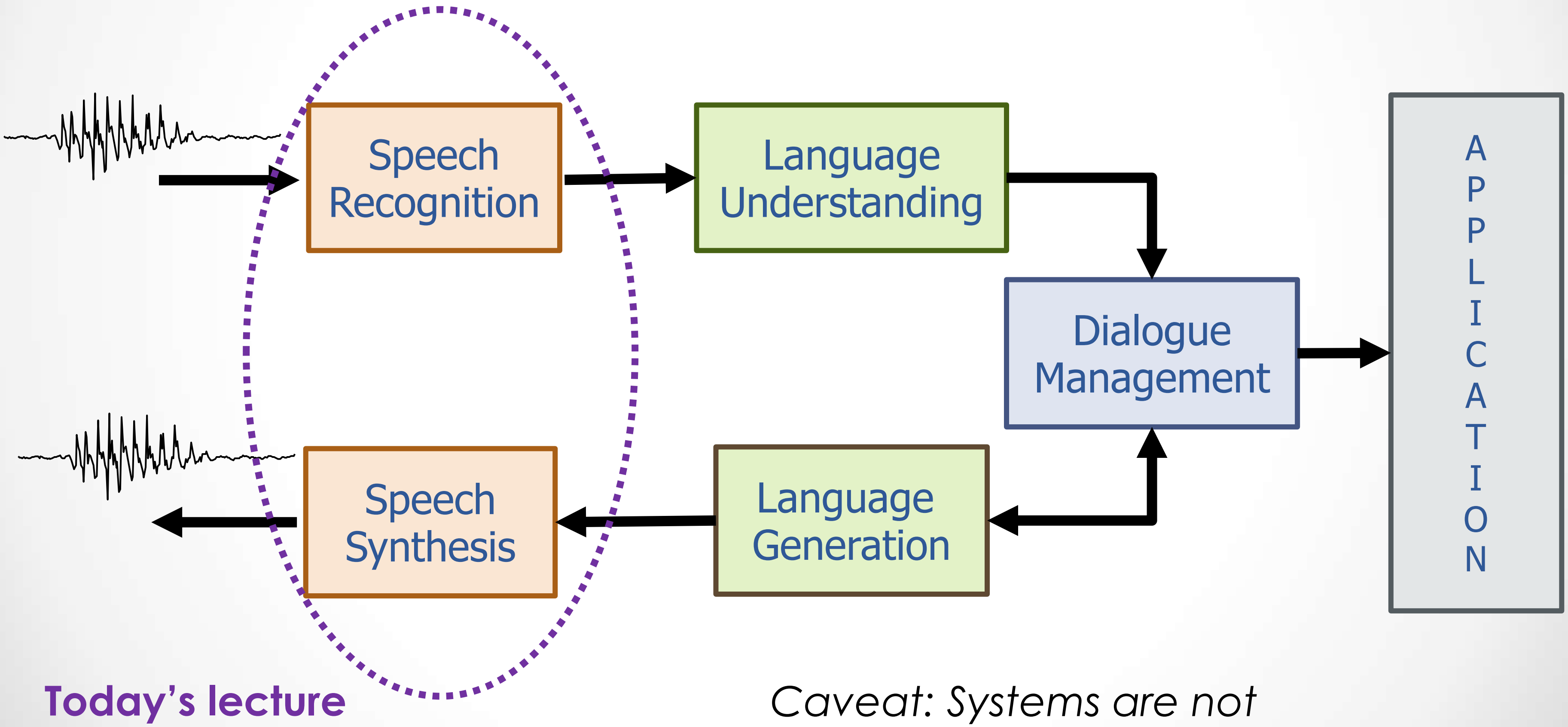
Speech Recognition and Synthesis for Conversational AI

Mari Ostendorf
University of Washington
EE596 – Spring 2018

Dialogue System Components



Caveat: Systems are not always quite so pipelined.

