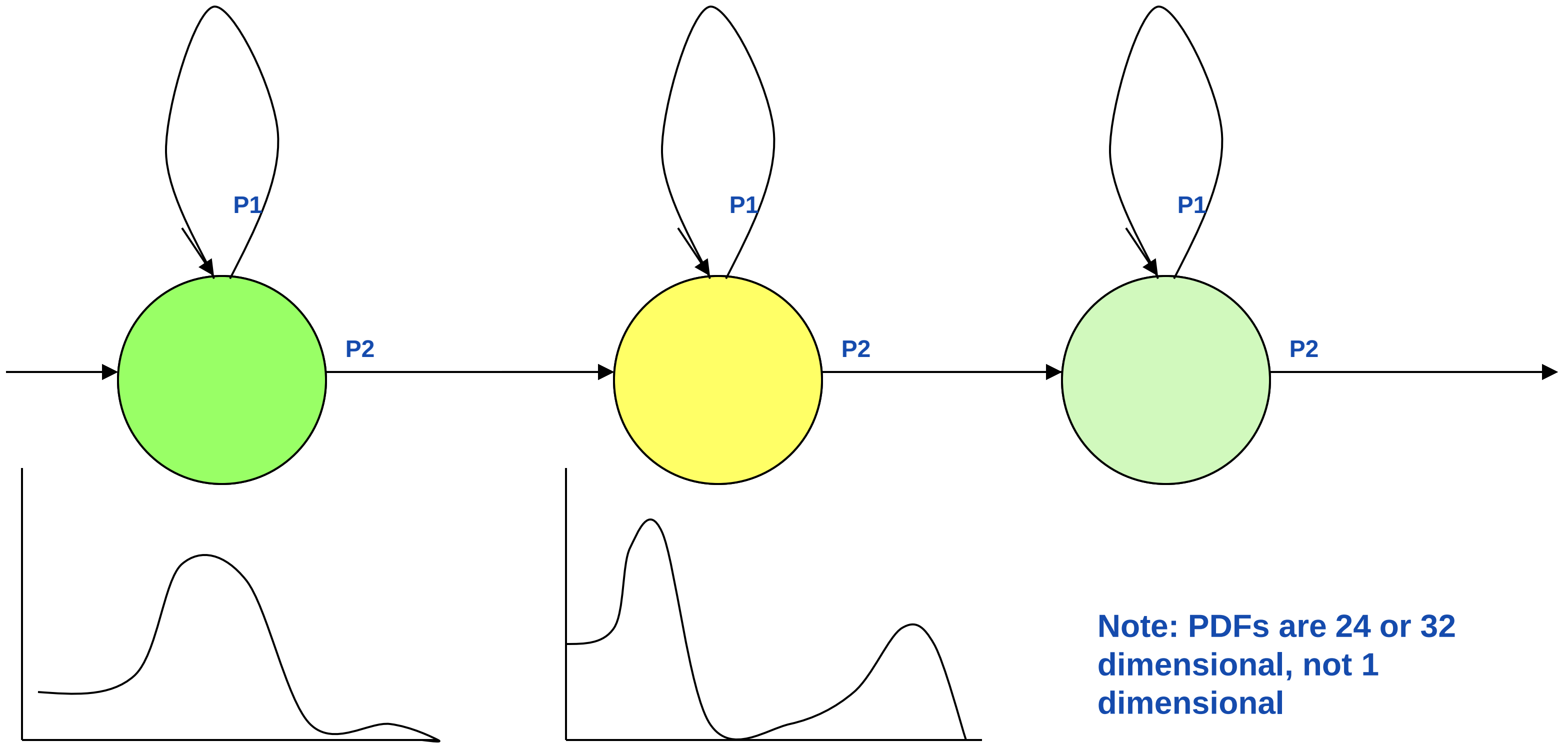


- Hidden **Markov Models**: FiniteStateMachine = : states, transitions, probabilities, and output symbols
- **Hidden** Markov Models: probabilities learned from data
- **Continuous Density** HMM: output symbols -> output vectors, (with probability densities)



Continuous Density Hidden Markov Models

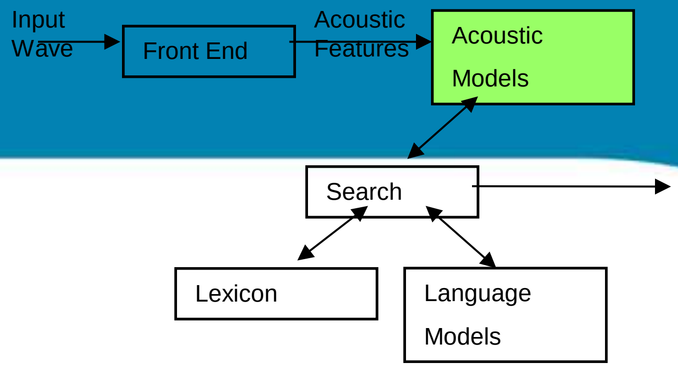
- Finite State Machines, with “hidden” probabilities
- Probabilities determined from training data
- Output: vectors, not symbols, with Probability Density Function



