A Brief Introduction to Automatic Speech Recognition

Jim Glass (glass@mit.edu)

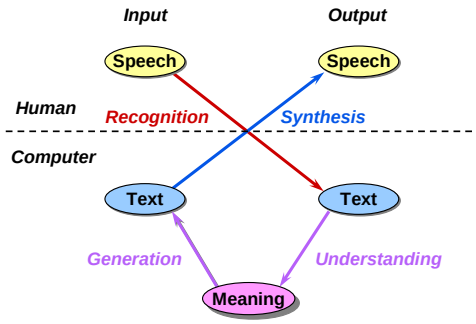
MIT Computer Science and Artificial Intelligence Laboratory

March 20, 2007

Overview

- Introduction
- Speech
- Models
- Search
- Representations

Communication via Spoken Language



Virtues of Spoken Language

Natural:	Requires no special training
Flexible:	Leaves hands and eyes free
Efficient:	Has high data rate
Economical:	Can be transmitted/received inexpensively

Speech interfaces are ideal for information access and management when:

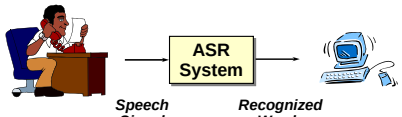
- The information space is broad and complex,
- The users are technically naive, or
- Only telephones are available.

video

Diverse Sources of Knowledge for Spoken Language Communication

Acoustic-Phonetic:	Let us pray Lettuce spray
Syntactic:	Meet her at the end of Main Street Meter at the end of Main Street
Semantic:	Is the baby crying Is the bay bee crying
Discourse Context:	It is easy to recognize speech It is easy to wreck a nice beach
Others:	I'm *flying* to Chicago tomorrow I'm flying to *Chicago* tomorrow

Automatic Speech Recognition



- An ASR system converts the speech signal into words
- The recognized words can be
  - The final output, or
  - The input to natural language processing

Application Areas for Speech Interfaces

- **Mostly input (recognition only)**
  - Simple command and control
  - Simple data entry (over the phone)
  - Dictation
- **Interactive conversation (understanding needed)**
  - Information kiosks
  - Transactional processing
  - Intelligent agents

Parameters that Characterize the Capabilities of ASR Systems

Parameters	Range
Speaking Mode:	Isolated word to continuous speech
Speaking Style:	Read speech to spontaneous speech
Enrollment:	Speaker-dependent to speaker-independent
Vocabulary:	Small (<20 words) to large (>50,000 words)
Language Model:	Finite-state to context-sensitive
Perplexity:	Low (<10) to high (>200)
SNR:	High (>30dB) to low (<10dB)
Transducer:	Noise-canceling microphone to cell phone

Read versus Spontaneous Speech

Filled and unfilled pauses:	read, spontaneous
Lengthened words:	read, spontaneous
False starts:	read, spontaneous

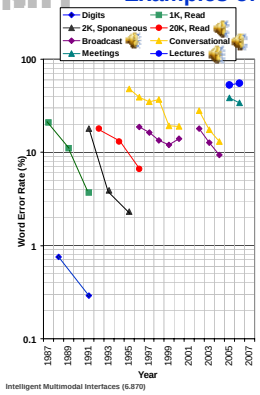
Speech Recognition: Where Are We Now?

- **High performance, speaker-independent speech recognition is now possible**
  - Large vocabulary (for cooperative speakers in benign environments)
  - Moderate vocabulary (for spontaneous speech over the phone)
- **Commercial recognition systems are now available**
  - Dictation (e.g., IBM, Microsoft, Nuance, etc.)
  - Telephone transactions (e.g., AT&T, Nuance, VST, etc.)
- **When well-matched to applications, technology is able to help perform real work**
- **Demos:**
  - Speaker-independent, medium-vocabulary, small footprint ASR
  - Dynamic vocabulary speech recognition with constrained grammar (http://web.sls.csail.mit.edu/city)
  - Academic spoken lecture transcription and retrieval (http://web.sls.csail.mit.edu/lectures)

video

video

Examples of ASR Performance

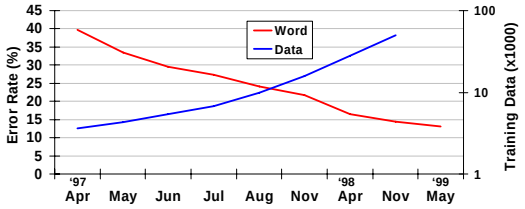


- **Telephone digit recognition has word error rates of 0.3%**
- **Error rate for spontaneous speech twice that of read speech**
- **Error rate cut in half every two years for moderate vocabularies**
- **Corpora range in size from tens to thousands of hours**
- **Conversational speech from many speakers with noise remains a research challenge**
  - Current focus on meetings & lectures