User-Interface Technologies

- Input side:
  - Acoustic processing
  - Automatic speech recognition
  - Natural language understanding
- Dialogue management
  - Problem or help request detection
  - Interaction with application
  - Context tracking
- Output side
  - Response generation
  - Text-to-speech synthesis

Speech

NLP

ASR

Acronyms

TTS

Overview

- General issues in speech processing
- Core recognition and synthesis technology
- What you need to know for working with commercial systems
- Recent advances & challenges

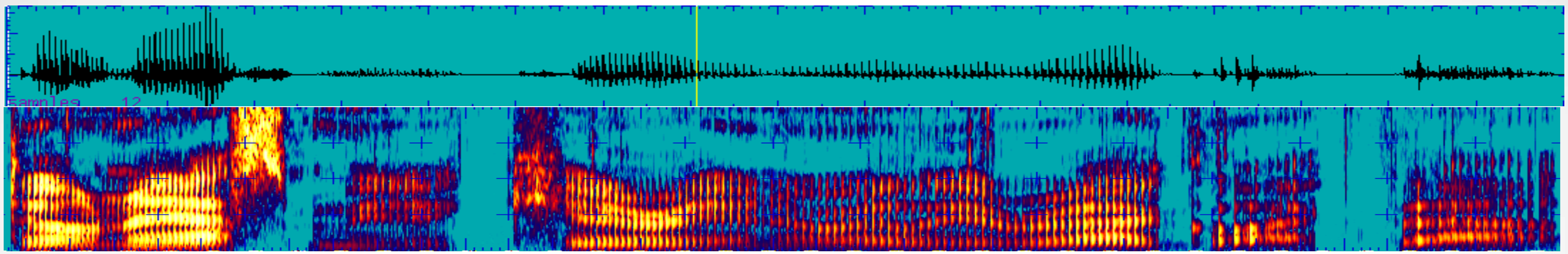
General Issues

Information in speech
Limitations of words
Modules & symbols

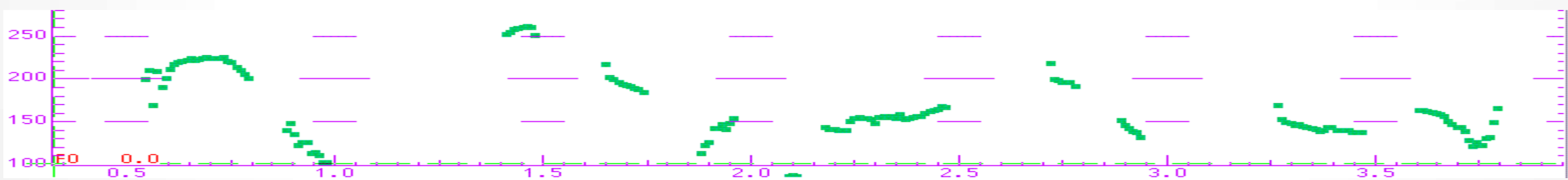
Information in Speech

- Spoken language carries information at many levels
  - Syntactic and semantic meaning
  - Emotion, affect
  - Speaker, dialect/sociolect
  - Social context, status, goals
- That information is reflected in both the audio signal and the choice of words

Information in Audio



- Spectral information:
  - Short term: phonemes that make up words
  - Long term: speaker characteristics, environment noise



- Prosodic information:
  - Short-term: constituent boundaries, intent, emphasis
  - Long-term: speaker, emotion, discourse structure

Problems with ASR Transcripts

- Speech/non-speech detection
- Speech recognition errors
- Speaker/sentence segmentation, punctuation
- Disfluencies (fillers, self corrections)

ok so what do you think well that's a pretty loaded topic absolutely well here in uh hang on just a second the dog is barking ok here in oklahoma we just went through a uh major educational reform…

Ok, so what do you think?

Well that's a pretty loaded topic.

Absolutely.

Well, here in …. Ok, here in Oklahoma, we just went through a major educational reform…

How we really talk…

A: and ***~~that~~*** that concerns me greatly. /

B: Well, ***~~I don't~~***, -/ yeah, / I'd certainly **uh** support Israel ***~~in in their their policy that~~*** in defending themselves and ***~~in~~*** **uh** in their handling of their foreign policy, / ***~~I think~~*** I think ***~~the stand they have,~~*** ***~~or~~*** **or** the way they command respect, ***~~I~~*** I support that. / I think that is ***~~a~~*** a positive thing for them after **um uh** thousands of years, / ***~~they have to,~~*** **uh**, ***~~they ha- I think they in -/~~*** when they ***~~be-~~*** became a country they ***~~more than or~~*** more or less decided they weren't going to take it anymore, / and **uh** -/

A: Well, they didn't have much choice, / they could either fight or die. /

B: Yeah, / exactly, exactly / ***~~and,~~*** **uh** **um** ***~~so~~*** ***gee, I lost my train of thought here.*** / ***~~But~~*** **uh um** ***~~so~~*** ***okay*** / so I can't say whether ***~~that~~*** that I'm pro Israel or anti Israel. / ….