Note: PDFs are 24 or 32 dimensional, not 1 dimensional

NUANCE

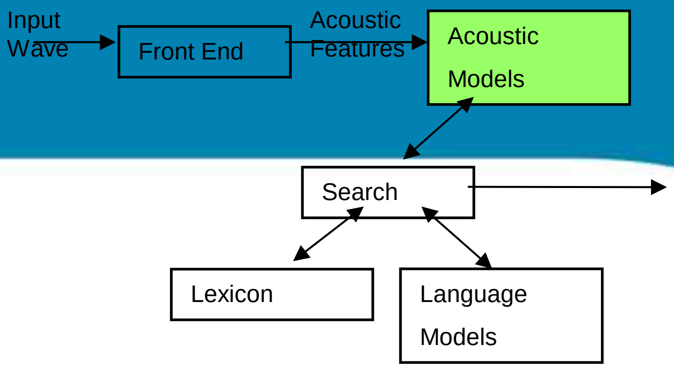
PDFs in HMMs: Gaussian Mixture Models, with variants



- N dimensional
- Gaussian Mixture Models—sum of N dimensional hills
- “genones”: pool of Gaussians, different weights.
- Variance structure
  - diagonal covariance (ellipsoids)
  - Full covariance (take into account correlations among variables)

NUANCE

Basic Acoustic Modeling



Standard Acoustic Modeling

- Words are sequences of (context dependent) phonemes
  - Phonetic context decision trees
    - Triphones: phonemes with left and right context
    - Clustered
  - "Simple" 3 state HMMs: statistical Gaussian mixture models
    - Hundreds of thousands of parameters

NUANCE

Training Acoustic Models

Input: lots of labeled data

- Hundreds of hours of speech, with word sequences

Output:

- LDA
- Phonetic context tree
- Gaussian Mixtures

Main step: segment data (into phones, or context dependent phones, then train Gaussian mixture models for each phone.

Bootstrap Recipe:

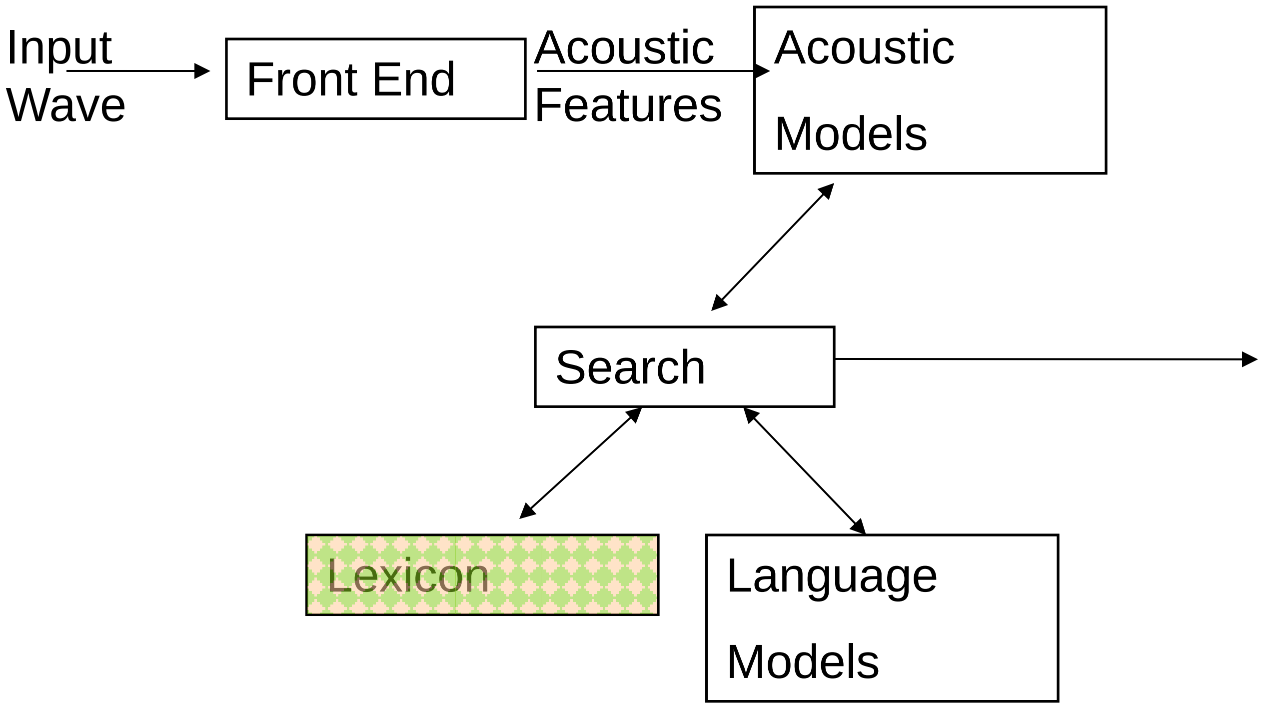
    - Flat start for phonemes
    - Cluster to make phonetic tree
    - Segment
    - Train models
- Redo

