SDS Systems & Components

Ling575
Spoken Dialog Systems
April 5, 2017

Roadmap

- Aspects of conversation
  - Turn-taking
  - Grounding
  - Speech Acts
  - Implicature
- SDS Pipeline & Components
- ASR
  - Basic approach
  - Recent developments

Conversation

- Intricate, joint activity

Conversation

- Intricate, joint activity
  - Constructed from consecutive turns

Conversation

- Intricate, joint activity
  - Constructed from consecutive turns

  - Joint activity between speakers, hearer

Conversation

- Intricate, joint activity
  - Constructed from consecutive turns
  - Joint activity between speakers, hearer
  - Involves inferences about intended meaning

Conversation

- Intricate, joint activity
  - Constructed from consecutive turns
  - Joint activity between speakers, hearer
  - Involves inferences about intended meaning
- SDS: simpler, but hopefully consistent

Turn-Taking

- Multi-party discourse
  - Need to trade off speaker/hearer roles
    - Interpret reference from sequential utterances
- When?

Turn-Taking

- Multi-party discourse
  - Need to trade off speaker/hearer roles
    - Interpret reference from sequential utterances
- When?
  - End of sentence?

Turn-Taking

- Multi-party discourse
  - Need to trade off speaker/hearer roles
    - Interpret reference from sequential utterances
- When?
  - End of sentence?
    - No: multi-utterance turns
  - Silence?

Turn-Taking

- Multi-party discourse
  - Need to trade off speaker/hearer roles
    - Interpret reference from sequential utterances
- When?
  - End of sentence?
    - No: multi-utterance turns
  - Silence?
    - No: little silence in smooth dialogue:< 250ms
      - Gaps less than actual sentence planning time - anticipate
  - When other starts speaking?

Turn-Taking

- Multi-party discourse
  - Need to trade off speaker/hearer roles
    - Interpret reference from sequential utterances
- When?
  - End of sentence?
    - No: multi-utterance turns
  - Silence?
    - No: little silence in smooth dialogue:< 250ms
      - Gaps less than actual sentence planning time - anticipate
  - When other starts speaking?
    - No: relatively little overlap face-to-face: ~5%

Turn-taking: Who & How

- At each TRP in each turn (Sacks 1974)
  - If speaker has selected A to speak, A must take floor
  - If speaker has selected no one to speak, anyone can
  - If no one else takes the turn, the speaker can
- Selecting speaker A:

Turn-taking: Who & How

- At each TRP in each turn (Sacks 1974)
  - If speaker has selected A to speak, A must take floor
  - If speaker has selected no one to speak, anyone can
  - If no one else takes the turn, the speaker can
- Selecting speaker A:
  - By explicit/implicit mention: What about it, Bob?
    - By gaze, function
- Selecting others:

Turn-taking: Who & How

- At each TRP in each turn (Sacks 1974)
  - If speaker has selected A to speak, A must take floor
  - If speaker has selected no one to speak, anyone can
  - If no one else takes the turn, the speaker can
- Selecting speaker A:
  - By explicit/implicit mention: What about it, Bob?
    - By gaze, function
- Selecting others: questions, greetings, closing
  - (Traum et al., 2003)

Turns and Structure

- Some utterances select others:

Turns and Structure

- Some utterances select others:
  - Adjacency pairs:
    - Greeting – Greeting, Question – Answer,
    - Compliment – Downplayer

Turns and Structure

- Some utterances select others:
  - Adjacency pairs:
    - Greeting – Greeting, Question – Answer,
    - Compliment – Downplayer
- Silence 'dispreferred' within adjacency pair
    - A: Is there something bothering you or not?
    - (1.0)
    - A: Yes or No?
    - (1.5)
    - A: Eh.
    - B: No.

Turn-taking in HCI

- Human turn end:
  -

Turn-taking in HCI

- Human turn end:
  - Detected by 250ms (or longer) silence
- System turn end:

Turn-taking in HCI

- Human turn end:
  - Detected by 250ms (or longer) silence
- System turn end:
  - Signaled by end of speech
  - Indicated by any human sound
    - Barge-in
- Continued attention:

Turn-taking in HCI

- Human turn end:
  - Detected by 250ms (or longer) silence
- System turn end:
  - Signaled by end of speech
  - Indicated by any human sound
    - Barge-in
- Continued attention:
  - No signal
- Design problems create ambiguous silences
  - Problematic for SDS users
    - (Stifelman et al., 1993), (Yankelovich et al, 1995)

Speech Acts

- Utterance:
  - Action performed by the speaker (Austin, 1962)

Speech Acts

- Utterance:
  - Action performed by the speaker (Austin, 1962)
    - Performatives: *name, second*
      - *I name this ship the Titanic.*
      - *I second that motion.*
    - Extend to all utterances

Utterances as 3 Act Types

- Locutionary act:
  - utterance with some meaning
  - *“You can’t do that!”*

Utterances as 3 Act Types

- Locutionary act:
  - utterance with some meaning
  - *"You can't do that!"*
- Illocutionary act:
  - Act of asking, promising, answering, in utterance

Utterances as 3 Act Types

- Locutionary act:
  - utterance with some meaning
  - *"You can't do that!"*
- Illocutionary act:
  - Act of asking, promising, answering, in utterance
  - *Protesting*
- Perlocutionary act:
  - Production of effects on feeling, beliefs of addressee

Utterances as 3 Act Types

- Locutionary act:
  - utterance with some meaning
  - *"You can't do that!"*
- Illocutionary act:
  - Act of asking, promising, answering, in utterance
  - *Protesting*
- Perlocutionary act:
  - Production of effects on feeling, beliefs of addressee
  - *Intend to prevent doing some action*
- Types: assertives, directives, commissives, expressives, declarations

The 3 levels of act revisited

	Locutionary Force	Illocutionary Force	Perlocutionary Force
Can I have the rest of your sandwich?			

The 3 levels of act revisited

	Locutionary Force	Illocutionary Force	Perlocutionary Force
Can I have the rest of your sandwich?	Question		

The 3 levels of act revisited

	Locutionary Force	Illocutionary Force	Perlocutionary Force
Can I have the rest of your sandwich?	Question	Request	

The 3 levels of act revisited

	Locutionary Force	Illocutionary Force	Perlocutionary Force
Can I have the rest of your sandwich?	Question	Request	Intent: You give me sandwich

The 3 levels of act revisited

	Locutionary Force	Illocutionary Force	Perlocutionary Force
Can I have the rest of your sandwich?	Question	Request	Intent: You give me sandwich
I want the rest of your sandwich			

The 3 levels of act revisited

	Locutionary Force	Illocutionary Force	Perlocutionary Force
Can I have the rest of your sandwich?	Question	Request	Intent: You give me sandwich
I want the rest of your sandwich	Declarative	Request	Intent: You give me sandwich
Give me your sandwich!			

The 3 levels of act revisited

	Locutionary Force	Illocutionary Force	Perlocutionary Force
Can I have the rest of your sandwich?	Question	Request	Intent: You give me sandwich
I want the rest of your sandwich	Declarative	Request	Intent: You give me sandwich
Give me your sandwich!	Imperative	Request	Intent: You give me sandwich

Collaborative Communication

- Speaker tries to establish and add to
  - “common ground” – “mutual belief”

Collaborative Communication

- Speaker tries to establish and add to
  - "common ground" – "mutual belief"

  - Presumed a joint, collaborative activity
    - Make sure "mutually believe" the same thing

Collaborative Communication

- Speaker tries to establish and add to
  - “common ground” – “mutual belief”

  - Presumed a joint, collaborative activity
    - Make sure “mutually believe” the same thing

  - Hearer must ‘ground’ speaker’s utterances
    - Indicate heard and understood

Closure

- Principle of closure:
  - Agents performing an action require evidence of successful performance
    - Also important to indicate failure or understanding

Closure

- Principle of closure:
  - Agents performing an action require evidence of successful performance
    - Also important to indicate failure or understanding
  - Non-speech closure:

Closure

- Principle of closure:
  - Agents performing an action require evidence of successful performance
    - Also important to indicate failure or understanding
  - Non-speech closure:
    - Push elevator button → Light turns on

Closure

- Principle of closure:
  - Agents performing an action require evidence of successful performance
    - Also important to indicate failure or understanding
  - Non-speech closure:
    - Push elevator button → Light turns on
  - Two step process:
    - Presentation (speaker)
    - Acceptance (listener)

Degrees of Grounding

- Weakest to strongest

Degrees of Grounding

- Weakest to strongest
- Continued attention

Degrees of Grounding

- Weakest to strongest
- Continued attention:
- Next relevant contribution

Degrees of Grounding

- Weakest to strongest
- Continued attention:
- Next relevant contribution
- Acknowledgment:
  - Minimal response, continuer: *yeah, uh-huh, okay; great*

Degrees of Grounding

- Weakest to strongest
- Continued attention:
- Next relevant contribution
- Acknowledgment:
  - Minimal response, continuer: *yeah, uh-huh, okay; great*
- Demonstrate:
  - Indicate understanding by reformulation, completion

Degrees of Grounding

- Weakest to strongest
- Continued attention:
- Next relevant contribution
- Acknowledgment:
  - Minimal response, continuer: *yeah, uh-huh, okay; great*
- Demonstrate:
  - Indicate understanding by reformulation, completion
- Display:
  - Repeat all or part

Dialog Example

C_1:	...I need to travel in May.
A_1:	And, what day in May did you want to travel?
C_2:	OK uh I need to be there for a meeting that's from the 12th to the 15th.
A_2:	And you're flying into what city?
C_3:	Seattle.
A_3:	And what time would you like to leave Pittsburgh?
C_4:	Uh hmm I don't think there's many options for non-stop.
A_4:	Right. There's three non-stops today.
C_5:	What are they?
A_5:	The first one departs PGH at 10:00am arrives Seattle at 12:05 their time. The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
C_6:	OK I'll take the 5ish flight on the night before on the 11th.
A_6:	On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
C_7:	OK.

