interACT

11-751/ 18-781

Speech Recognition and Understanding

Florian Metze

August 26, 2013

lti

Carnegie Mellon

Welcome!

interACT

Course web-site:	(https://sites.google.com/a/is.cs.cmu.edu/11-751-fall-2012/) and Blackboard
Lecturers:	Florian Metze (fmetze@cs.cmu.edu; 407 SCRG or GHC 5703 by appointment)
TA:	Yajie Miao (ymiao@cs.cmu.edu; 407 SCRG)
Time:	4:30pm Mon and Wed, GHC 4102
Guest Lecturers:	Richard Stern (ECE), Monika Woszczyna (M*Modal), Alan Black (LTI)

- Please check the web site frequently, we will provide slides and information
- Please make yourself known if you‘re visiting

Term Projects and Homework

Term project

- Can be performed in groups
  - Ideally self-organized, we'll accept project suggestions
- Ideas presented soon
- Submit proposal before end of September
- Work, presentation and report

Homework

- Four homework assignments
- Individual work – **do not work with others**
- Closely related to topic
- Two weeks to work on each

Important Dates (tentative)

- **Oct 1st** Project proposals due
- **Oct 24st** Project progress presentations
- **Dec 5th** Final term project reports due, term project presentations
- **Dec ?** Exam
- 4 homeworks every two weeks

Grading/ Exams

Grading for course:

- 40% weight on Final Exam (mid-december)
- 30% on Homework
- 30% on Term Project

Details on requirements etc as we go along

Specific requirements from your side?

Literature (see papers on Blackboard)

- Xuedong Huang, Alex Acero and Hsiao-wuen Hon, Spoken Language Processing, Prentice Hall PTR, NJ, 2001
- Rabiner and Juang, Fundamentals of Speech Recognition, Prentice Hall Signal Processing Series, Englewood Cliffs, NJ, 1993
- Jurafsky and Martin, Speech and Language Processing, 2nd ed. Prentice Hall, 2008.
- Jelinek, Statistical Methods for Speech Recognition, MIT Press, Cambridge, MA, 1997
- Schultz and Kirchhoff, Multilingual Speech Processing, Elsevier, Academic Press, 2006
- Waibel and Lee, Readings in Speech Recognition, Morgan Kaufman Publishers, San Mateo, CA, 1990

Follow-up Courses

Speech Courses in spring semester:

- 11-753: Advanced Lab in Speech Recognition and Understanding (S14, Florian Metze)
- 11-783: Rich Interaction in Virtual Worlds Lab (S14, Florian Metze)

Agenda

- Practicalities
- What is 11-751 / 18-781 about? Whom is it for?
- Why Speech Recognition and Understanding? Why is it interesting and difficult?
- How to Approach Speech Recognition (not Understanding)
- State of the Art
- Speech Production

Whom is 11-751 for?

- Primarily for graduate students in LTI, CS, Robotics, ECE, HCI, Psychology, or Computational Linguistics. Others by prior permission of instructor
- No prior experience with speech recognition is necessary, but a solid background in mathematics, computer science, or signal processing will help
- The course is suitable for graduate students with some background in computer science, electrical engineering, Human-computer interaction or natural language processing, as well as for advanced undergraduates

Course Overview I

- **ASR – The Big Picture**
  - Evaluation
  - Speech Production
  - Linguistics and Phonetics
- Pattern Recognition and Classification
- Template-based Recognition
- Speaker Identification and Meta-Data Classification

A general overview on how statistical methods can be used to recognize speech and what else can be done using the same methods. How can ASR systems be evaluated and compared?

Course Overview II

interACT

- Signal Processing
- Hidden Markov Models
- Acoustic Modeling
- Language Modeling
- Search: Tree Search and wFSTs
- Discriminative Training
- Adaptation
- Deep Learning

This covers the **state-of-the art** in today's ASR systems. We will treat theoretical methods and some tricks of the trade and cover some of the active current research areas.

Course Overview III

interACT

- Speech Dialog Systems
- Multi-modal Interaction
- Spoken Language Understanding
- Question Answering
- Industrial Applications

This section covers **ASR** (aka speech to text) **as part of a bigger system**, which can translate speech into foreign languages, answer questions, understand languages – including your term project.

Why Speech Recognition?