Recent extension: "lightly supervised" training

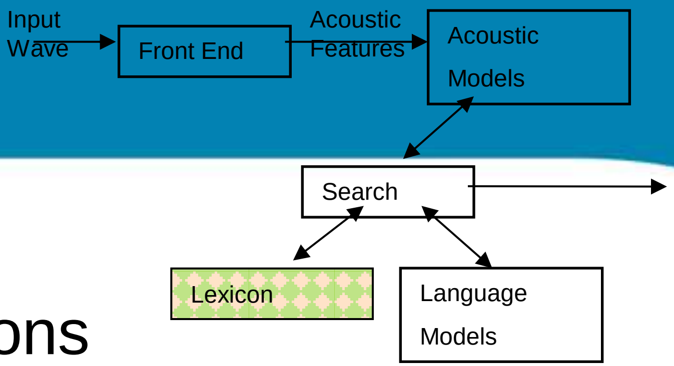
NUANCE

Lexicon

Psychologically inspired

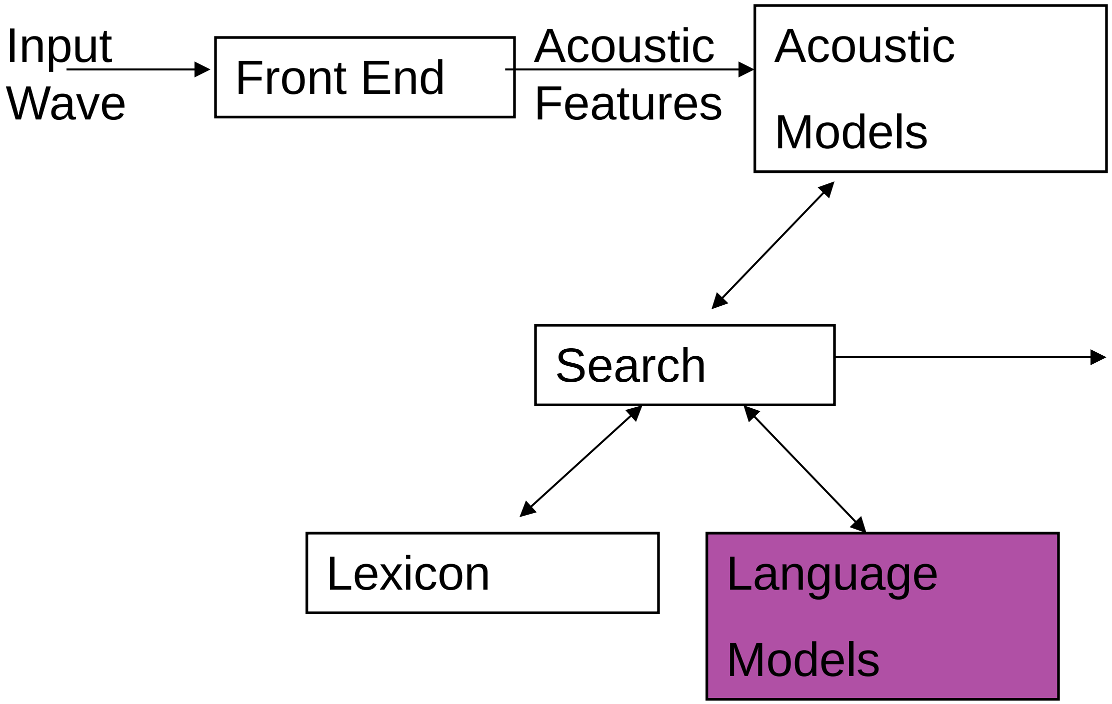


Basic Lexicon



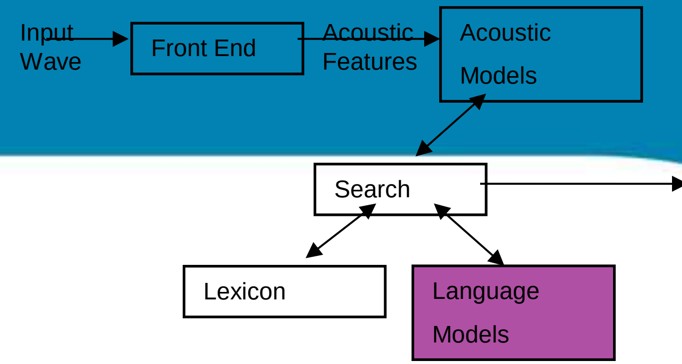
- A list of spellings and pronunciations
  - Canonical pronunciations
  - And a few others
  - Limited to 64k entries
  - Support simple stems and suffixes
- Linguistically extremely naïve
  - No phonological rewrites
  - Doesn’t support all languages
- Large injection of linguistic information into the system.