The Importance of Data

- **We need data for analysis, modeling, training, and evaluation**
  - "There is no data like more data"
- **However, we need to have the right kind of data**
  - From real users
  - Solving real problems
- **Conduct research within the context of real application domains**
  - Forces us to confront critical technical issues (e.g., rejection, new word problem)
  - Provides a rich and continuing source of useful data
  - Demonstrates the usefulness of the technology
  - Facilitates technology transfer

(Real) Data Improves Performance



- **Longitudinal evaluations show improvements**
- **Collecting real data improves performance:**
  - Enables increased complexity and improved robustness for acoustic and language models
  - Better match than laboratory recording conditions
- **Users come in all kinds**

Real Data will Dictate Technology Needs

TECHNOLOGY REQUIRED	EXAMPLE
Simple word spotting	Um, Braintree
Complex word spotting	Eh yes, Avis rent-a-car in Boston
	Hello, please Brighton, uh, can I have the number of Earthscape, in, uh, on Nonantum Street
Speech understanding	Woburn, uh, Somerville. I'm sorry

Important Lessons Learned

- **Statistical modeling and data-driven approaches have proved to be powerful**
- **Research infrastructure is crucial:**
  - Large amounts of linguistic data
  - Evaluation methodologies
- **Availability and affordability of computing power lead to shorter technology development cycles and real-time systems**
- **Performance-driven paradigm accelerates technology development**
- **Interdisciplinary collaboration produces enhanced capabilities (e.g., spoken language understanding)**

ASR Trends*: Then and Now

	before mid 70's	mid 70's - mid 80's	after mid 80's
Recognition Units:	whole-word and sub-word units	sub-word units	sub-word units
Modeling Approaches:	heuristic and ad hoc	template matching	mathematical and formal
	rule-based and declarative	deterministic and data-driven	probabilistic and data-driven
Knowledge Representation:	heterogeneous and complex	homogeneous and simple	homogeneous and simple
Knowledge Acquisition:	intense knowledge engineering	embedded in simple structure	automatic learning

*** There are, of course, many exceptions.**

But We Are Far from Done!

Corpus	Speech Type	Lexicon Size	Word Error Rate (%)	Human Error Rate (%) *
Digit Strings (phone)	spontaneous	10	0.3	0.009
Resource Management	read	1000	3.6	0.1
ATIS	spontaneous	2000	2	--
Wall Street Journal	read	~20K	6.6	1
Broadcast News	mixed	~64K	9.4	--
Switchboard (phone)	conversation	~25K	13.1	4
Meetings	conversation	~25K	30	--

* Lippmann, 1997

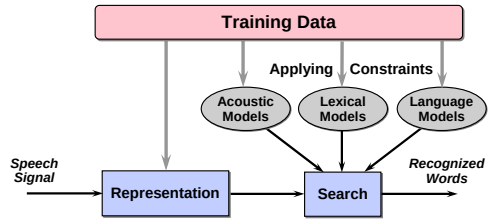
What Makes Speech Recognition Hard?

