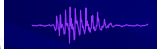
Speech Recognition Introduction I

E.M. Bakker

Speech Recognition

- Some Applications
- An Overview
- General Architecture
- Speech Production
- Speech Perception

Speech Recognition



Speech Signal

Speech Recognition

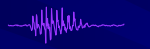
Words

"How are you?"

Goal: Automatically extract the string of words spoken from the speech signal

Speech Recognition

Goal: Automatically extract the string of words spoken from the speech signal



Speech Signal

Speech Recognition

Words

"How are you?"

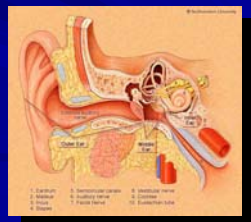
How is SPEECH produced?
⇒ Characteristics of Acoustic Signal

Speech Recognition

Goal: Automatically extract the string of words spoken from the speech signal

Speech Signal → Speech Recognition → Words "How are you?"

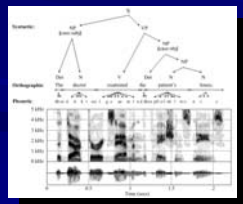
How is SPEECH perceived?
=> Important Features



Speech Recognition

Goal: Automatically extract the string of words spoken from the speech signal

Speech Signal → Speech Recognition → Words "How are you?"



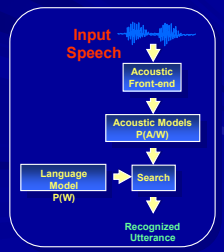
What LANGUAGE is spoken?
=> Language Model

Speech Recognition

Goal: Automatically extract the string of words spoken from the speech signal

Speech Signal → Speech Recognition → Words "How are you?"

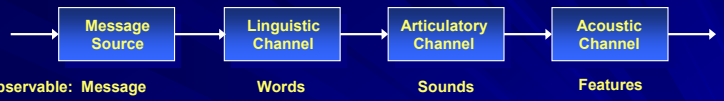
What is in the BOX?



Important Components of General SR Architecture

- Speech Signals
- Signal Processing Functions
- Parameterization
- Acoustic Modeling (Learning Phase)
- Language Modeling (Learning Phase)
- Search Algorithms and Data Structures
- Evaluation

Recognition Architectures
A Communication Theoretic Approach



Speech Recognition Problem: $P(W|A)$, where A is acoustic signal, W words spoken

Objective: minimize the word error rate
Approach: maximize $P(W|A)$ during training

Bayesian formulation for speech recognition:

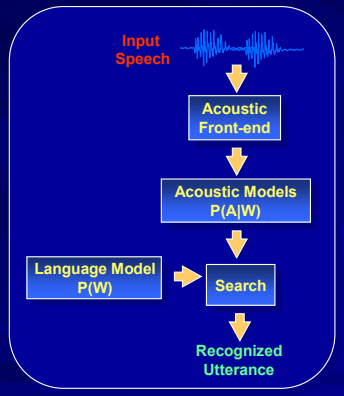
- $P(W|A) = P(A|W)\,P(W) / P(A)$, A is acoustic signal, W words spoken

Components:

- $P(A|W)$: acoustic model (hidden Markov models, mixtures)
- $P(W)$: language model (statistical, finite state networks, etc.)

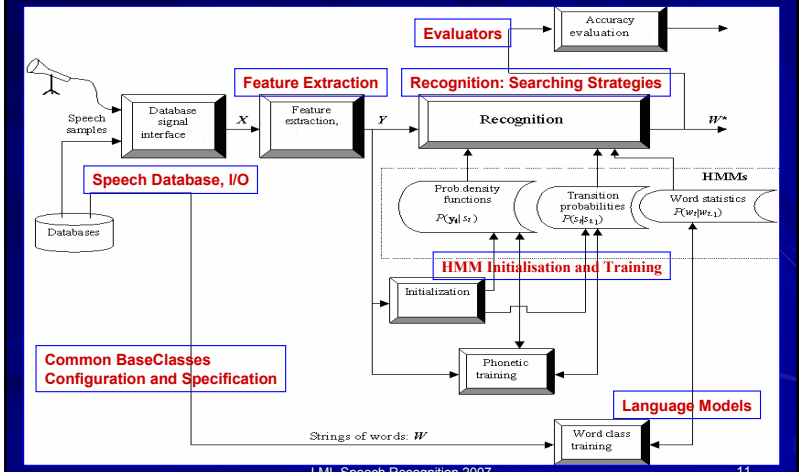
The language model typically predicts a small set of next words based on knowledge of a finite number of previous words (N-grams).