Basic speech recognizer—Language Model



∈

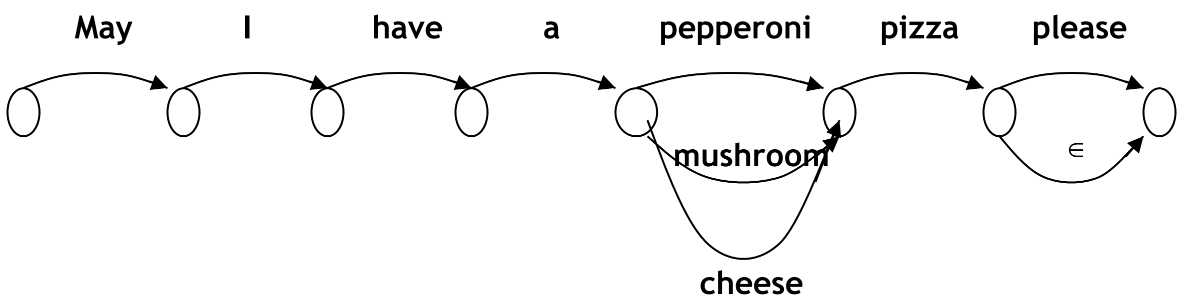
Grammars



Constrains what the user can say

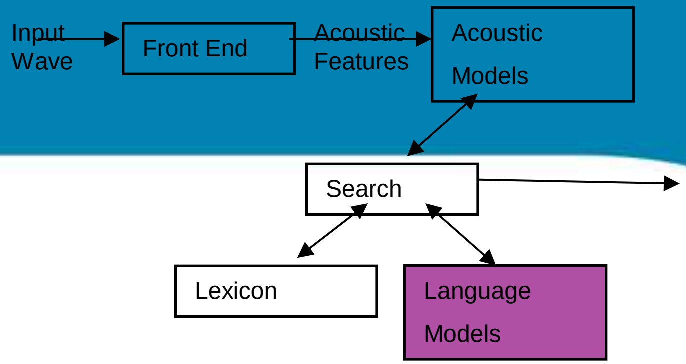
- Reduces search space substantially
- Related to interpretation problem
- Usually expressed as CFG or FSM

One version: FSM



NUANCE

Language Modeling

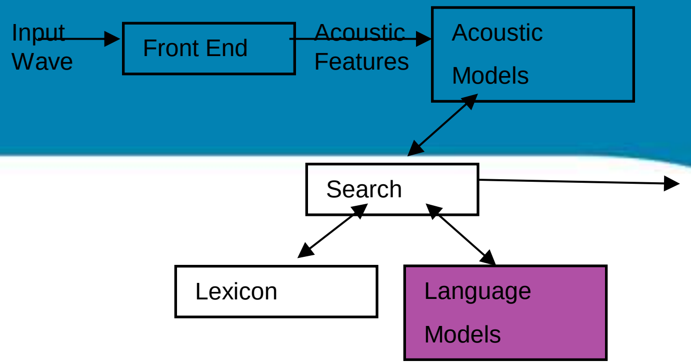


- Application Grammars
  - A ***grammar*** is a set of ***words*** together with ***constraints*** specifying how words are combined to form valid sentences.
  - Some word sequences are in a grammar:
    - “I’d like to order a pepperoni pizza please”
  - Some are not
    - “please pizza pepperoni a order to like I’d”
    - “Do you fly to Jacksonville?”

- SLMs: Stochastic LMs
  - Rather than enumerating legal sequences, any sequence is legal, but some more likely than others.
  - Ngrams
    - Word NGrams
    - POS Ngrams
    - really big LMs, including the web

NUANCE

Ngram sLMs



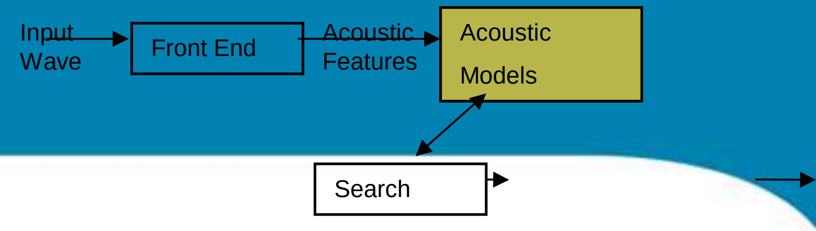
- Trigrams: p(W3 | W1 W2)

  e.g. p("rose" | "stock prices") vs.

  p("prose" | "stock prices")