- **Phonological variations**
  - Local and global contexts, …
- **Individual differences**
  - Anatomy, socio-linguistic factors, …
- **Environmental factors**
  - Transducers, noise, …
- **Diversity of language use**
  - Syntax, semantics, discourse, …
- **Real-world issues**
  - Disfluencies, new words, …
- **. . .**

ASR Is All About Utilizing Constraints

- **Acoustic**
  - Speech signal is generated by the human vocal apparatus
- **Phonetic**
  - /s/ in word initial /st/ cluster is unaspirated (e.g. “stay”)
- **Phonological**
  - /s/-/S/ sequence can turn into a long /S/ (e.g., “gas shortage”)
- **Lexical**
  - Words in a language are limited (e.g., “blit” and “vnuk” are not English words)
- **Language**
  - Probability of a word depends on its predecessors (e.g., “you” is the most likely word to follow “thank”)
  - A sentence must be syntactically and semantically well formed (e.g., subject-verb agreement)
- **. . .**

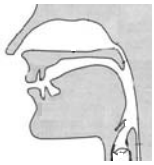
Major Components in a Speech Recognizer



- **Speech recognition is the problem of deciding on**
  - How to *represent* the signal
  - How to *model* the constraints
  - How to *search* for the most optimal answer

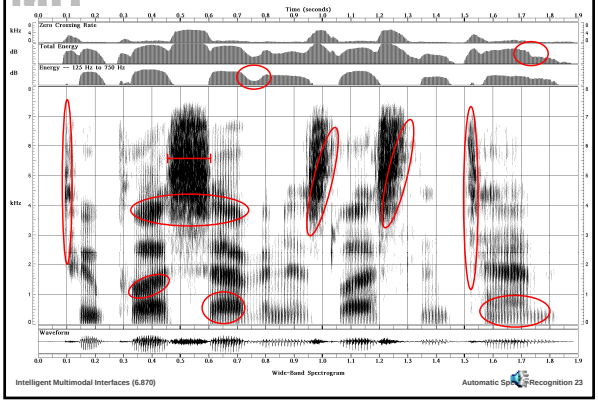
Speech

Speech Production



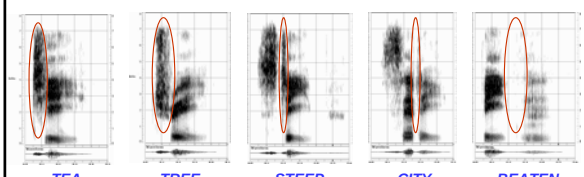
- **Speech produced via coordinated movement of articulators**
- **Spectral characteristics of speech influenced by source, vocal tract shape, and radiation characteristics**
- **Speech articulation characterized by manner and place**
  - Vowels: No significant constriction in the vocal tract; usually voiced
  - Fricatives: turbulence produced at a narrow constriction
  - Stops: complete closure in the vocal tract; pressure build up
  - Nasals: velum lowering results in airflow through nasal cavity
  - Semivowels: constriction in vocal tract, no turbulence

A Wide-Band Spectrogram



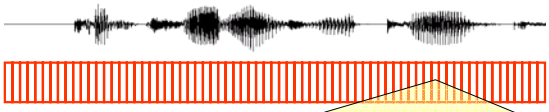
Phonological Variation

- **The acoustic realization of a phoneme depends strongly on the context in which it occurs**

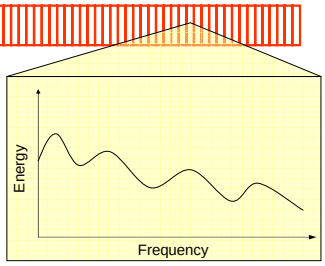


Signal Processing

Waveform

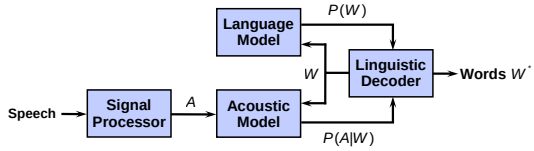


- **Frame-based spectral feature vectors (typically every 10 milliseconds)**
- **Efficiently represented with Mel-frequency scale cepstral coefficients**
  - Typically ~13 MFCCs used per frame



Models

Statistical Approach to ASR



- **Given acoustic observations, *A*, choose word sequence, *W**, which maximizes *a posteriori* probability, *P(W |A)***

$$W^* = \arg\max_W P(W \mid A)$$

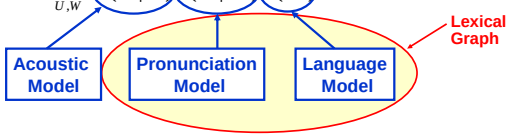
- **Bayes rule is typically used to decompose *P(W |A)* into acoustic and linguistic terms**

$$P(W \mid A) = \frac{P(A \mid W)P(W)}{P(A)}$$

Probabilistic Framework

- **Words are typically represented as sequence of phonetic units**
- **Using phonetic units, U, expression expands to:**

$$\max_{U,W} P(A \mid U)P(U \mid W)P(W)$$



- **Search must efficiently find most likely U and W**
- **Pronunciation and language models encoded in a graph**