Grounding

- Display:
  - C: I need to travel in May.
  - A: And *what day* in May did you want to travel?

Grounding

- Display:
  - C: I need to travel in May.
  - A: And *what day* in May did you want to travel?

- Acknowledgment + Next relevant contribution:
  - *And* what day in May did you want to travel?
  - *And* you are flying into what city?
  - *And* what time would you like to leave Pittsburgh?

Travel Planning

S1: Hello. You've reached the [Communicator]. Tell me your full name
U2: *hi i'd like to fly to seattle tuesday morning*
S3: Traveling to Seattle on Tuesday, August eleventh in the morning. Your full name?
U4: *[John Doe]*
S5: Hello, [John], I'm glad to hear from you again. You can now ask me for help.
S6: OK. Let's see I have a United flight departing Pittsburgh at ten forty five a.m. arriving into Seattle at one fifty seven p.m. Is that OK?
U7: *Yes*
S8: Will you return to Pittsburgh from Seattle?
U9: *no i will go to Los Angeles*
S10: Traveling to Los Angeles. on what date
U11: *Sorry can we go back to talk about the previous leg again*
S12: For the first leg, you are on a United flight 1446 departing Pittsburgh at ten forty five A.M. tomorrow, arriving in Seattle at one fifty seven P.M. Do you want to take that?
U13: *actually i want to leave on wednesday*
S14: a flight on wednesday.
S15: I have a flight departing Pittsburgh at ten forty five a.m., arrives Seattle at one fifty seven p.m. Is that OK?
U16: *Yes*

Figure 19.1 The travel domain: a fragment from a successful conversation between a user (U) and the Communicator system (S) of Xu and Rudnicky (2000).

Grounding in HCI

- Key factor in HCI:
  - Users confused if system fails to ground, confirm
    - (Stifelman et al., 1993), (Yankelovich et al, 1995)

    - S: Did you want to review some more of your profile?
    - U: No.
    - S: What's next?

Grounding in HCI

- Key factor in HCI:
  - Users confused if system fails to ground, confirm
    - (Stifelman et al., 1993), (Yankelovich et al, 1995)

    - S: Did you want to review some more of your profile?
    - U: No.
    - S: What's next?

    - S: Did you want to review some more of your profile?
    - U: No.
    - S: Okay, what's next?

Conversational Implicature

- Meaning more than just literal contribution

  - *A: And, what day in May did you want to travel?*
  - *C: OK uh I need to be there for a meeting the 12-15th*
    - Appropriate?

Conversational Implicature

- Meaning more than just literal contribution

  - *A: And, what day in May did you want to travel?*
  - *C: OK uh I need to be there for a meeting the 12-15th*
    - Appropriate? Yes
    - Why?

Conversational Implicature

- Meaning more than just literal contribution
  - *A: And, what day in May did you want to travel?*
  - *C: OK uh I need to be there for a meeting the 12-15th*
    - Appropriate? Yes
    - Why?
  - Inference required

Grice's Maxims

- Cooperative principle:
  - Tacit agreement b/t conversants to cooperate

Grice's Maxims

- Cooperative principle:
  - Tacit agreement b/t conversants to cooperate
- Grice's Maxims
  - Quantity: Be as informative as required

Grice's Maxims

- Cooperative principle:
  - Tacit agreement b/t conversants to cooperate
- Grice's Maxims
  - Quantity: Be as informative as required
  - Quality: Be truthful
    - Don't lie, or say things without evidence

Grice's Maxims

- Cooperative principle:
  - Tacit agreement b/t conversants to cooperate
- Grice's Maxims
  - Quantity: Be as informative as required
  - Quality: Be truthful
    - Don't lie, or say things without evidence
  - Relevance: Be relevant
  - Manner: "Be perspicuous"
    - Don't be obscure, ambiguous, prolix, or disorderly

Relevance

- Client: **I need to be there for a meeting that's from the 12th to the 15th**
  - Hearer thinks:

Relevance

- Client: **I need to be there for a meeting that's from the 12th to the 15th**
  - Hearer thinks: **Speaker is following maxims, would only have mentioned meeting if it was relevant. How could meeting be relevant? If client meant me to understand that he had to depart in time for the mtg.**

Quantity

- A: How much money do you have on you?
- B: I have 5 dollars
  - Implication

Quantity

- A: How much money do you have on you?
- B: I have 5 dollars
  - Implication: not 6 dollars
- A: Did you do the reading for today's class?
- B: I intended to
  - Implication:

Quantity

- A: How much money do you have on you?
- B: I have 5 dollars
  - Implication: not 6 dollars
- A: Did you do the reading for today's class?
- B: I intended to
  - Implication: No
  - B's answer would be true if B intended to do the reading AND did the reading, but would then violate maxim

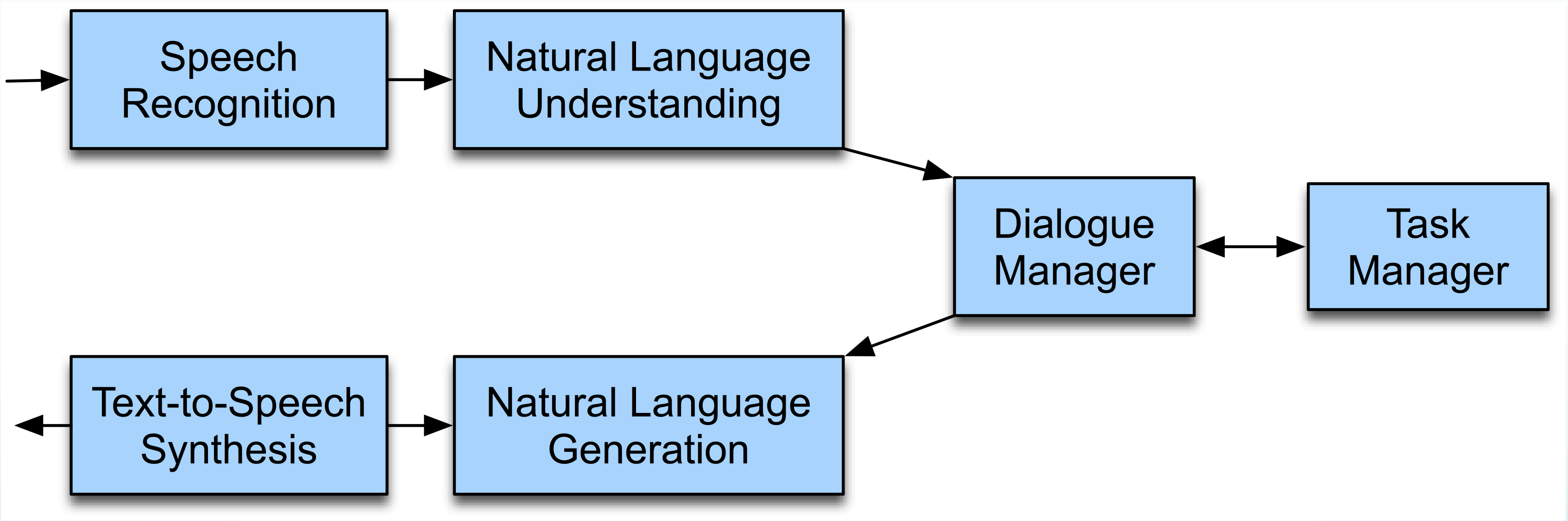
From Human to Computer

- Conversational agents
  - Systems that (try to) participate in dialogues
  - Examples: Directory assistance, travel info, weather, restaurant and navigation info
- Issues:

From Human to Computer

- Conversational agents
  - Systems that (try to) participate in dialogues
  - Examples: Directory assistance, travel info, weather, restaurant and navigation info
- Issues:
  - Limited understanding: ASR errors, interpretation
  - Computational costs

Dialogue System Architecture



Speech Recognition

- (aka ASR)
- Input: acoustic waveform
  - Telephone, microphone, and smartphone

Speech Recognition

- (aka ASR)
- Input: acoustic waveform
  - Telephone, microphone, and smartphone
- Output: recognized word string

Speech Recognition

- (aka ASR)
- Input: acoustic waveform
  - Telephone, microphone, and smartphone
- Output: recognized word string
- Requirements:

Speech Recognition

- (aka ASR)
- Input: acoustic waveform
  - Telephone, microphone, and smartphone
- Output: recognized word string
- Requirements:
  - Acoustic models: map acoustics to phone [ae] [k]
  - Pronunciation dictionary: words to phones: cat: [k][ae][t]
  - Grammar: legal word sequences
  - Search procedure: best word sequence given audio

Recognition in SDS

Recognition in SDS

- Create domain specific vocabulary, grammar
  - Typically hand-crafted in most commercial systems
  - Based on human-human interactions
  - Grammars: finite-state, context-free, language model

Recognition in SDS

- Create domain specific vocabulary, grammar
  - Typically hand-crafted in most commercial systems
  - Based on human-human interactions
  - Grammars: finite-state, context-free, language model

- Activate only portion of grammar based on dialog state
  - E.g. Where are you leaving from?