- Very simple model for language, which we know is "wrong" (because it ignores all previous words, except the last two)
- As an engineering approximation for speech recognition, has been very hard to beat.
- sparse data problem requires back-off strategy:
  - many trigrams in testing data are new

Language Learning: About Data



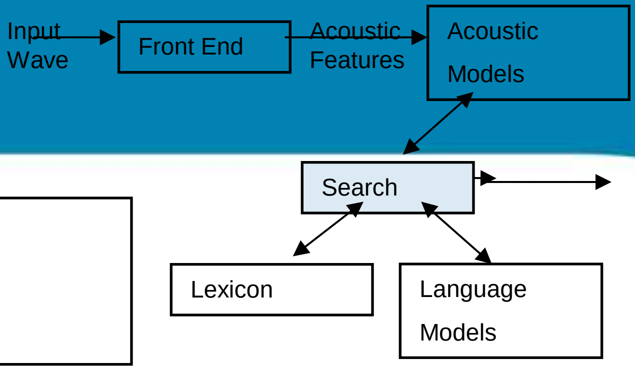
"There is no data like more data" (Mercer at Arden House, 1985)

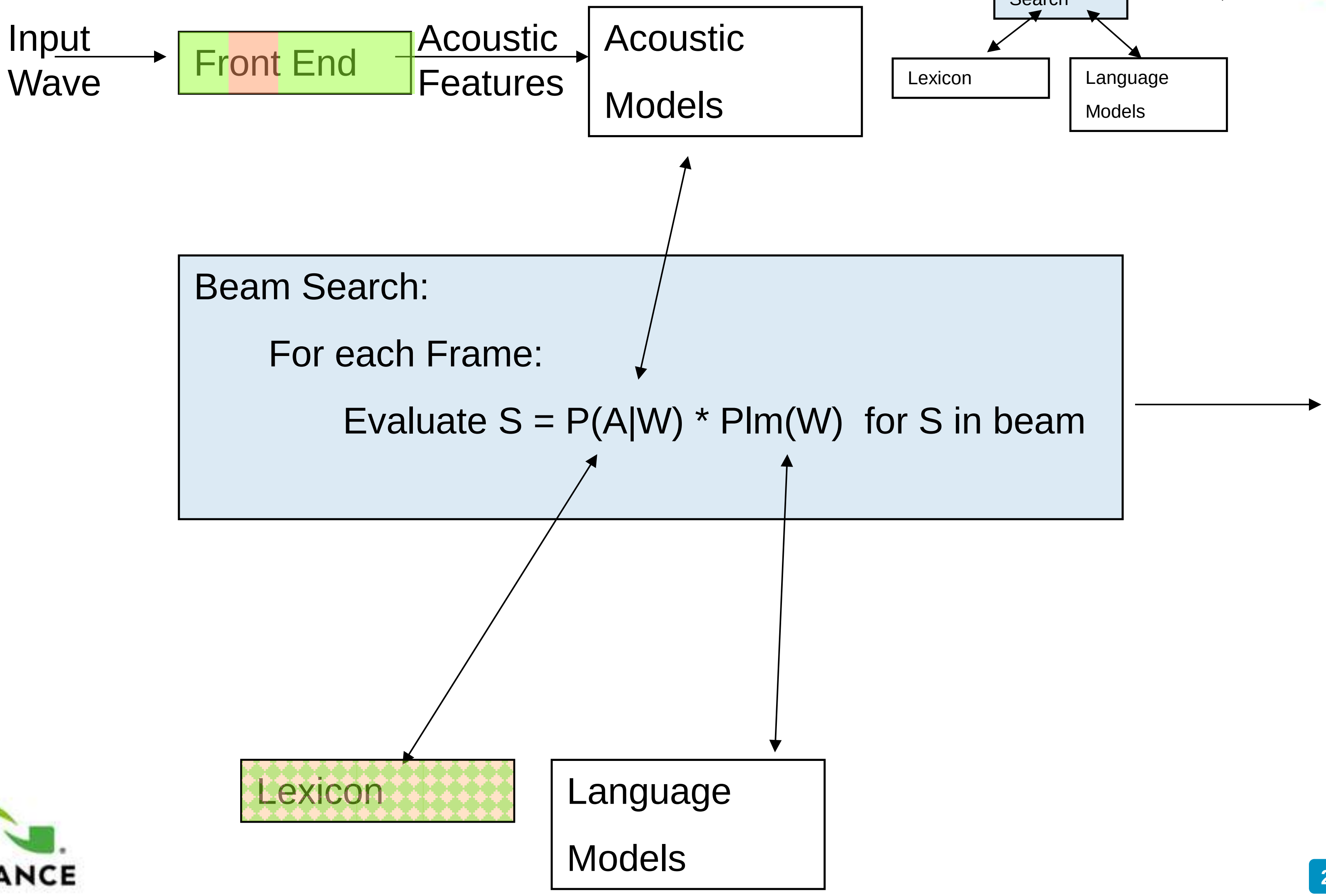
"More data is more important than better algorithms" (Brill's opinion)

from: Some of my Best Friends are Linguists (LREC 2004)
Frederick Jelinek, Johns Hopkins University
http://www.lrec-conf.org/lrec2004/doc/jelinek.pdf