Language Modeling

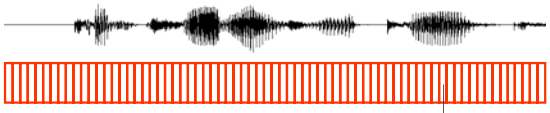
- **ASR systems constrain possible word combinations by way of simple, but powerful, language models:**
  - Finite-state network
  - Deterministic, sequential constraints (e.g., word-pair)
  - Probabilistic, sequential constraints (e.g., bigram, trigram)
- **Trigram is the dominant language model for ASR:**

$$P(w_n \mid w_{n-2}, w_{n-1})$$

  - Much effort has gone into smoothing techniques for sparse data
- **Task difficulty is measured by perplexity**

Acoustic Modeling

Waveform

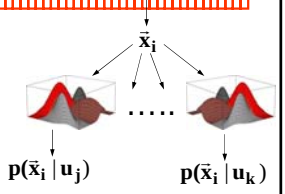


- **Feature vector scoring:**

$$P(A \mid U) = \prod_{i=0}^{N} P(\vec{x}_i \mid u_i)$$

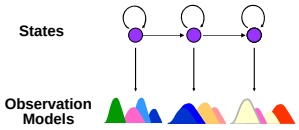
- **Each phonetic unit modeled w/ a mixture of Gaussians:**

$$P(\vec{x} \mid u) = \sum_{j=0}^{M} w_j N(\vec{x} \mid \mu_j, \Sigma_j)$$



Hidden Markov Models

- **Dominant modeling framework used for speech recognition**
- **Generative model that predicts likelihood of observation sequence O being generated by state sequence Q**
  - Either discrete or continuous observation models can be used



- **HMMs can model words or sub-words (e.g., phones)**
  - Sub-word HMMs concatenated to create larger word-based HMMs

Phonological Modeling

- **Words described by phonemic baseforms**
- **Phonological rules expand baseforms into graph, e.g.,**
  - Deletion of stop bursts in syllable coda (e.g., *laptop*)
  - Deletion of /t/ in various environments (e.g., *intersection*, *crafts)*
  - Gemination of fricatives and nasals (e.g., *this side, in nome*)
  - Place assimilation (e.g., *did you* (/d ih jh uw/))

batter : b ae tf er

This can be realized phonetically as:

bcl b ae tcl t er — Standard /t/

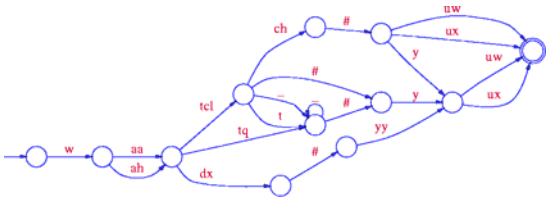
or as:

bcl b ae dx er — Flapped /t/

- **Arc probabilities can be trained (i.e., *P(U|W)*)**

Phonological Example

- **Example of "what you" expanded with phonological rules**
  - Final /t/ in "what" can be realized as released, unreleased, palatalized, or glottal stop, or flap



"what" "you"

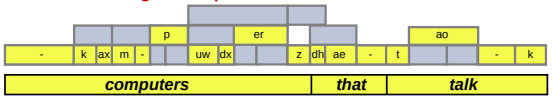
Search

A Simple View of Speech Recognition

Waveform

Frame-based measurements

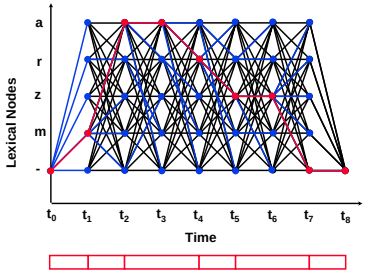
Acoustic models generate phonetic likelihoods