Recognition in SDS

- Create domain specific vocabulary, grammar
  - Typically hand-crafted in most commercial systems
  - Based on human-human interactions
  - Grammars: finite-state, context-free, language model

- Activate only portion of grammar based on dialog state
  - E.g. Where are you leaving from?
  - {I want to (leave|depart) from} CITYNAME {STATENAME}
  - ‘Yes/No’ grammar for confirmations

Natural Language Understanding

- Most systems use frame-slot semantics
  Show me morning flights from Boston to SFO on Tuesday
  Alternatives:
    - Full parser with semantic attachments
    - Domain-specific analyzers
- SHOW:
- FLIGHTS:
  - ORIGIN:
    - CITY: Boston
    - DATE:
      - DAY-OF-WEEK: Tuesday
    - TIME:
      - PART-OF-DAY: Morning
  - DEST:
    - CITY: San Francisco

Generation and TTS

- Generation:
  - Identify concepts to express
  - Convert to words
  - Assign appropriate prosody, intonation

Generation and TTS

- Generation:
  - Identify concepts to express
  - Convert to words
  - Assign appropriate prosody, intonation
- TTS:
  - Input words, prosodic markup
  - Synthesize acoustic waveform

Generation

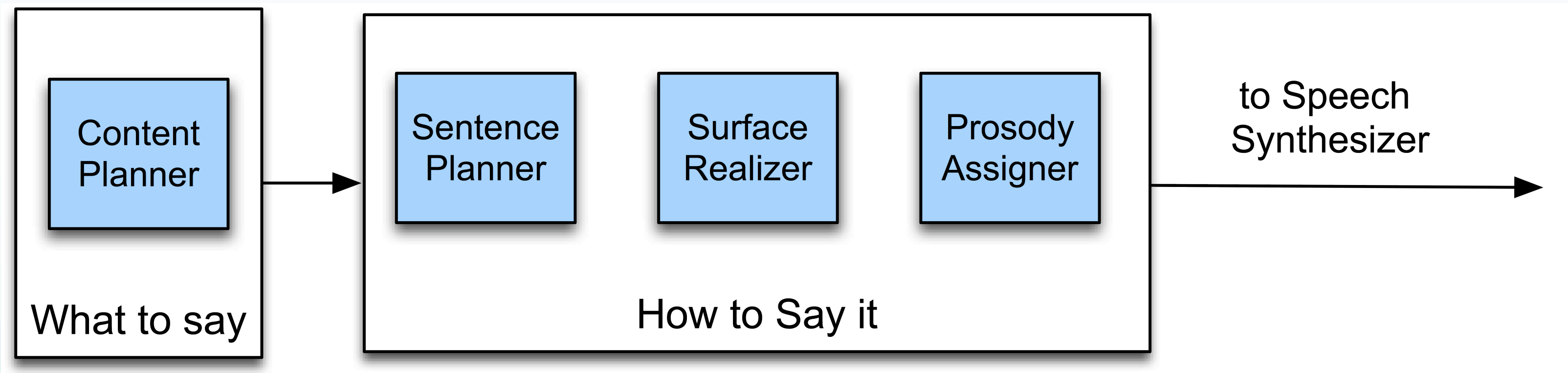
- Content planning:
  - What to say:
    - Question, answer, etc?
    - Often merged with dialog manager

Generation

- Content planning:
  - What to say:
    - Question, answer, etc?
    - Often merged with dialog manager
- Language generation:
  - How to say it
    - Select syntactic structure and words
    - Most common: Template-based generation (prompts)
      - Templates with variable: When do you want to leave CITY?

Full NLG

- Converts representation from dialog manager



Dialogue Manager

- Holds system together: Governs interaction style

Dialogue Manager

- Holds system together: Governs interaction style
  - Takes input from ASR/NLU

Dialogue Manager

- Holds system together: Governs interaction style
  - Takes input from ASR/NLU

  - Maintains dialog state, history
    - Incremental frame construction
    - Reference, ellipsis resolution
    - Determines what system does next

Dialogue Manager

- Holds system together: Governs interaction style
  - Takes input from ASR/NLU
  - Maintains dialog state, history
    - Incremental frame construction
    - Reference, ellipsis resolution
    - Determines what system does next
  - Interfaces with task manager/backend app

Dialogue Manager

- Holds system together: Governs interaction style
  - Takes input from ASR/NLU
  - Maintains dialog state, history
    - Incremental frame construction
    - Reference, ellipsis resolution
    - Determines what system does next
  - Interfaces with task manager/backend app
  - Formulates basic response, passes to NLG,TTS

Dialog Management Types

- Finite-State Dialog Management
- Frame-based Dialog Management
- Information State Manager
- Statistical Dialog Management

Designing Dialog

- Apply user-centered design

Designing Dialog

- Apply user-centered design
  - Study user and task: How?

Designing Dialog

- Apply user-centered design
  - Study user and task: How?
    - Interview potential users, record human-human tasks
  - Study how the user interacts with the system

Designing Dialog

- Apply user-centered design
  - Study user and task: How?
    - Interview potential users, record human-human tasks
  - Study how the user interacts with the system
    - But it’s not built yet….

Designing Dialog

- Apply user-centered design
  - Study user and task: How?
    - Interview potential users, record human-human tasks
  - Study how the user interacts with the system
    - But it's not built yet....
      - Wizard-of-Oz systems: Simulations
        - User thinks they're interacting with a system, but it's driven by a human
      - Prototypes

Designing Dialog

- Apply user-centered design
  - Study user and task: How?
    - Interview potential users, record human-human tasks
  - Study how the user interacts with the system
    - But it's not built yet....
      - Wizard-of-Oz systems: Simulations
        - User thinks they're interacting with a system, but it's driven by a human
      - Prototypes
  - Iterative redesign:
    - Test system: see how users really react, what problems occur, correct, repeat

SDS Evaluation

- Goal: Determine overall user satisfaction
  - Highlight systems problems; help tune

SDS Evaluation

- Goal: Determine overall user satisfaction
  - Highlight systems problems; help tune
- Classically: Conduct user surveys

SDS Evaluation

- Goal: Determine overall user satisfaction
  - Highlight systems problems; help tune
- Classically: Conduct user surveys

TTS Performance	Was the system easy to understand ?
ASR Performance	Did the system understand what you said?
Task Ease	Was it easy to find the message/flight/train you wanted?
Interaction Pace	Was the pace of interaction with the system appropriate?
User Expertise	Did you know what you could say at each point?
System Response	How often was the system sluggish and slow to reply to you?
Expected Behavior	Did the system work the way you expected it to?
Future Use	Do you think you'd use the system in the future?

Figure 24.14 User satisfaction survey, adapted from Walker et al. (2001).

SDS Evaluation

- User evaluation issues:

SDS Evaluation

- User evaluation issues:
  - Expensive; often unrealistic; hard to get real user to do
- Create model correlated with human satisfaction
- Criteria:

SDS Evaluation

- User evaluation issues:
  - Expensive; often unrealistic; hard to get real user to do
- Create model correlated with human satisfaction
- Criteria:
  - Maximize task success
    - Measure task completion: % subgoals; Kappa of frame values
  - Minimize task costs
    - Efficiency costs: time elapsed; # turns; # error correction turns
    - Quality costs: # rejections; # barge-in; concept error rate