Recognition Architectures



- The signal is converted to a sequence of feature vectors based on spectral and temporal measurements.
- Acoustic models represent sub-word units, such as phonemes, as a finite-state machine in which:
  - states model spectral structure and
  - transitions model temporal structure.
- The language model predicts the next set of words, and controls which models are hypothesized.
- Search is crucial to the system, since many combinations of words must be investigated to find the most probable word sequence.

ASR Architecture



Signal Processing Functionality

- Acoustic Transducers
- Sampling and Resampling
- Temporal Analysis
- Frequency Domain Analysis
- Ceps-tral Analysis
- Linear Prediction and LP-Based Representations
- Spectral Normalization

Acoustic Modeling: Feature Extraction

Fourier Transform

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

Cepstral Analysis

- Measure features 100 times per sec.
- Use a 25 msec window for frequency domain analysis.
- Include absolute energy and 12 spectral measurements.
- Time derivatives to model spectral change.

Perceptual Weighting → Time Derivative → Time Derivative

Energy + Mel-Spaced Cepstrum

Delta Energy + Delta Cepstrum

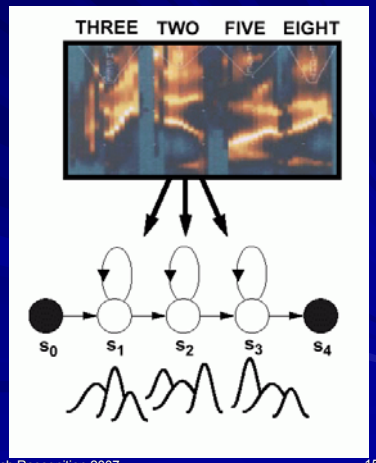
Delta-Delta Energy + Delta-Delta Cepstrum

Acoustic Modeling

- Dynamic Programming
- Markov Models
- Parameter Estimation
- HMM Training
- Continuous Mixtures
- Decision Trees
- Limitations and Practical Issues of HMM

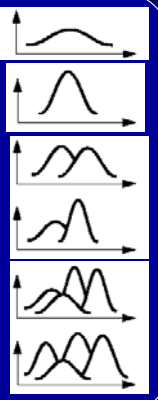
Acoustic Modeling
Hidden Markov Models

- Acoustic models encode the temporal evolution of the features (spectrum).
- Gaussian mixture distributions are used to account for variations in speaker, accent, and pronunciation.
- Phonetic model topologies are simple left-to-right structures.
- Skip states (time-warping) and multiple paths (alternate pronunciations) are also common features of models.
- Sharing model parameters is a common strategy to reduce complexity.



Acoustic Modeling: Parameter Estimation

- Initialization
- Single Gaussian Estimation
- 2-Way Split
- Mixture Distribution Reestimation
- 4-Way Split
- Reestimation ...

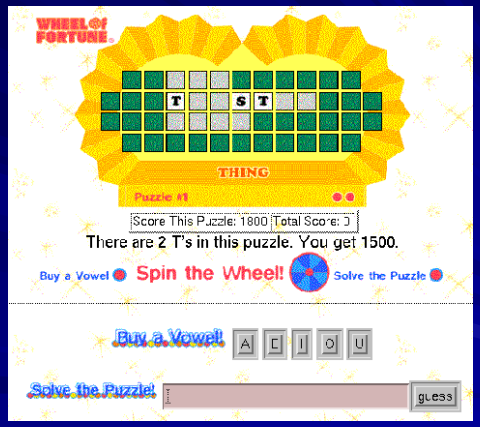


- Closed-loop data-driven modeling supervised from a word-level transcription.
- The expectation/maximization (EM) algorithm is used to improve our parameter estimates.
- Computationally efficient training algorithms (Forward-Backward) have been crucial.
- Batch mode parameter updates are typically preferred.
- Decision trees are used to optimize parameter-sharing, system complexity, and the use of additional linguistic knowledge.