Speech as a Communication Medium

- Speech is the most natural and powerful form of communication between humans
  - **Natural:** No additional training required
  - **Flexible:** Adapt to dialogue partners, environment and situations
  - **Efficient:** Communicate large amounts of information
    - Information about speaker, their cognitive-state, environment
  - Good for large amounts of information

Input Speeds (Characters per Minute)

These numbers are of course approximate

Mode	Standard	Best
Handwriting	200	500
Typewriter	200	1.000
Stenography	500	2.000
Speech	1.000	4.000

Speech for Human Computer Interaction

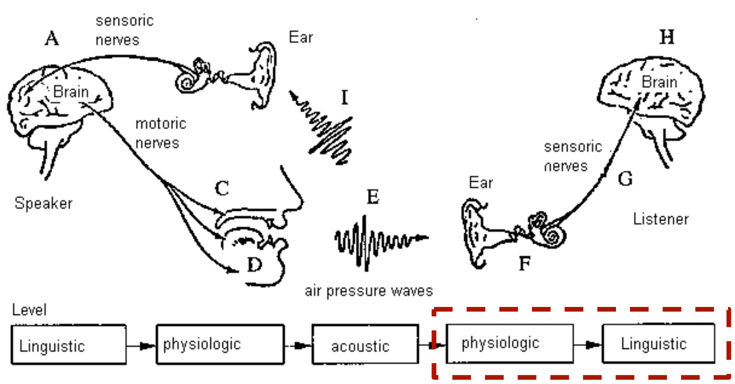
- **Usability:** Novice users can complete complex tasks with little additional training, same interface can be used be expert users to quickly complete task
- **Ubiquity:** Only require cellular phone to access information
- Suitable for busy environments ("hands/ eyes free")
  - Information retrieval & device operation (in car)
  - Voice-based manual reference during maintenance (NASA) or in warehouses (Vocollect)
  - Speech-translation for medical, military tasks (checkpoints)
- Can effectively combine with other modalities of interaction

Current ASR Technologies?

Dictation
(mildly useful)

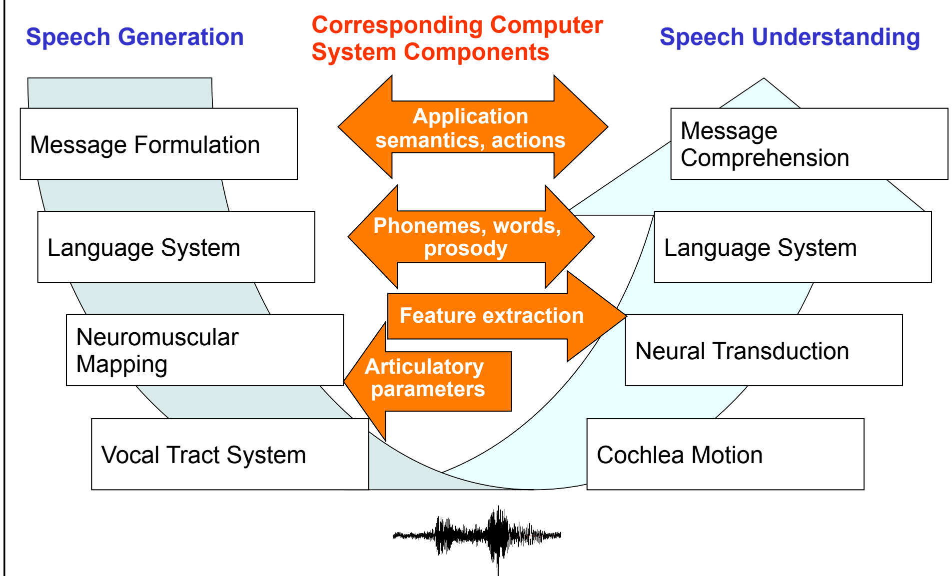
Intelligent IVR Systems
(very frustrating)

The Speech Chain (Human to Human)



a) Neuro-physiological process in the speaker‘s brain
b) Electrical process in the efferent nerves (impulses **from** the central nervous system)
c) Resulting position and movement of articulatory apparatus
d) Acoustical production of acoustic speech signal in vocal tract
e) Acoustical transmission of the speech signal
f) Mechanical process in the middle ear, hydro-mechanical process in the inner ear
g) Electrical signals on the afferent nerves (impulses **to** the central nervous system)
h) Neuro-physiological process in the listener‘s brain
i) Acoustic feedback to the speaker‘s ear

The Speech Chain (Human to Computer)