PARADISE Model

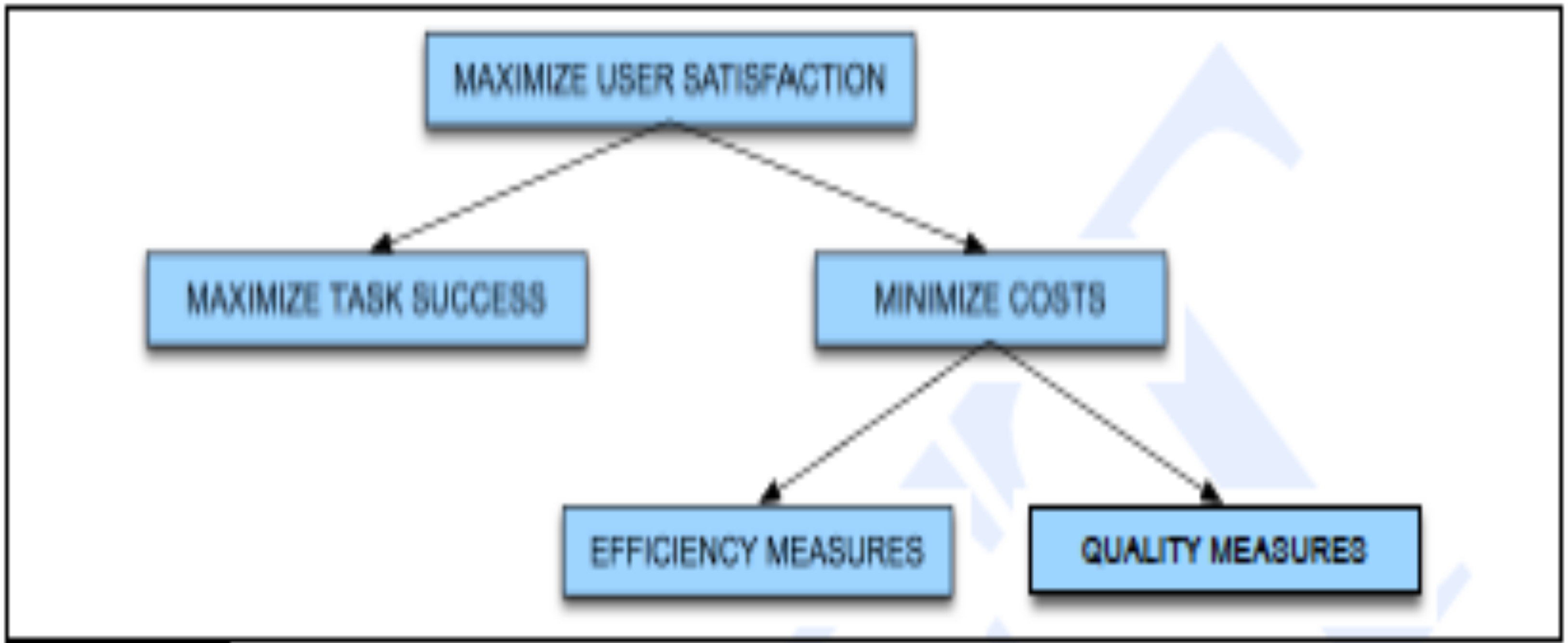


Figure 24.15 PARADISE's structure of objectives for spoken dialogue performance. After Walker et al. (1997).

PARADISE Model

- Compute user satisfaction with questionnaires
- Extract task success and costs measures from corresponding dialogs
  - Automatically or manually
- Perform multiple regression:
  - Assign weights to all factors of contribution to Usat
  - Task success, Concept accuracy key
- Allows prediction of accuracy on new dialog

Summary

- Spoken Dialogue Systems:
  - Build on existing text-based NLP techniques, but
  - Incorporate dialogue specific factors:
    - Turn-taking, grounding, dialogue acts
  - Affected by computational and modal constraints
    - Recognition errors, processing speed, etc.
    - Speech transience, slowness
  - Becoming more widespread and more flexible

Components: ASR

Drawing heavily on resource slides from Speech and Language Processing, Jurafsky and Martin

Speech Recognition

- Applications of Speech Recognition (ASR)
  - Dictation
  - Telephone-based Information (directions, air travel, banking, etc)
  - Hands-free (in car)
  - Speaker Identification
  - Language Identification
  - Second language ('L2') (accent reduction)
  - Audio archive searching

LVCSR

- Large Vocabulary Continuous Speech Recognition
- ~20,000-64,000 words
- Speaker independent (vs. speaker-dependent)
- Continuous speech (vs isolated-word)

Current error rates

Ballpark numbers; exact numbers depend very much on the specific corpus

Task	Vocabulary	Error Rate%
Digits	11	0.5
WSJ read speech	20K	3
Broadcast news	64,000+	10
CTS SWBD (GMM) 300hrs	64,000+	23-27
CTS SWBD (DNN) 300hrs	64,000+	11-15
CTS SWBD (GMM) >1000hr	64,000+	17-18
CTS SWBD (DNN) >>1000hr	64,000+	5.9
Google Voice > 5800hrs		12
YouTube > 1,400hrs		47

(some results Hinton 2012)

HSR versus ASR

Task	Vocab	ASR	Hum SR
Continuous digits	11	.5	.009
WSJ 1995 clean	5K	3	0.9
WSJ 1995 w/noise	5K	9	1.1
SWBD 2004	65K	5.9	4

- Conclusions:
  - Machines about 5 times worse than humans
  - Gap increases with noisy speech
  - These numbers are rough, take with grain of salt

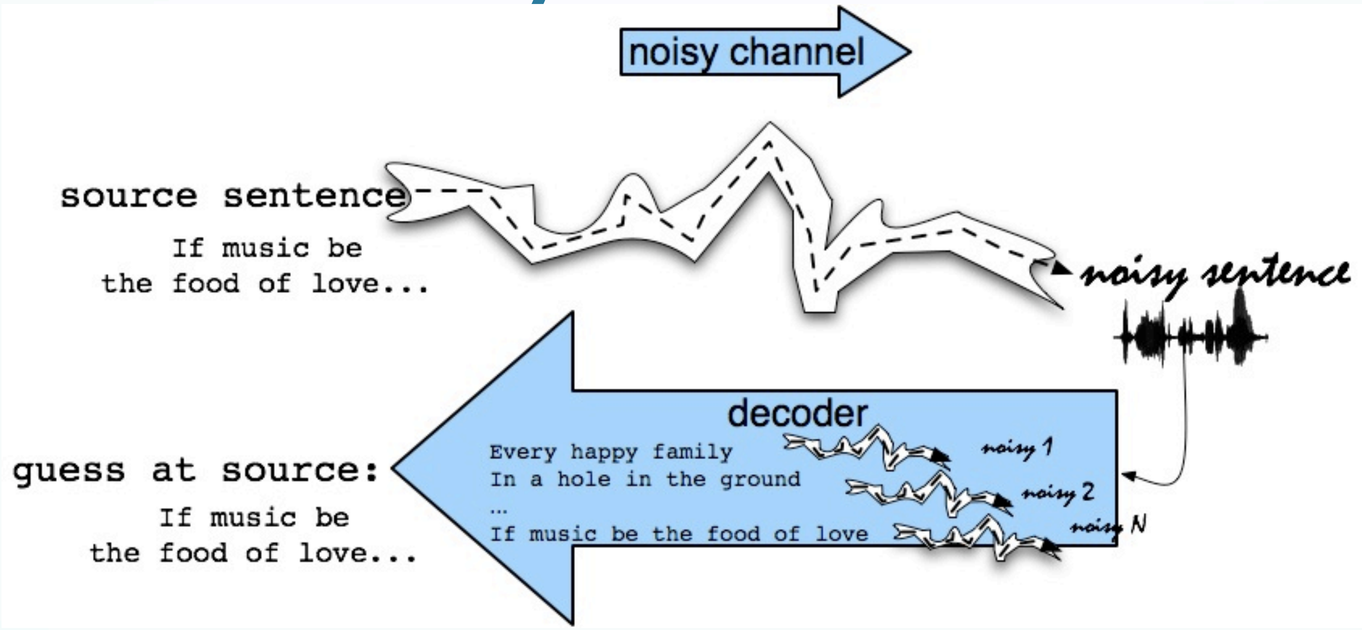
Why is conversational speech harder?

- A piece of an utterance without context
- The same utterance with more context

LVCSR Design Intuition

- Build a statistical model of the speech-to-words process
- Collect lots and lots of speech, and transcribe all the words.
- Train the model on the labeled speech
- Paradigm: Supervised Machine Learning + Search

The Noisy Channel Model



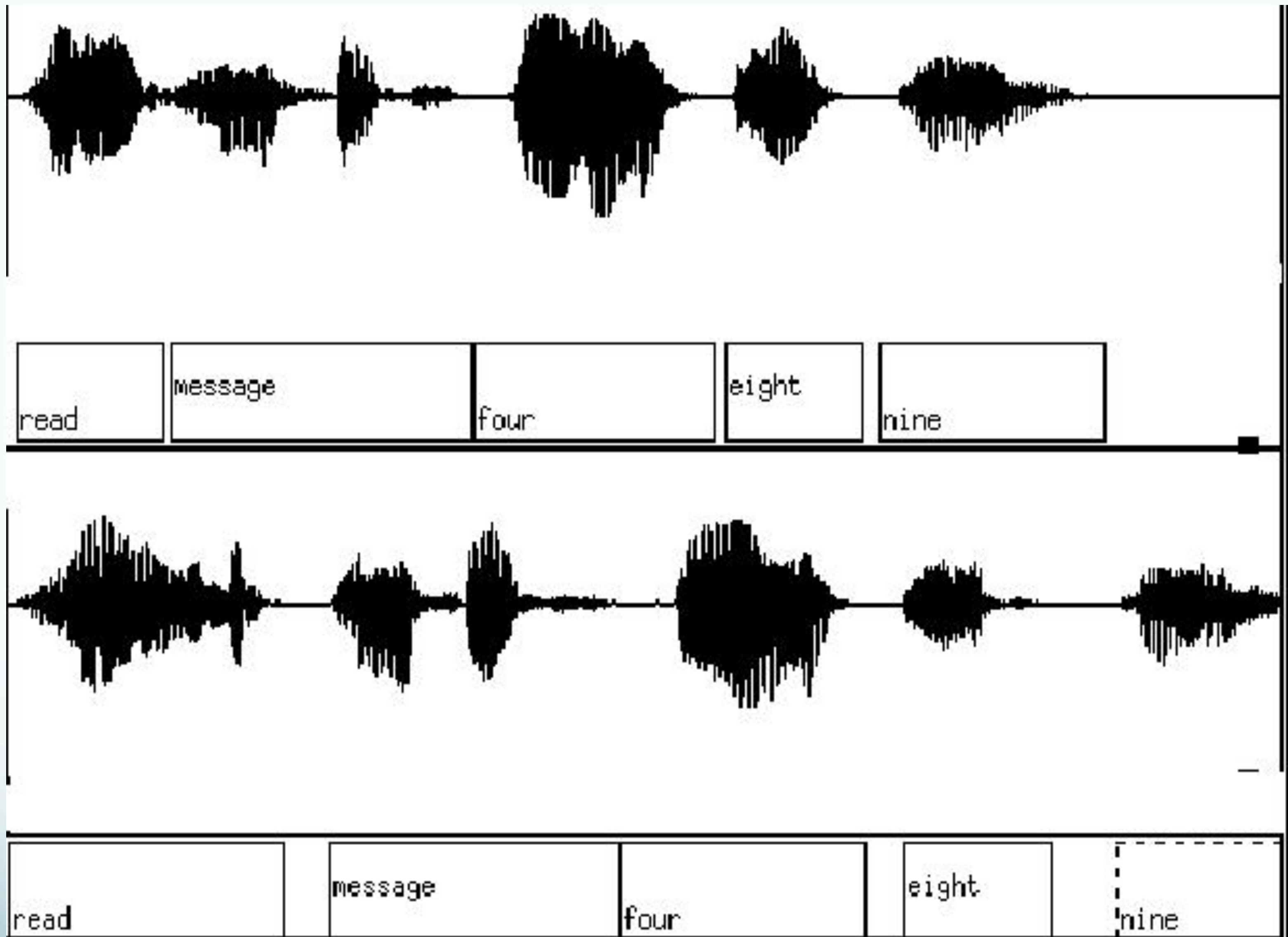
- Search through space of all possible sentences.
- Pick the one that is most probable given the waveform.