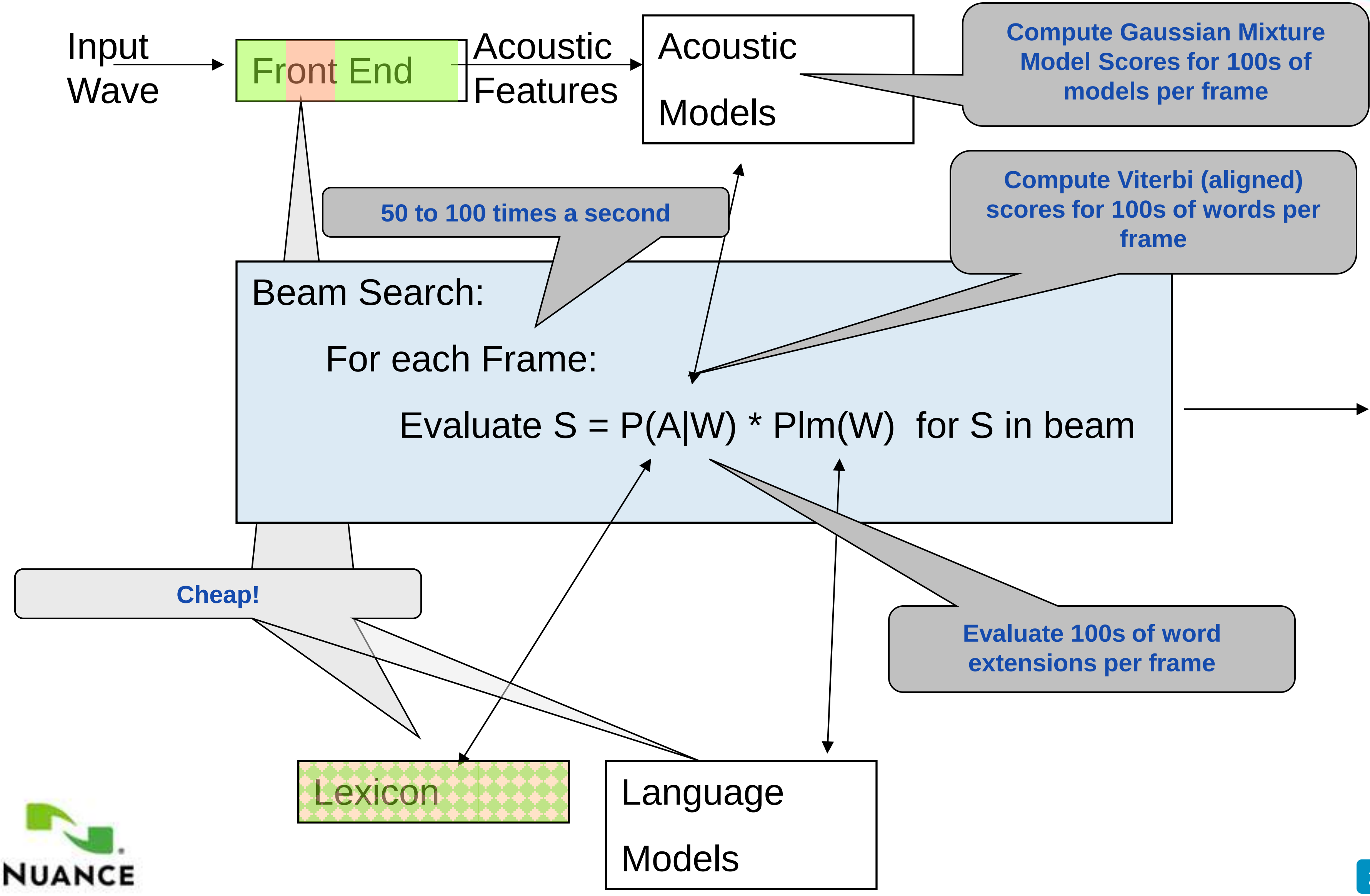
NUANCE

Basic Speech Recognition, 1990: Search





Solving crosswords at 10,000 Words Per Second



NUANCE

Speech Recognition: Basic -> 2010

$$\Delta \rightarrow \Delta\Delta \rightarrow StackedFrames$$

- 3->5:
  - Triphones -> Quinphones
  - Trigrams -> Pentagrams
- Bigger acoustic models:
  - More parameters
  - More mixtures
- Bigger lexicons
  - 65k -> 256k
- Bigger language models
  - More data
  - More parameters