… as do justices and lawyers

Underwood: And this Court said it wasn't sufficient in Buckley, and observed that that's ~~***part of why the***~~ part of what justifies the limit on individual ***um uh*** contributions in a campaign, the total limit, not

Rehnquist: ~~***Is is***~~ is the argument, General Underwood, ~~***it***~~ it is not that the party is corrupted, I take it, because that would seem just fatuous, but the party is kind of a means to corrupting the candidate himself?

Underwood: Yes. ~~***That***~~ that is ~~***there there***~~ ***uh uh*** there are two arguments about the risk of corruption.
At the moment the argument that I'm talking about is ~~***that the party is a means that that to that***~~ that the ***um*** contribution limits on individual donors are justified as a means of preventing ***uh*** corruption and the risk of corruption donor to candidate, and that the party, ~~***as an in-***~~ as an intermediary, ~~***can facilitate,***~~ can essentially undermine that mechanism that the individuals can exceed their contribution limits.

Disfluencies are Common

- Multiple studies find disfluency rates of 6% or more in human-human speech
- People have some control over their disfluency rate, but ***everyone*** is disfluent
- People aren't usually conscious of disfluencies, so transcripts may miss them
- But they use them as speakers & listeners; evidence in fMRI studies

Disfluencies as…

Noise

- Degraded transcripts hurt readability for humans
- Word fragments are difficult to handle in speech recognition
- Grammatical “interruptions” create problems for parsing (and NLP more generally)

Information

- Listeners use disfluencies as cues to corrections
- Speakers use “um” in turntaking
- Silent & filled pauses indicate speaker confidence
- Disfluency rate reflects cognitive load, emotion (stress, anxiety)

Word Ambiguity

- Many sources of ambiguity in language
  - Word sense ambiguities can be resolved from lexical context
  - Intent ambiguities require prosody
    - "yeah" as agreement vs. "I'm listening" vs. sarcasm
    - Many other examples impact dialog: ok, thank you
- Problem for speech technology
  - Understanding ambiguities
  - TTS: Sounding Board vs. Sounding bored

Modules and Symbols

- Speech is inherently continuous; language is communicated with discrete symbols
- Speech recognition and synthesis involves mapping between these domains
- Historically, the mapping is broken into stages with symbolic communication
  - Advantages: more efficient training, more control over experiments
  - Disadvantages: hard decision error propagation, missed interactions

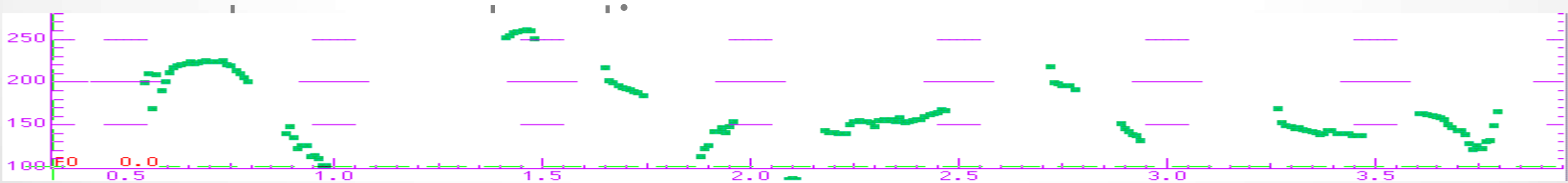
Prosody: Symbol and Signal

- Two representations of prosody
- Symbolic level: prosodic phrase structure, word prominence, tonal patterns

Wanted: Chief Justice of the Massachusetts Supreme Court.

- Continuous parameters:

  fundamental frequency (F0), energy, segmental

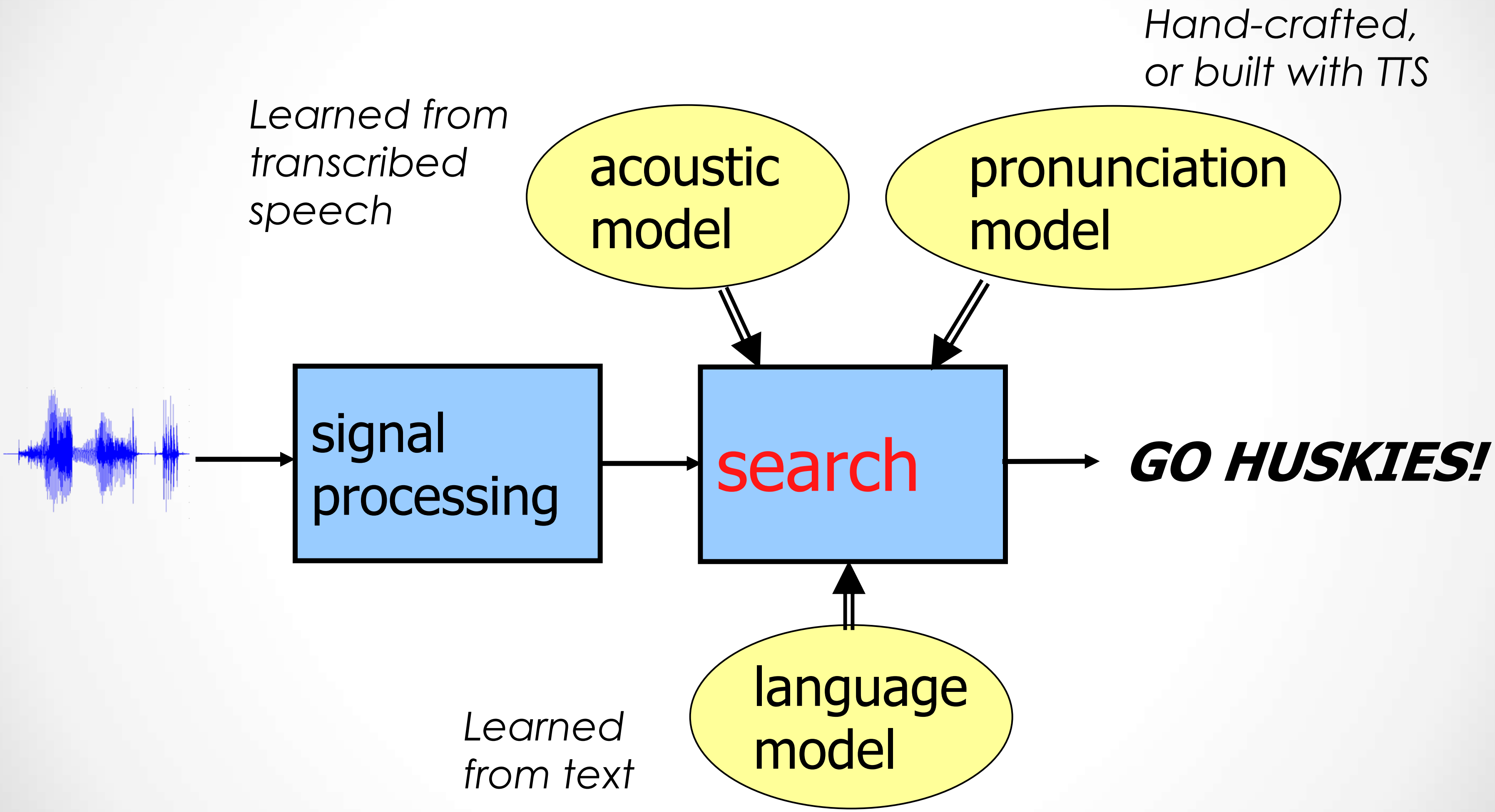


Core Speech Technology

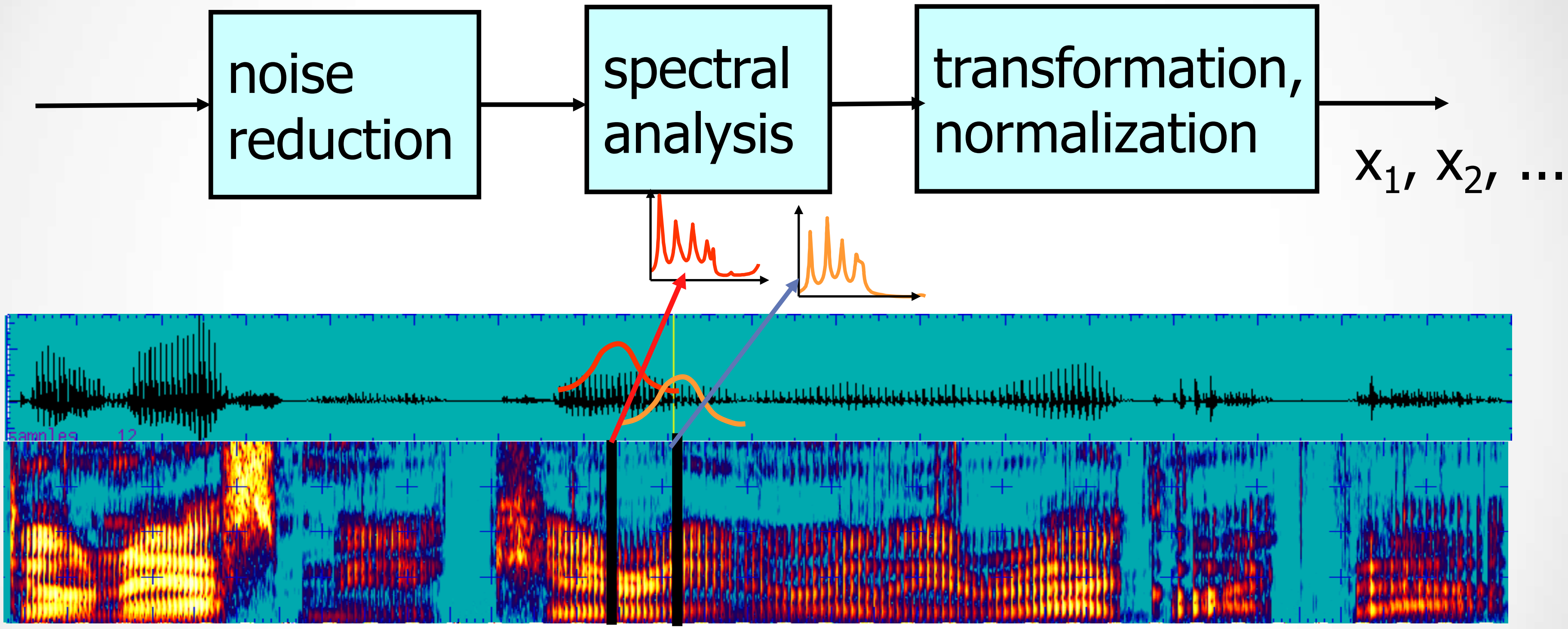
Speech Recognition

Speech Synthesis

Classical ASR



Signal Processing



- Noise reduction often involves multi-mic beamforming
- Spectral analysis can involve time & frequency slices
- Normalization accounts for channel variation, speaker differences

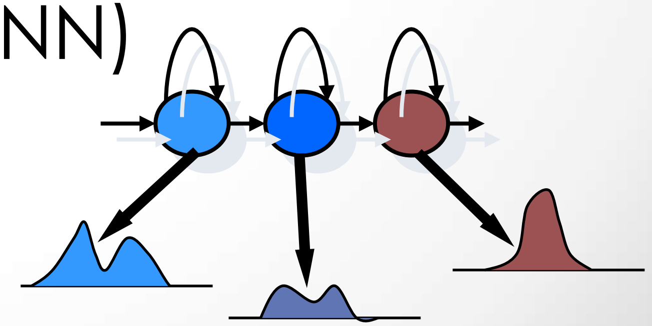
Language Model

- Goal: describe the probabilities of sequences of words
  - $p(w) = \Pi_i \, p(w_i \mid \text{history})$
- Needed to discriminate similar sounding words
  - "Write to Mrs. Wright right now."
- Most common language model: trigram $p(w_n \mid w_{n-2}, w_{n-1})$
  - actually quite powerful, e.g. p(? | president, donald)
  - Difficult parameter estimation problem (e.g., 60k words, 2.16e14 entries)

Acoustic Model

- Words are built from "phones" (aa, ow, ih, s, t, m, ….) using hidden Markov models (HMMs) to capture feature & time variation.
- Each phone is characterized as a sequence of "states", depending on the neighboring phonemes, that form a "template" to match against dynamically.
- Each state q_t represents a feature x_t using a mixture of Gaussians (or DNN)

(ignorance modeling)