Decomposing Speech Recognition

- Q1: What speech sounds were uttered?
  - Human languages: 40-50 phones
    - Basic sound units: b, m, k, ax, ey, … (arpabet)
    - Distinctions categorical to speakers
      - Acoustically continuous
    - Part of knowledge of language
      - Build per-language inventory

Decomposing Speech Recognition

- Q2: What words produced these sounds?
  - Look up sound sequences in dictionary
  - Problem 1: Homophones
    - Two words, same sounds: too, two
  - Problem 2: Segmentation
    - No “space” between words in continuous speech
    - “I scream”/”ice cream”, “Wreck a nice beach”/”Recognize speech”
- Q3: What meaning produced these words?
  - NLP (But that’s not all!)


read
message
four
eight
nine
read
message
four
eight
nine

The Noisy Channel Model (II)

- What is the most likely sentence out of all sentences in the language L given some acoustic input O?
- Treat acoustic input O as sequence of individual observations
  - $O = o_1,o_2,o_3,\ldots,o_t$
- Define a sentence as a sequence of words:
  - $W = w_1,w_2,w_3,\ldots,w_n$

Noisy Channel Model (III)

- Probabilistic implication: Pick the highest prob S = W:

$$\hat{W} = \underset{W \in L}{\arg\max}\, P(W \mid O)$$

- We can use Bayes rule to rewrite this:

$$\hat{W} = \underset{W \in L}{\arg\max}\, \frac{P(O \mid W)P(W)}{P(O)}$$

- Since denominator is the same for each candidate sentence W, we can ignore it for the argmax:

$$\hat{W} = \underset{W \in L}{\arg\max}\, P(O \mid W)P(W)$$

Noisy channel model

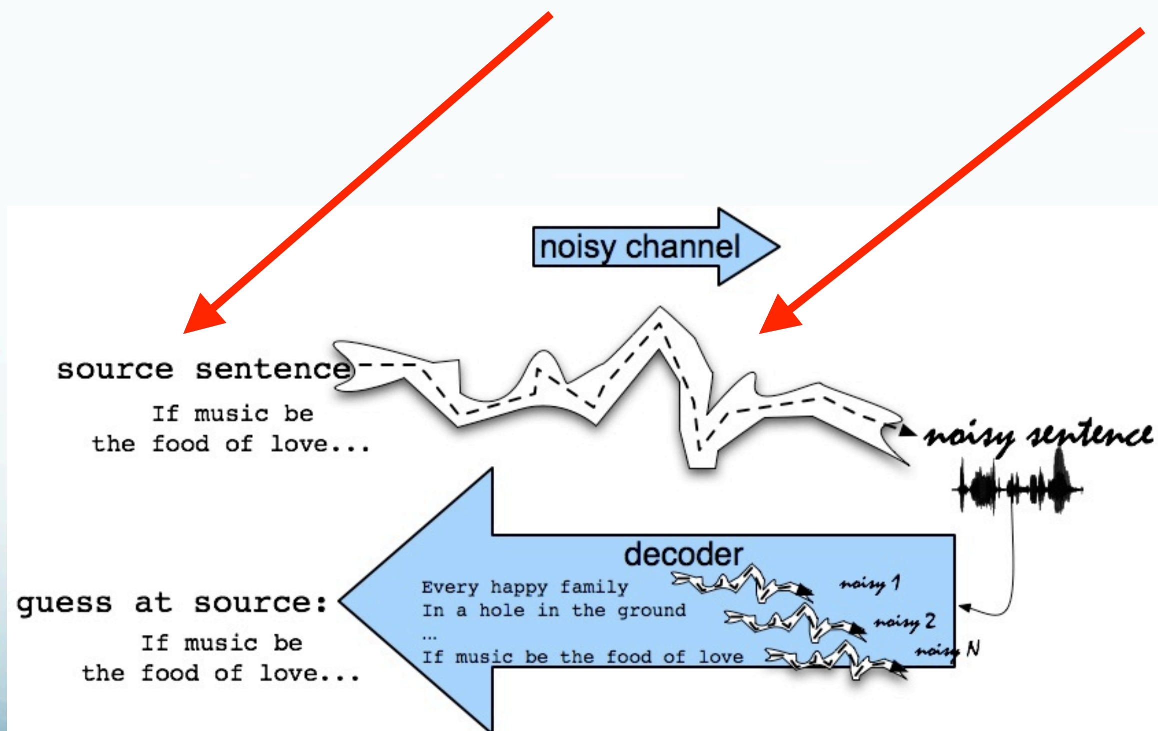
Acoustic Model likelihood → $P(O \mid W)$

Language Model prior → $P(W)$

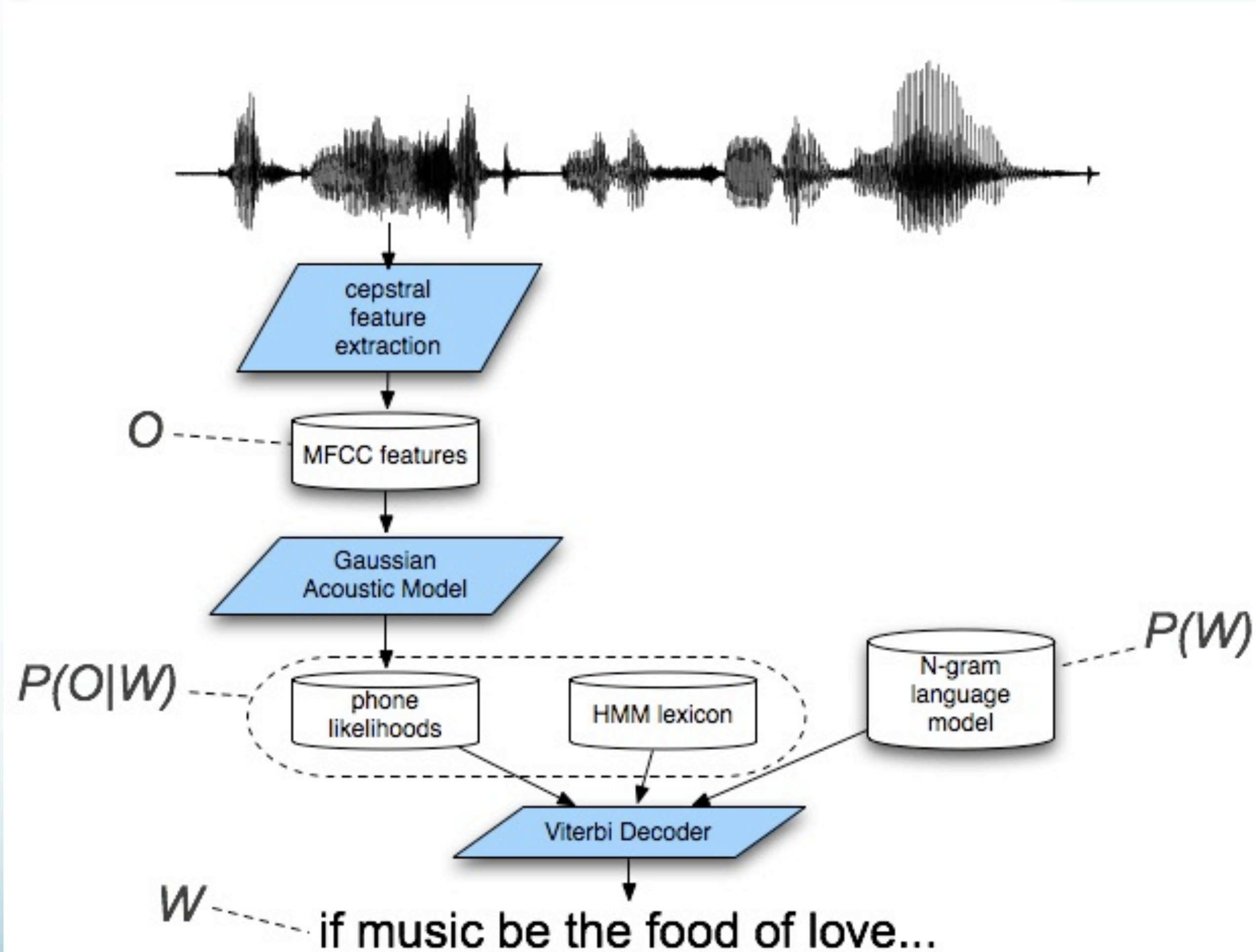
$$\hat{W} = \underset{W \in L}{\arg\max}\, P(O \mid W)\, P(W)$$