Probabilistic search finds most likely phone & word strings

Viterbi Search Example

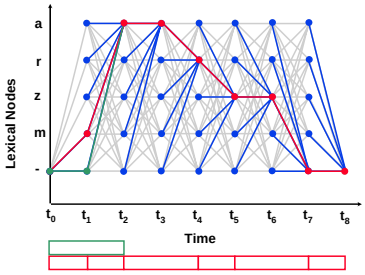
- **Viterbi search typically used in first-pass to find best path**



- **Relative and absolute thresholds used to speed-up search**

A* Search Example

- **Second pass uses backwards *A** search to find *N*-best paths**
- **Viterbi backtrace is used as future estimate for path scores**



Search Issues

- **Search often uses forward and backward passes, e.g.,**
  - Forward Viterbi search using bigram
  - Backwards A* search using bigram to create a word graph
  - Rescore word graph with trigram (i.e., subtract bigram scores)
  - Backwards A* search using trigram to create *N*-best outputs
- **Search relies on two types of pruning:**
  - Pruning based on relative likelihood score
  - Pruning based maximum number of hypotheses
  - Pruning provides tradeoff between speed and accuracy
- **Multiple searches is a form of successive refinement**
  - More sophisticated models can be used in each iteration

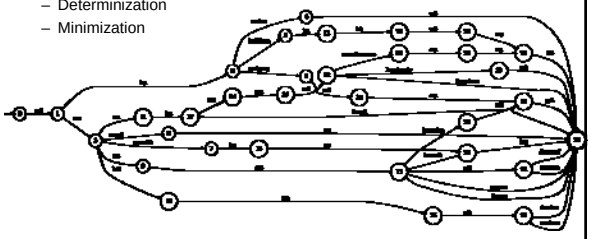
Representations

Finite-State Transducers

- **Most speech recognition constraints and results can be represented as finite-state automata:**
  - Language models (e.g., n-grams and word networks)
  - Lexicons
  - Phonological rules
  - N-best lists
  - Word graphs
  - Recognition paths
- **Common representation and algorithms desirable**
  - Consistency
  - Powerful algorithms can be employed throughout system
  - Flexibility to combine or factor in unforeseen ways
- **Finite-state transducers (FSTs) are effective for defining weighted relationships between regular languages**
  - Extend FSAs by enabling transduction between input and output strings
  - Pioneered by researchers at AT&T for use in speech recognition

Example FST Operations

- **Construction (produce new functionality)**
  - Union: A U B
  - Composition: A o B
- **Optimization (retain original functionality)**
  - Determinization
  - Minimization



Speech Recognition as Cascade of FSTs

- **Cascade of FSTs**

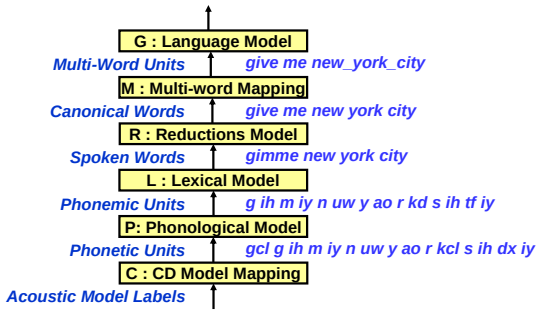
O o (M o P o L o G)

  - G: language model (weighted words ← words)
  - L: lexicon (phonemes ← words)
  - P: phonological rule application (phones ← phonemes)
  - M: model topology (e.g., HMM) (states ← phones)
  - O: observations with acoustic model scores

- **(M o P o L o G) is *single* FST seen by search**
- **Search performs composition of O with (M o P o L o G)**
- **Gives great flexibility in how components are combined**

Expanded FST Representation

- FST representation can be expanded for more efficient representation of lexical variation



Related Areas of Research

- Speech understanding and spoken dialogue
- Multimodal interaction
- Audio-visual analysis (e.g., AVSR)
- Spoken document retrieval
- Speaker identification and verification
- Paralinguistic analysis (e.g., emotion)
- Acoustic scene analysis (e.g., CASA)
- …