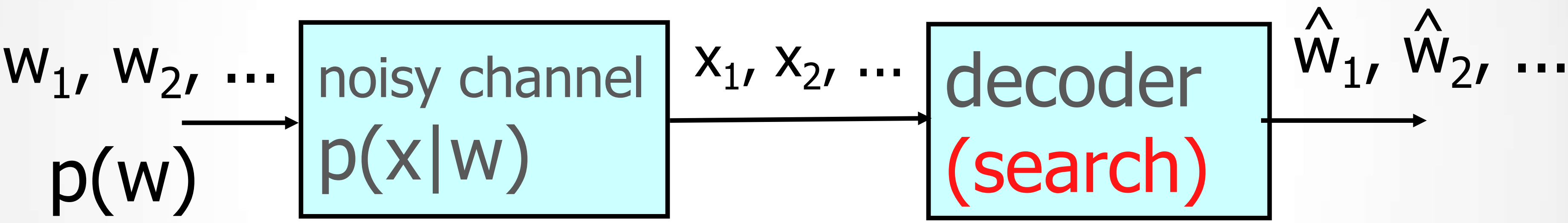


Pronunciation Model

- Simple approach: list alternatives
  - e.g. “and” -- “ae n d”, “eh n d”, “ae n”, “n”, …..
- Need probabilities to reduce confusability between words (e.g. “and” vs. “an”)
- Pronunciation model must handle speaking style, dialect, foreign accent, etc.

Search: Brute Force Approach

- Speech recognition formulated as a communications theory problem:



$$\hat{w} = \underset{w}{\operatorname{argmax}}\, p(w|x) = \underset{w}{\operatorname{argmax}}\, p(x|w)p(w)$$

- … means try everything, requires lots of computing

Words are Not Enough

o- ohio state's pretty big isn't it yeah yeah I mean oh it's you know we're about to do like the the uh fiesta bowl there oh yeah

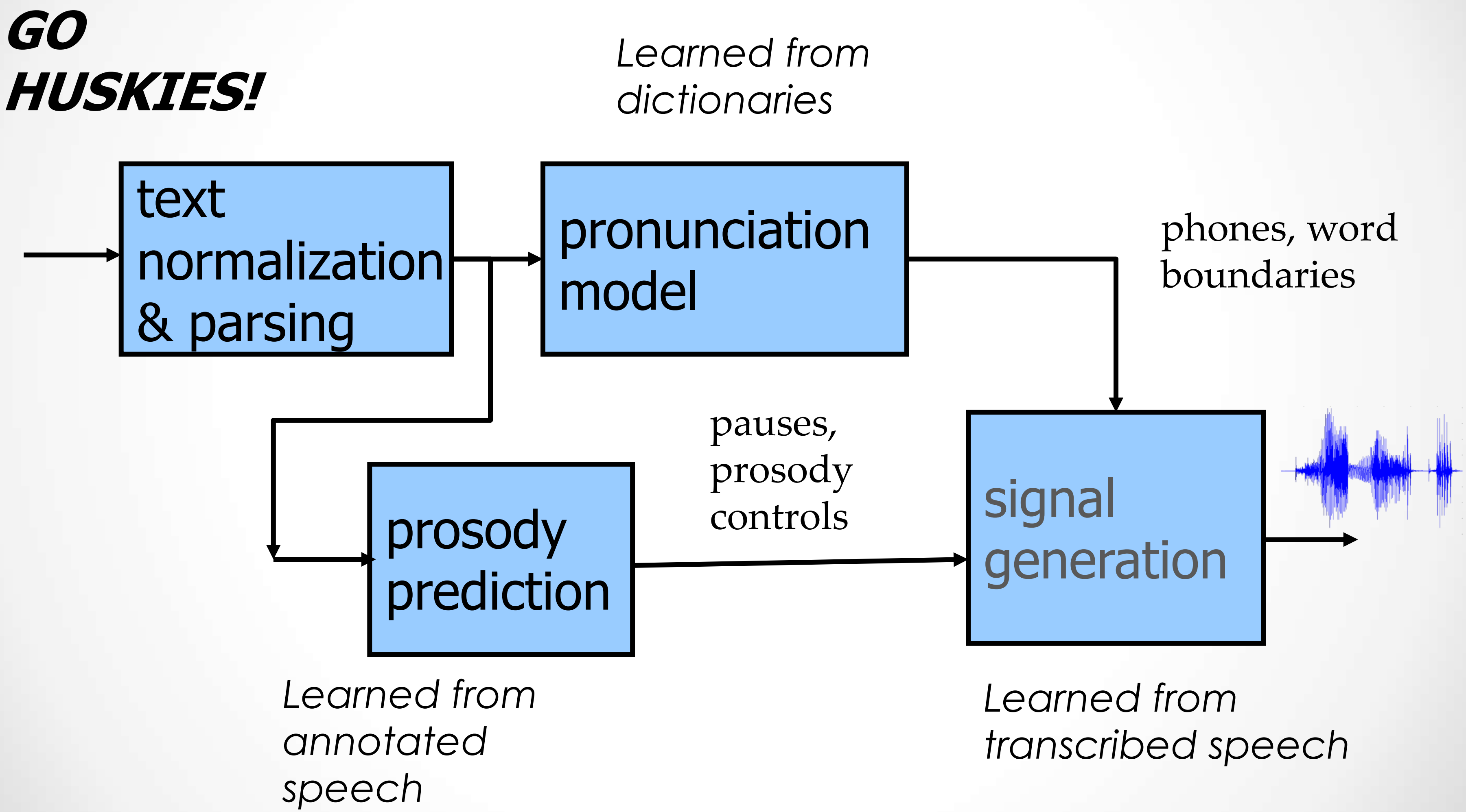
A: O- Ohio State's pretty big, isn't it?
B: Yeah. Yeah. I mean- oh it's you know- we're about to do like the the uh Fiesta Bowl there.
A: Oh, yeah.