NUANCE

Speech Recognition: Recent improvements

- Deep belief nets
  - Largest improvement in several years
    - Augments (or replaces) acoustic model
  - Breakthrough realization:
    - Geoff Hinton, U. Toronto
    - “loading” with unsupervised data
    - Evaluates large, deep neural nets
    - Efficient computation with GPUs
- Training from lots and lots of lightly labeled data

NUANCE

Speech recognition observations

- Total victory of data-based, statistical approaches
- Standard statistical problems
  - Curse of dimensionality, Long tails
  - Desirability of Priors
- Quite sophisticated statistical models
  - Advances due increased size and sophistication of models
- Similar to Moore's law: no breakthroughs, dozens of small incremental advances
  - Substantial continuous improvement in recognition technology
- Tiny impact of linguistic theories
    - except naïve theory: speech is an unconstrained sequence of words, words are sequences of phonemes

NUANCE

Are we there yet?

- **Whatever Happened to Voice Recognition?**
- **June 21, 2010   (Jeff Atwood)**
- Remember that Scene in Star Trek IV where Scotty tried to use a Mac Plus?

NUANCE

Are we there yet?

Can speech recognition match human performance?

- Dictation performance
  - Some speakers (David Pogue) < 1%
  - Many radiologists, <2%
  - Most people aren’t David Pogue or Radiologists!

- Performance for 2 speakers, single channel, simple grammar
  - Human: 22.3% Error rate; IBM: 21.6

People still better than machines in noise:

- audio CAPTCHAs
  - present a digit sequence in noise

Multi microphones will allow speech recognition to exceed

Beyond speech recognition: intents

- When you use a computer, how often is it write a text?
- (v.s. how often do you want to do something?
  - Listen to some music
  - Check a fact (is my flight on time?)
  - Fill out a form

- Call centers, IVR
  - Can we recognize what the user wants:
    - “What can I do for you”? speak freely for call routing
- SIRI:
  - Carrying out actions

NUANCE

How to map sentences to intents?

- Use similar techniques as speech recognition:
  - Labeled data
  - Learn associations between words and intents
- Similar problems:
  - Expensive to get labeled data
- Want more than just “what is this about”, but “exactly what do you want?

Work very much in progress.

NUANCE

Speech Recognition in use

We're a company most people don't realize they already have a relationship with

AIRFRANCE / SKYTEAM
xerox
amazon.com
Cerner
intuit
LG
Time Warner Cable
USAir
SAMSUNG
Canon
BARCLAYS
UPMC
University of Pittsburgh Medical Center
IBM
Disney
Deutsche Bank
Australian Government
Department of Broadband, Communications and the Digital Economy
htc
quietly brilliant
hp
Vonage
E*TRADE
WELLS FARGO
DELTA
Epic
citi
intel
BMW
USAA

NUANCE

Nuance Products

- Dragon
  - Dragon Naturally Speaking
  - Dictate (Mac)
- Healthcare
  - eScription: machine assisted medical transcription
  - PowerScribe: “Front end” dictation by radiologists. Instant reports!
- Mobile:
  - Handsets: recognition on the phone
  - Handsets: key pad software: T9, Swype
  - Automotive
- Enterprise:
  - Call Centers

NUANCE

5 billion
mobile cloud transactions daily

3,900
patents & applications

65+
countries

12 billion
customer calls served annually

12,000
employees

70+
languages

800 million
mobile keyboards shipped annually

13,000
mobile app developers

1,200
voice and language scientists and engineers

5 billion
lines of medical data transcribed annually

25 million
voice-enabled cars sold annually

NUANCE

Nuance Products

- Dragon
  - Dragon Naturally Speaking 21 million customers
  - Dictate (Mac)
- Healthcare 450,000 physicians, 10,000 organizations
  - eScription: machine assisted medical transcription
  - PowerScribe: “Front end” dictation by radiologists. Instant reports!
- Mobile:
  - Handsets: recognition on the phone 500 million devices
  - Handsets: key pad software: T9, Swype 7 billion devices
  - Automotive 70 million cars
- Enterprise:
  - Call Centers 8,000 systems, 10 billion caller interactions per year

NUANCE

And you can try it: (if you have a smart phone)

- iPhone: app Store
  - Dragon Dictation
  - Dragon Go
- Android app Store
  - Dragon Hands Free Assistant
  - Dragon Go!