Problems and Research Questions

- Speech Recognition (“speech-to-text”)
  - Finding Robust Representations of Speech
  - Acoustic Modeling (how do things sound)
  - Dictionary Learning (how to decompose words)
  - Language Modeling (what is likely to be said)
  - Decoding (how to get an answer in finite time)
- Meta-data extraction (what is not in text)
  - Speaker identification (age, gender, ...)
  - Emotions, personalities, ...
  - Languages, dialects, ...
- Adaptation of models and techniques to changing conditions
- Integration and proper optimization of models to go from speech-to-text towards “speech-to-meaning” or “speech-to-action”

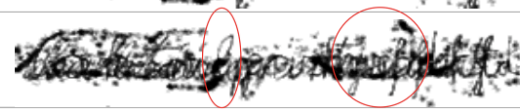
True or Not?

interACT

At an international conference on speech processing, a speech scientist once held up a tube of toothpaste (whose brand was "Signal") and, squeezing it in front of the audience, coined the phrase: “This is speech synthesis; speech recognition is the art of pushing the toothpaste back into the tube.”

Why is Speech Recognition Difficult?

interACT

written text:	Why is speech Recognition so Difficult?
spontaneous:	why's speech recognition so difficult
continuous:	whysspeechrecognitionsodifficult
pronunciation:	whazbeechregnizhnsadifcld
acoustic variability:	
noise:	
Cocktail party-Effect:	

Which Factors Influence Difficulty?

COMPLEXITY	
amount of data:	typically 32000 bytes per second (16khz)
class inventory:	50 phonemes, 5000 sounds, 100.000 words
combinatorial explosion:	exponential growth of possible sentences
SEGMENTATION	
our perception:	Phones, syllables, words, sentences
actually there are:	no boundary markers, continuous flow of samples
VARIABILITY	
speaker:	anatomy of vocal tract, speed, loudness, acoustic stress, mood, dialect, speaking style, context
channel, environment:	noise, microphones, channel conditions
AMBIGUITY	
Homophones:	two vs. too,
Word Boundaries:	interface vs. in her face,
Semantics:	He saw the Grand Canyon flying to New York,
Pragmatics:	Time flies like an arrow.

So, Why Is It Easy for Humans?

"The main prerequisite of the uniquely human communication is that speaker and listener must have a common understanding that out of all possible sounds man can produce and hear, only a few have linguistic significance."

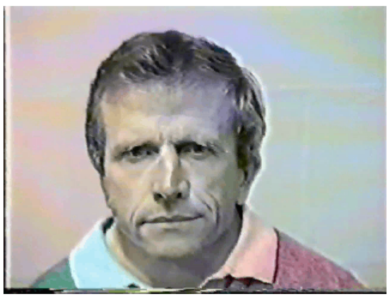
(Olli Aaltonen& Esa Uusipaikka: Why Speaking Is so Easy? – Because Talking Is Like Walking with a Mouth)

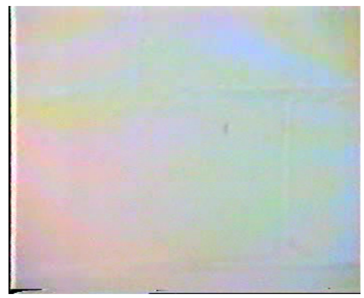
- Important feature of speech perception: we hear sounds either as speech or non-speech
- Once defined as speech we hear them a sequence of vowels and consonants not as buzzes and hisses, the segmentation into words happens on the fly
- Abstract away from sound variability - we use an enormous database of background knowledge: phonotactics, morphology, syntax, semantics, pragmatic knowledge
- But: beware …

Multimodal Perception

Humans use many contextual ques to understand speech

- **McGurk Effect:** Speech Interpreted using both Acoustic and Visual information





McGurk Effect Explanation

- “My bab pop me poo brive”, dubbed onto the video
- “My gag kok me koo grive”, with the expected McGurk effect of perceiving
- “My dad taught me to drive”

- Also: “Bateson Experiment”
  - Random eye gaze during conversations reduced in noise
  - (Vatikiotis-Bateson, 1998)

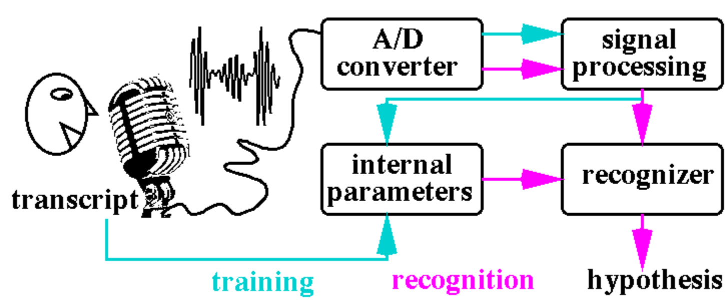
How to Approach Speech Recognition

How does ASR work?