The noisy channel model

- Ignoring the denominator leaves us with two factors: P(Source) and P(Signal|Source)



Speech Recognition Architecture



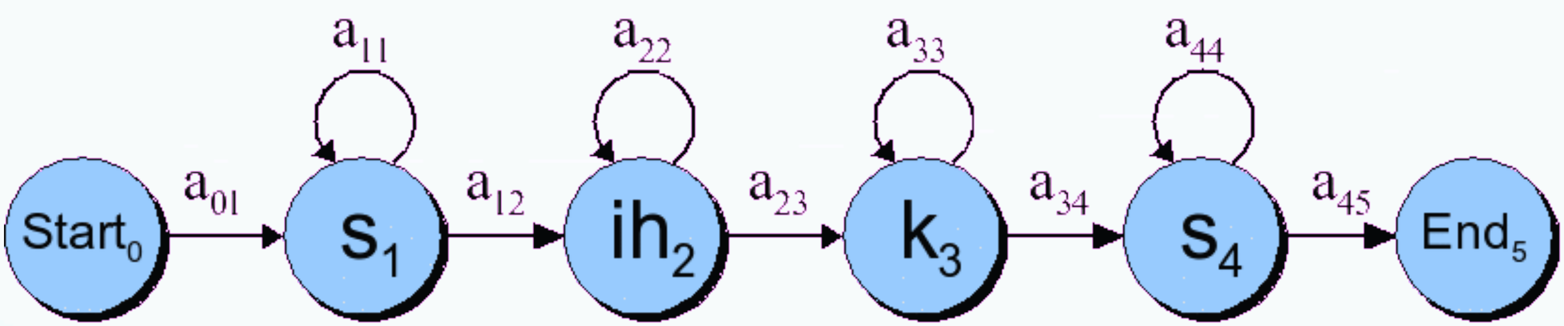
ASR Components

- Lexicons and Pronunciation:
  - Hidden Markov Models
- Feature extraction
- Acoustic Modeling
- Decoding
- Language Modeling:
  - Ngram Models

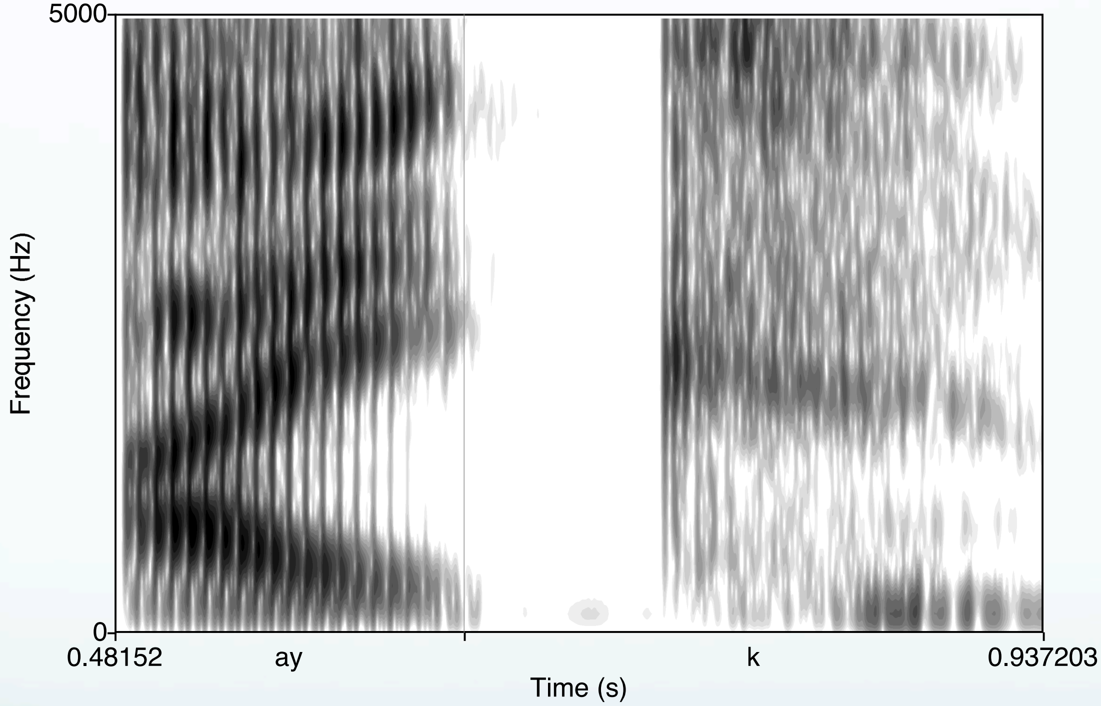
Lexicon

- A list of words
- Each one with a pronunciation in terms of phones
- We get these from on-line pronunciation dictionary
- CMU dictionary: 127K words
  - http://www.speech.cs.cmu.edu/cgi-bin/cmudict
- We'll represent the lexicon as an HMM

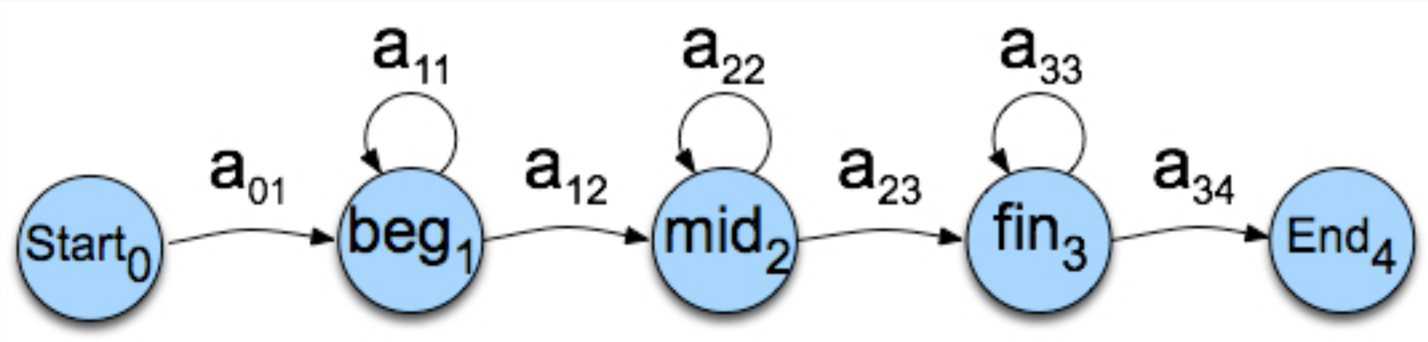
HMMs for speech: the word “six”



Phones are not homogeneous!

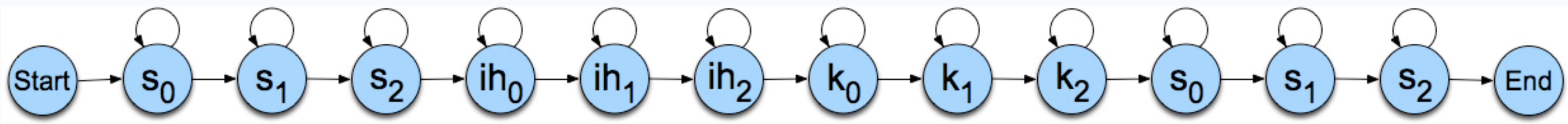


Each phone has 3 subphones



HMM word model for "six"