Language Modeling

- Formal Language Theory
- Context-Free Grammars
- N-Gram Models and Complexity
- Smoothing

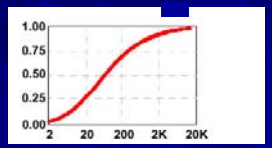
Language Modeling



Language Modeling: N-Grams

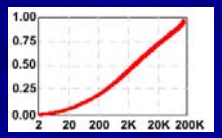
Unigrams (SWB):

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



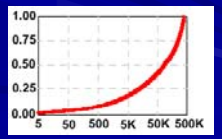
Bigrams (SWB):

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

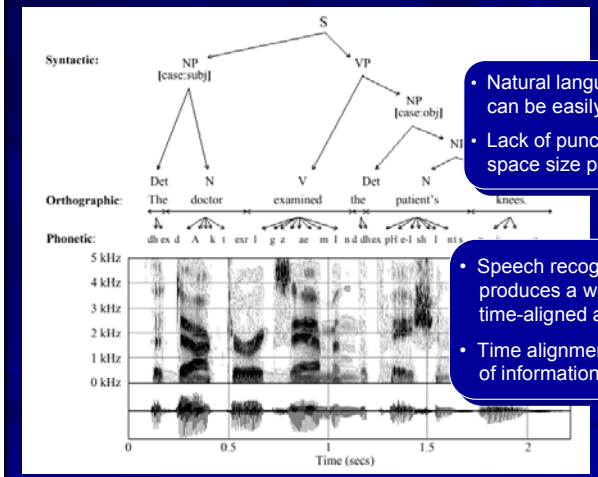


Trigrams (SWB):

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



LM: Integration of Natural Language



- Natural language constraints can be easily incorporated.
- Lack of punctuation and search space size pose problems.

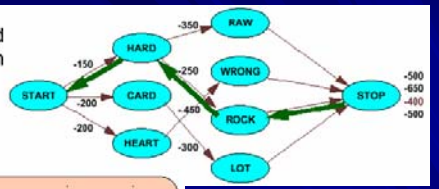
- Speech recognition typically produces a word-level time-aligned annotation.
- Time alignments for other levels of information also available.

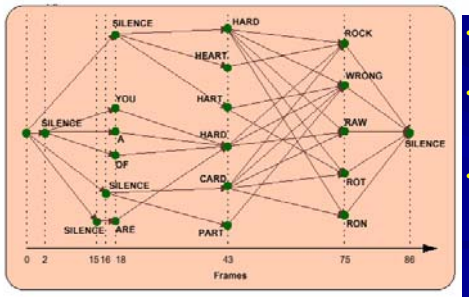
Search Algorithms and Data Structures

- Basic Search Algorithms
- Time Synchronous Search
- Stack Decoding
- Lexical Trees
- Efficient Trees

Dynamic Programming-Based Search

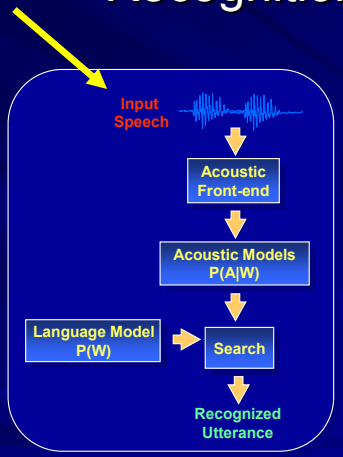
- Dynamic programming is used to find the most probable path through the network.
- Beam search is used to control resources.





- 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.

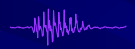
Recognition Architectures



- The signal is converted to a sequence of feature vectors based on spectral and temporal measurements.
- Acoustic models represent sub-word units, such as phonemes, as a finite-state machine in which:
  - states model spectral structure and
  - transitions model temporal structure.
- The language model predicts the next set of words, and controls which models are hypothesized.
- Search is crucial to the system, since many combinations of words must be investigated to find the most probable word sequence.

Speech Recognition

Goal: Automatically extract the string of words spoken from the speech signal



Speech Signal → Speech Recognition → Words *"How are you?"*

How is SPEECH produced?