- 1 billion words recognized

NUANCE

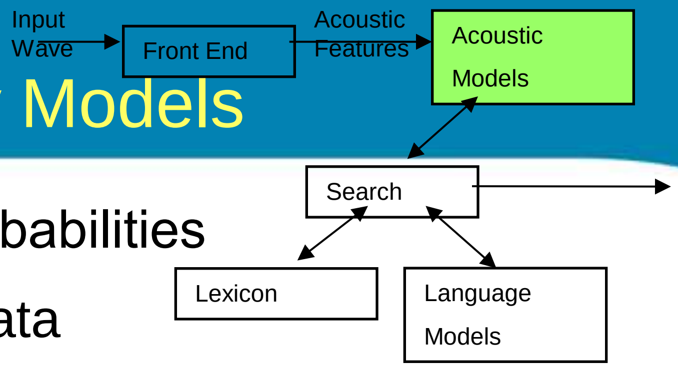
Summary, Conclusions

- Speech recognition is very hard
    - Engineering a very human capability
  - Substantial progress
    - Not as good as people. Yet.
  - Ongoing progress
    - No fundamental limits seen
- Intent recognition is harder
  - Unknown how far we can go

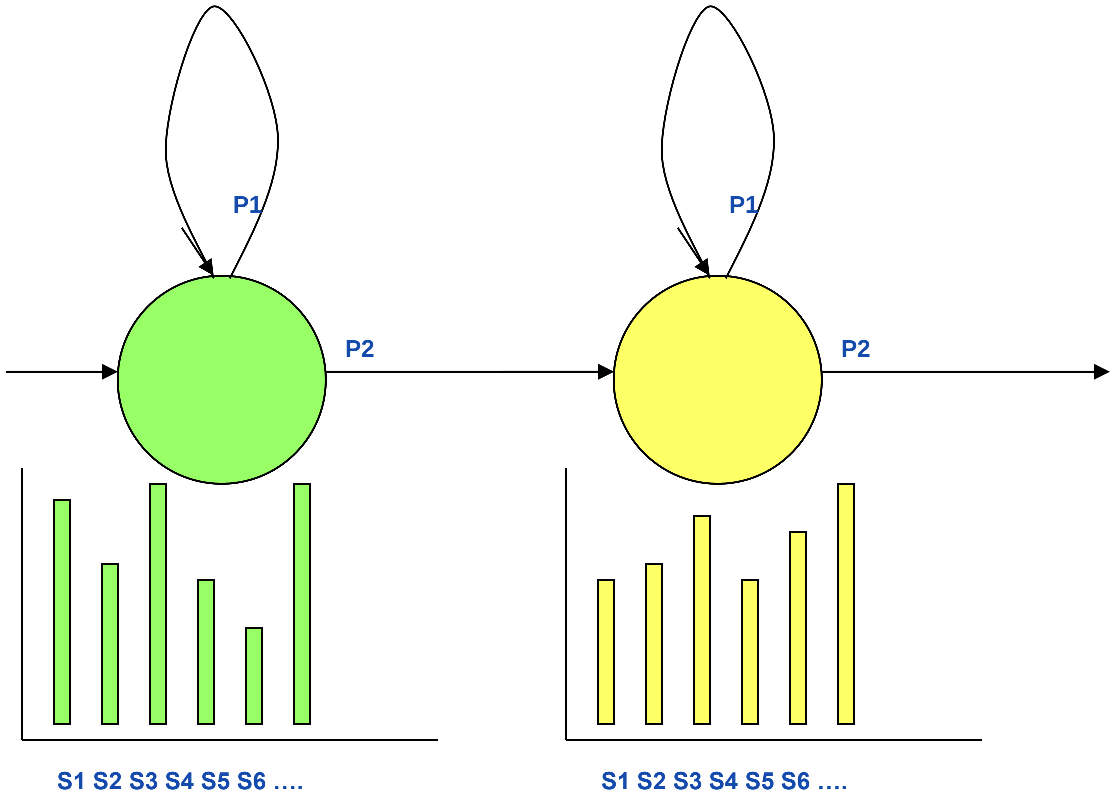
Questions?

NUANCE

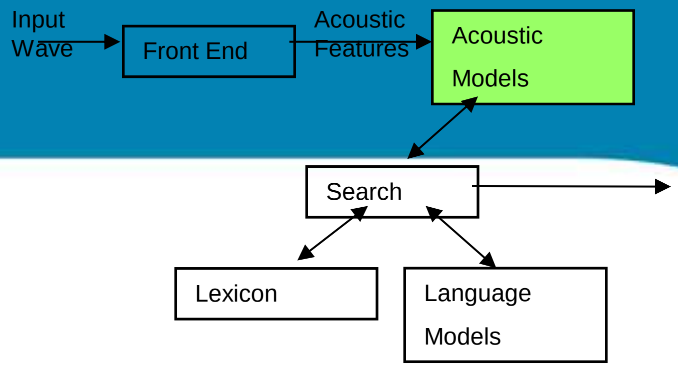
(discrete density) Hidden Markov Models



- Finite State Machines, with "hidden" probabilities
- Probabilities determined from training data
- Input: symbols, 100 symbols second: F1, F2,…



Hidden Markov Models



- Finite State Machine
- = : states, transitions, probabilities, output symbols