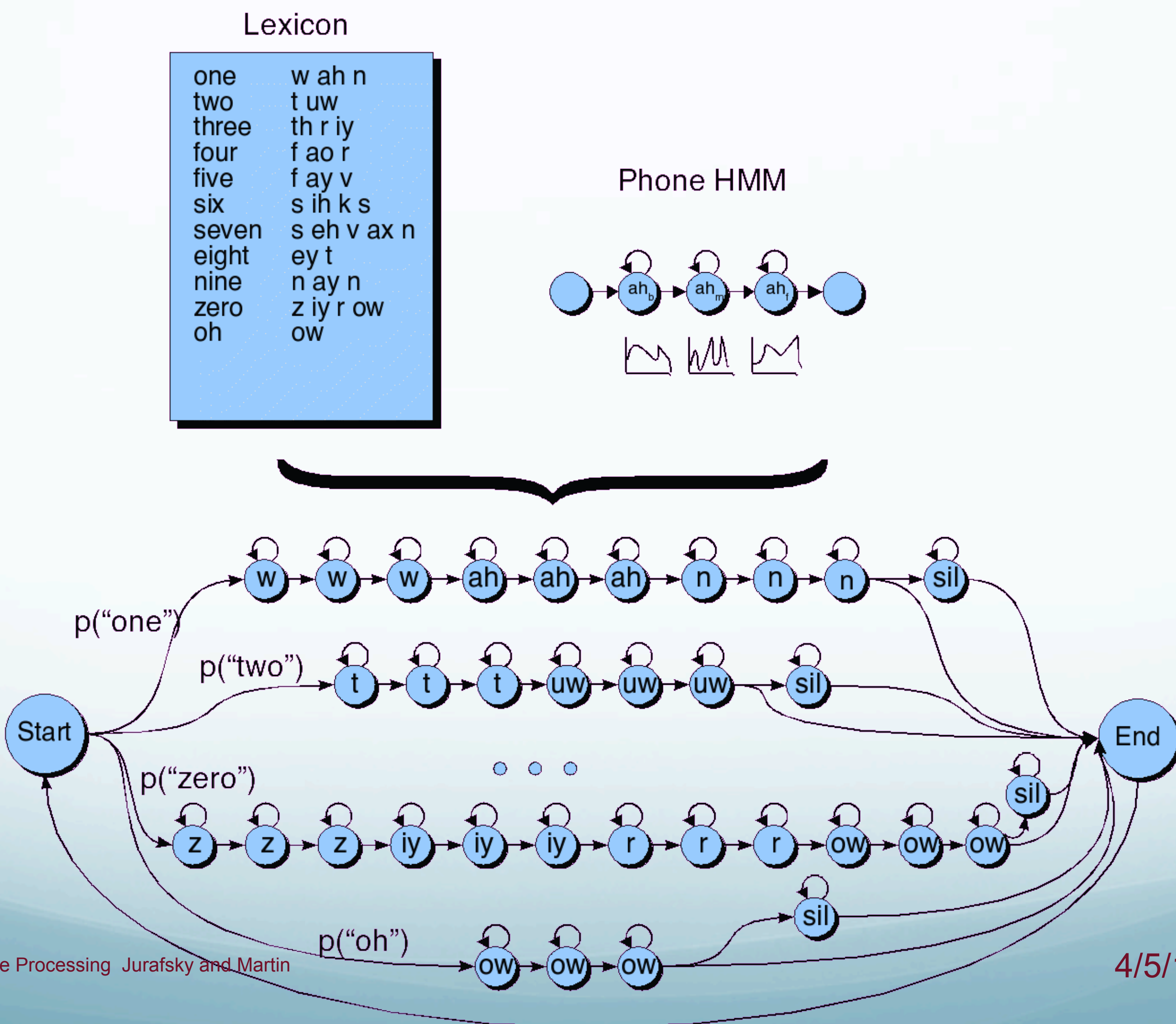
- Resulting model with subphones



HMMs for speech

$Q = q_1 q_2 \ldots q_N$	a set of **states** corresponding to **subphones**
$A = a_{01} a_{02} \ldots a_{n1} \ldots a_{nn}$	a **transition probability matrix** A, each a_{ij} representing the probability for each subphone of taking a **self-loop** or going to the next subphone. Together, Q and A implement a **pronunciation lexicon**, an HMM state graph structure for each word that the system is capable of recognizing.
$B = b_i(o_t)$	A set of **observation likelihoods:**, also called **emission probabilities**, each expressing the probability of a cepstral feature vector (observation o_t) being generated from subphone state i.

HMM for the digit recognition task



Typical MFCC features

- Window size: 25ms
- Window shift: 10ms
- Pre-emphasis coefficient: 0.97
- MFCC:
  - 12 MFCC (mel frequency cepstral coefficients)
  - 1 energy feature
  - 12 delta MFCC features
  - 12 double-delta MFCC features
  - 1 delta energy feature
  - 1 double-delta energy feature
- Total 39-dimensional features

Why is MFCC so popular?

- Efficient to compute
- Incorporates a perceptual Mel frequency scale
- Separates the source and filter
- Fits well with HMM modelling

Decoding

- In principle:

$$\hat{W} = \underset{W \in \mathcal{L}}{\operatorname{argmax}} \overbrace{P(O|W)}^{\text{likelihood}} \overbrace{P(W)}^{\text{prior}}$$

- In practice:

$$\hat{W} = \underset{W \in \mathcal{L}}{\operatorname{argmax}} P(O|W)P(W)^{LMSF}$$

$$\hat{W} = \underset{W \in \mathcal{L}}{\operatorname{argmax}} P(O|W)P(W)^{LMSF}\, WIP^{N}$$

$$\hat{W} = \underset{W \in \mathcal{L}}{\operatorname{argmax}} \log P(O|W) + LMSF \times \log P(W) + N \times \log WIP$$

Why is ASR decoding hard?

[ay d ih s hh er d s ah m th ih ng ax b aw m uh v ih ng r ih s en l ih]

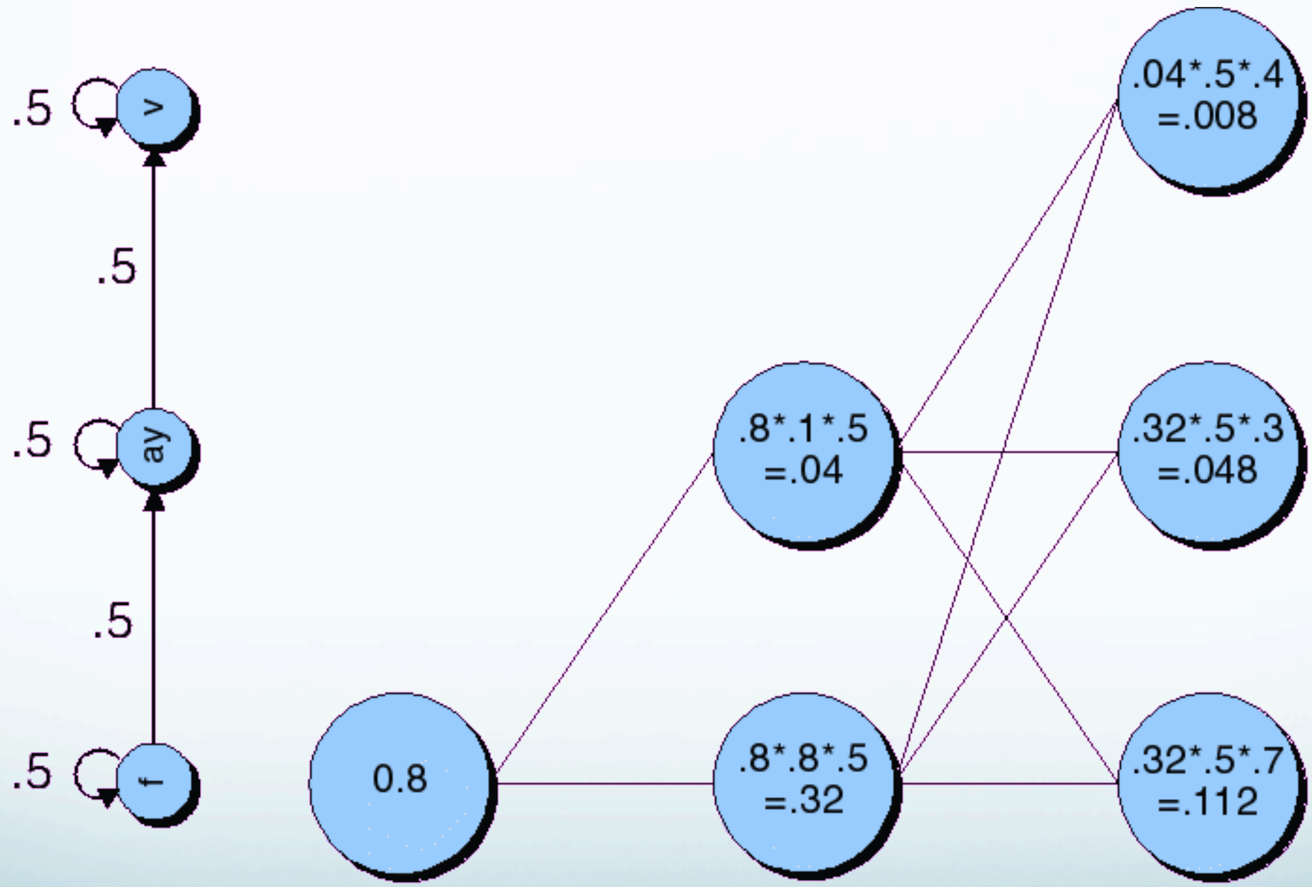
The Evaluation (forward) problem for speech

- The observation sequence O is a series of MFCC vectors
- The hidden states W are the phones and words
- For a given phone/word string W, our job is to evaluate $P(O|W)$
- Intuition: how likely is the input to have been generated by just that word string W

Evaluation for speech: Summing over all different paths!