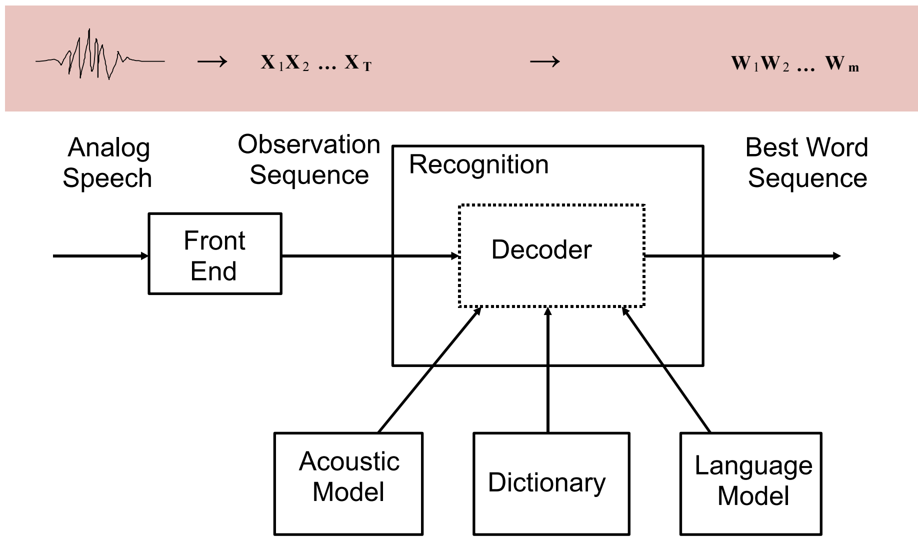
Two-stage process for statistical-based ASR (Automatic Speech Recognition):

1. **Train** statistical model (Maximum Likelihood or discriminative approach)
2. **Test** on unknown data

Output is **most likely** hypothesis according to internal model



Speech Recognition Components



When Is a Recognizer Good?

Typical criteria for the evaluation of modern large vocabulary recognisers are

Word-Error-Rate: WER = #Errors / #Spoken_Words

Word Accuracy: WA = 1 – WER

#Errors = #substitutions + #deletions + #insertions (alignment errors taken)

Example: WER = ¾ = 75%

Reference:		`SHOW`	`ME`	`THE`	`INTERFACE`
Hypothesis:	`I`	`SHOW`	`ME`		`FACE`
Alignment:	`I`			`D`	`S`

Alignment is not unique, but error count is (`FACE` could also be aligned with `THE`)

Note that we cannot optimize for this directly!

Fundamental Equation of Speech Recognition

interACT

Given: an observation (ADC, FFT) $X = x_1, x_2, \ldots, x_T$

Wanted: the corresponding word sequence $W = w_1, w_2, \ldots, w_m$

Search: the most likely word sequence W'

$$W' = \arg\max_W P(W \mid X) = \arg\max_W \frac{p(X \mid W) P(W)}{p(X)} = \arg\max_W p(X \mid W) P(W)$$

(Bayes)

$p(X|W)$ = The **acoustic model**
(how likely is it to observe X when W is spoken)

$P(W)$ = The **language model**
(how likely is it that W is spoken a-priori)

Fundamental Problem of Speech Recognition

interACT

- We want to minimize the Word Error Rate (WER): $<P(w_i)>$ with $W=w_1, w_2, \ldots$
  - Can be evaluated automatically
  - Typically correlates with application-specific optimality criteria
- But: we typically do not solve for $<P(w_i)>$ during recognition
- $<P(W/X)>$ (as in the Fundamental Equation of Speech Recognition) expresses the highest proportion of correct sentences (not words)
- These issues can be addressed (will discuss)
- The criterion for which $P(W)$ and $P(X/W)$ are trained are also “ad-hoc”
- This is why speech recognition is still somewhat a “black art”

Speech Recognition Conundrum

- We can “kind of” convert speech to text
  - Spoken language is different from written language, needs different processing
  - “Uhm, he was like, you know, like totally, uhm, yeah, really nice”
  - But we cannot just download tons of text from the Internet
- In some cases, “superhuman performance” can be achieved
- But some things we are still doing fundamentally wrong
- Proof: “hyper-articulated” speech
  - Talk to a speech dialog system, it will often fail to understand you
  - If you try to speak extra clearly, it will typically understand you even less
  - We have no idea how to model the changes in speech that occur
  - This means our modeling assumptions are fundamentally wrong