A: Ohio State's pretty big, isn't it?
B: Yeah. Yeah. We're about to do the Fiesta Bowl there.
A: Oh, yeah.

Rich Transcription of Speech

- Goals:
  - Endow speech with characteristics that make text easy to manage, *AND*
  - Represent (don't discard) the extra information that makes speech more valuable to humans
- Recognizing the spoken words and …
  - Story segmentation
  - Speaker segmentation and ID
  - Sentence segmentation & punctuation
  - Disfluencies
  - Prosodic phrase boundaries, emphasis
  - Syntactic structure
  - Speech acts (question, statement, disagree, …)
  - Mood (e.g. in talk shows)

Classical TTS



Acoustic Models

- Model-based synthesis
  - Source-filter vocoder
  - Generative recognition models
- Concatenative (unit selection)
  - Large inventory of annotated speech snippets (time-marked speech)
  - Dynamic programming search to minimize loss function (unit match & concatenation cost)
  - Synthesis with juncture smoothing

Practical Issues

Lexical uncertainty

Error handling

Situation-sensitive synthesis

Typical Commercial System

- The ASR interface provides only word transcripts
  - No sentence boundaries, may or may not have pauses, probably no word times
  - No access to audio for privacy reasons
  - Typically some sort of confidence indicator
- The TTS interface takes only word transcripts (with punctuation)
  - Speech generated with a reading style
  - Optionally some simple prosody controls

Lexical Uncertainty

- ASR uncertainty modeling
  - In decoding, systems often build a lattice of possible word hypotheses.
  - Each arc in the lattice can be associated with a likelihood
- Simple representations of uncertainty
  - N best sentence hypotheses + sentence-level confidence score
  - Confusion network + word-level confidences

Options for using Confidence

- Sentence-level confidence:
  - Criterion for rejecting the transcript (ask the user to repeat or change the topic)
  - Intent classification using a weighted combination of ASR and NLU confidences
- Word-level confidence
  - Feature for detecting out-of-vocabulary words
  - Criterion for ignoring a word in slot filling or asking the user to confirm something
  - Weighted bag-of-words input to vector space model
  - Confidence-weighted rules in parsing

Confirmation & Error Handling

- Two types of errors:
  - ASR confidence tells you that the transcript is bad
  - What the user is saying suggests that the system made a mistake
- Considerations:
  - Errors derail the dialog but too many confirmations are annoying
  - Asking for a repeat may give the same error; asking for confirmation of one thing may give better results
  - Apologies are helpful if not too frequent

Situation-Sensitive Synthesis

- SSML = speech synthesis mark-up language
  - Pronunciation: ‘say as’
  - Prosody
    - Symbolic (break, emphasis)
    - Continuous (rate, pitch, volume)
- When would you use SSML:
  - the TTS pronunciation is wrong,
  - the default prosody is not appropriate (emphasis or pauses in the wrong place),
  - you want to add some enthusiasm or empathy

Recent Advances

Paradigm shift
Technical trends

Providing perspective…

A view of "the future of natural user interfaces" from 2004

Speech Tech Paradigm Shifts

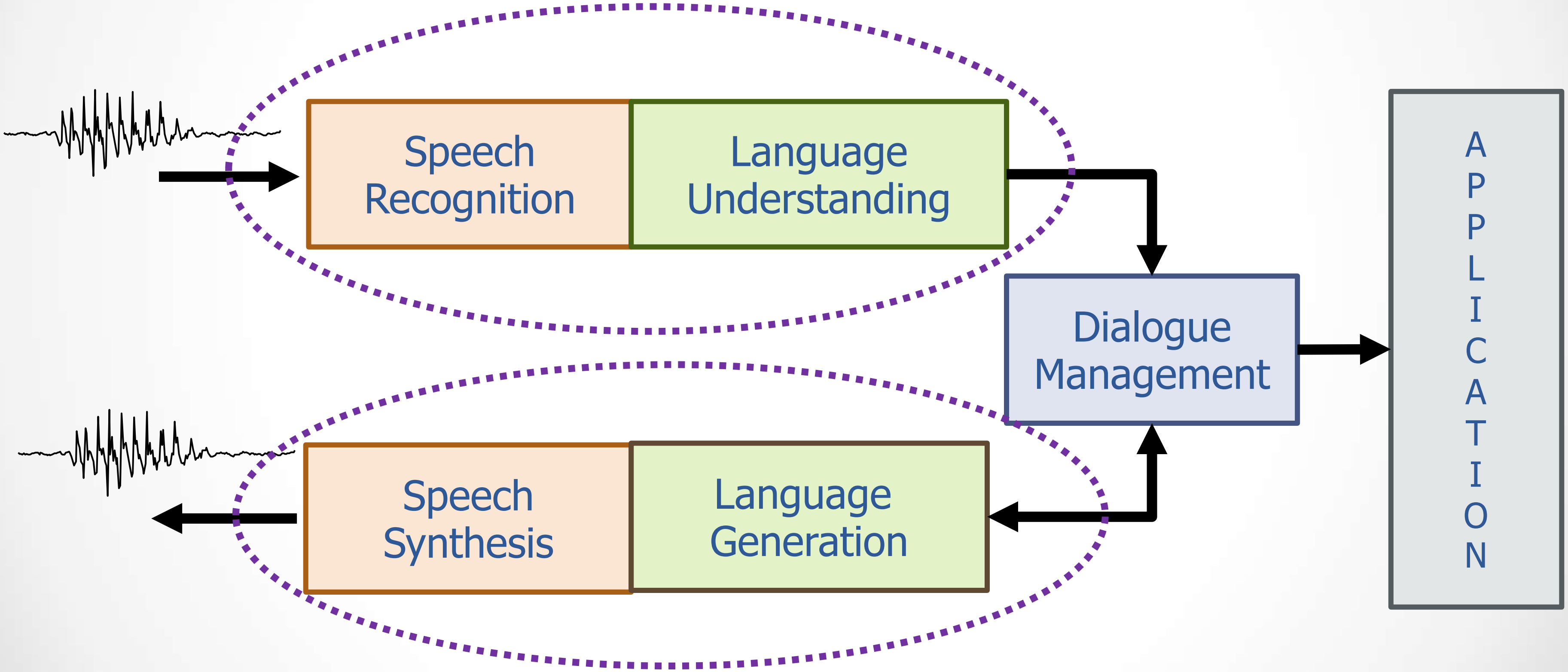
Major changes in speech technology in the past 5 (or so) years:

- Deep learning → improved performance
- People actually use it → more data to learn from
- More natural systems

Impact of Better ASR/TTS

- People (unconsciously) expect more human-like capabilities, and
- Computer-directed speech becomes more like human-directed speech
  - Evidence: increasing rate of disfluencies
- Challenges evolve
  - Speech recognition → speech understanding
  - Simple dialogs → interactive conversation
  - Speech synthesis → speech generation
- Prosody will matter more

Dialogue System Components



Big Trends

- End-to-end systems
- Affective systems

Summary

Summary (I)

- General issues
  - There are many levels of information in speech, characterized by words and prosody
  - Disfluencies create noise & information
  - Symbolic representations are used in many ways
- Core speech technology
  - Speech recognition: signal processing, acoustic model, language model, dictionary → search
  - Rich transcripts: sentences, disfluencies, …
  - Speech synthesis: text norm, prosody prediction, pron prediction → search

Summary (II)

- Practical issues
  - Use word confidence to improve error handling
  - Problems that arise from NLU or dialog errors have different signals
  - SSML can make the conversation more natural
- Advanced speech technology
  - End-to-end systems
  - Affective systems

Thanks!