- f ay ay ay ay v v v v
- f f ay ay ay ay v v v
- f f f f ay ay ay ay v
- f f ay ay ay ay ay ay v
- f f ay ay ay ay ay ay ay ay v
- f f ay v v v v v v v

Viterbi trellis for "five"



Viterbi trellis for "five"

V	0	0	0.008	0.0072	0.00672	0.00403	0.00188	0.00161	0.000667	0.000493
AY	0	0.04	0.048	0.0448	0.0269	0.0125	0.00538	0.00167	0.000428	8.78e-05
F	0.8	0.32	0.112	0.0224	0.00448	0.000896	0.000179	4.48e-05	1.12e-05	2.8e-06
Time	1	2	3	4	5	6	7	8	9	10
B	*f* 0.8	*f* 0.8	*f* 0.7	*f* 0.4	*f* 0.4	*f* 0.4	*f* 0.4	*f* 0.5	*f* 0.5	*f* 0.5
	ay 0.1	*ay* 0.1	*ay* 0.3	*ay* 0.8	*ay* 0.8	*ay* 0.8	*ay* 0.8	*ay* 0.6	*ay* 0.5	*ay* 0.4
	v 0.6	*v* 0.6	*v* 0.4	*v* 0.3	*v* 0.3	*v* 0.3	*v* 0.3	*v* 0.6	*v* 0.8	*v* 0.9
	p 0.4	*p* 0.4	*p* 0.2	*p* 0.1	*p* 0.1	*p* 0.1	*p* 0.1	*p* 0.1	*p* 0.3	*p* 0.3
	iy 0.1	*iy* 0.1	*iy* 0.3	*iy* 0.6	*iy* 0.6	*iy* 0.6	*iy* 0.6	*iy* 0.5	*iy* 0.5	*iy* 0.4

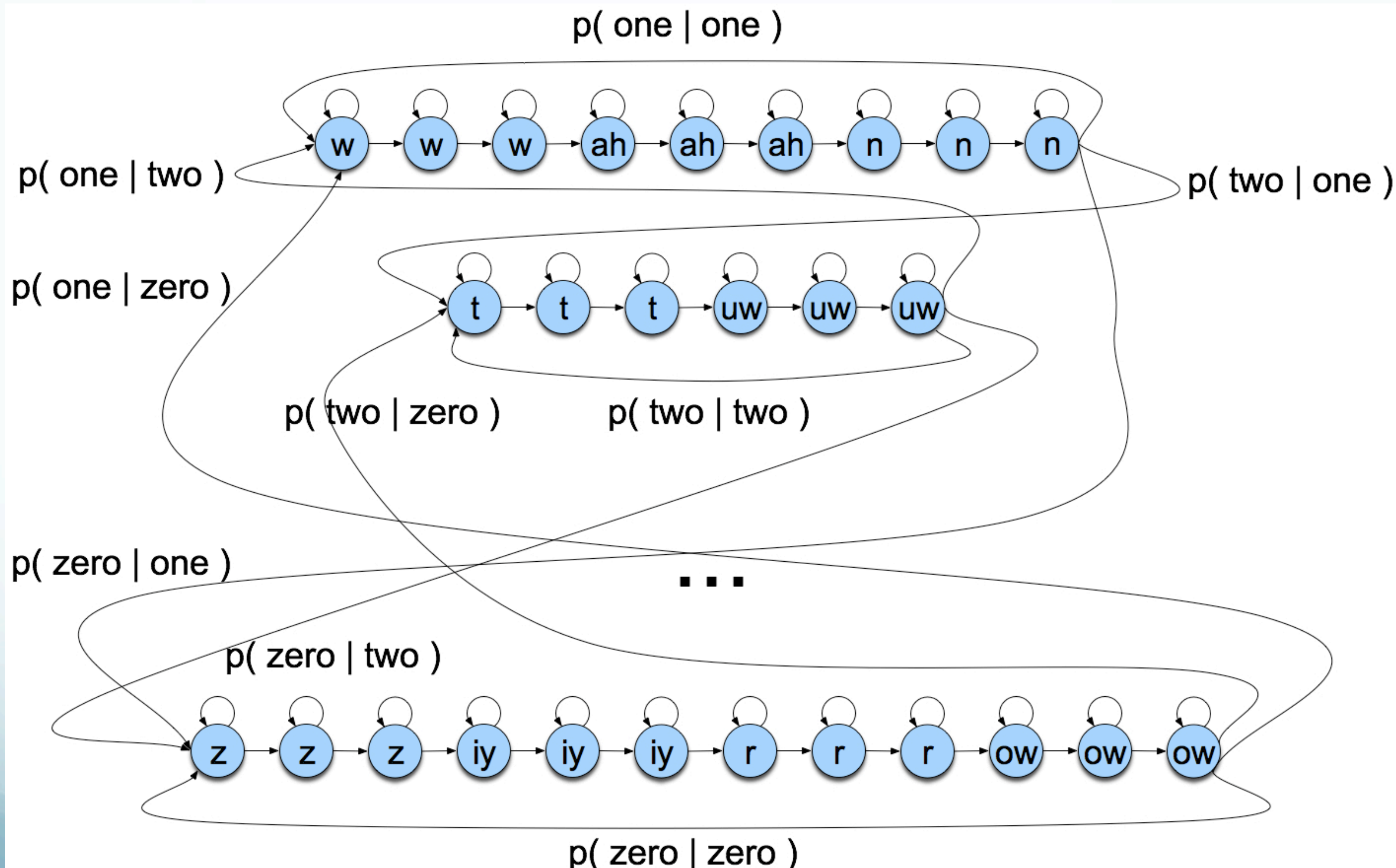
Language Model

- Idea: some utterances more probable

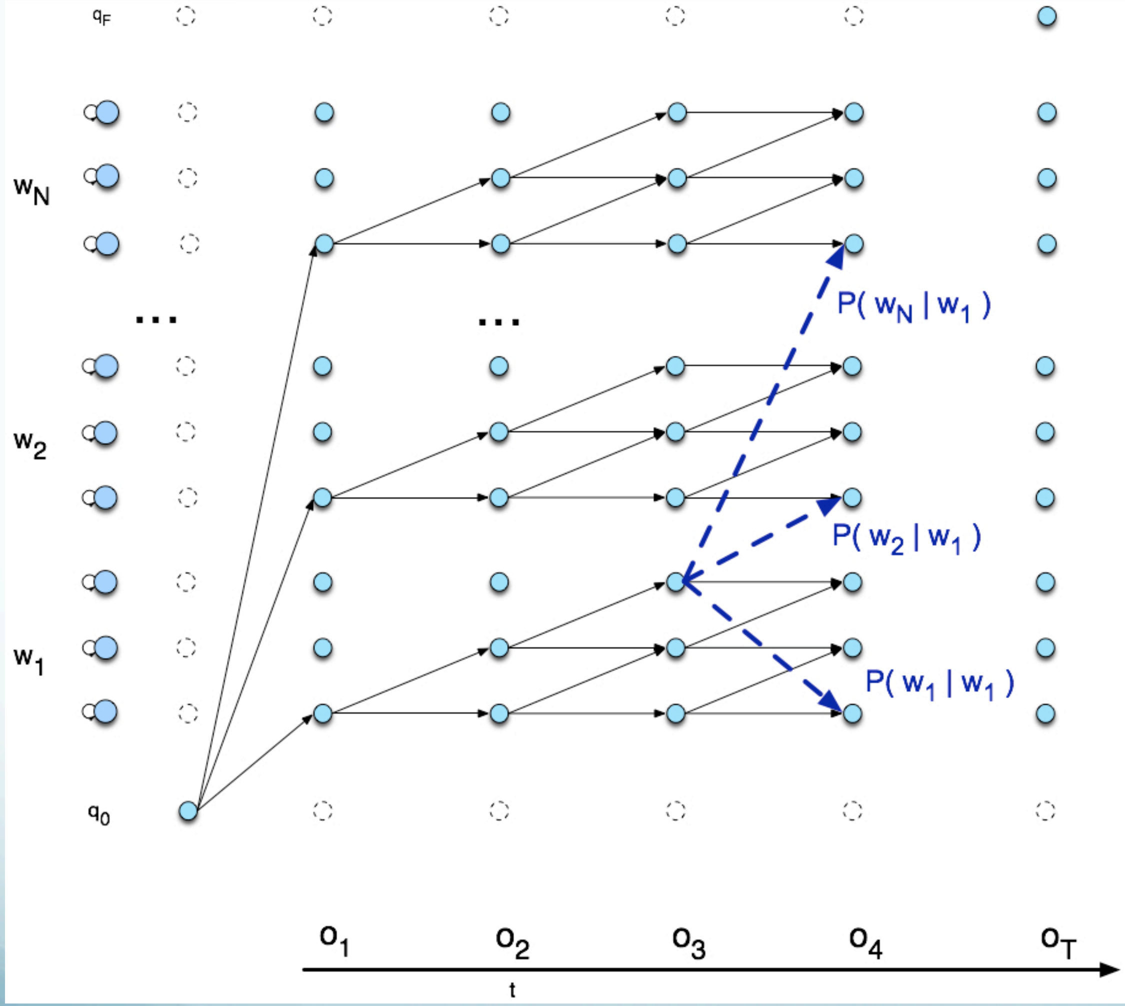
- Standard solution: "n-gram" model
  - Typically tri-gram: $P(w_i | w_{i-1}, w_{i-2})$
    - Collect training data from large side corpus
      - Smooth with bi- & uni-grams to handle sparseness
  - Product over words in utterance:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k \mid w_{k-1}, w_{k-2})$$

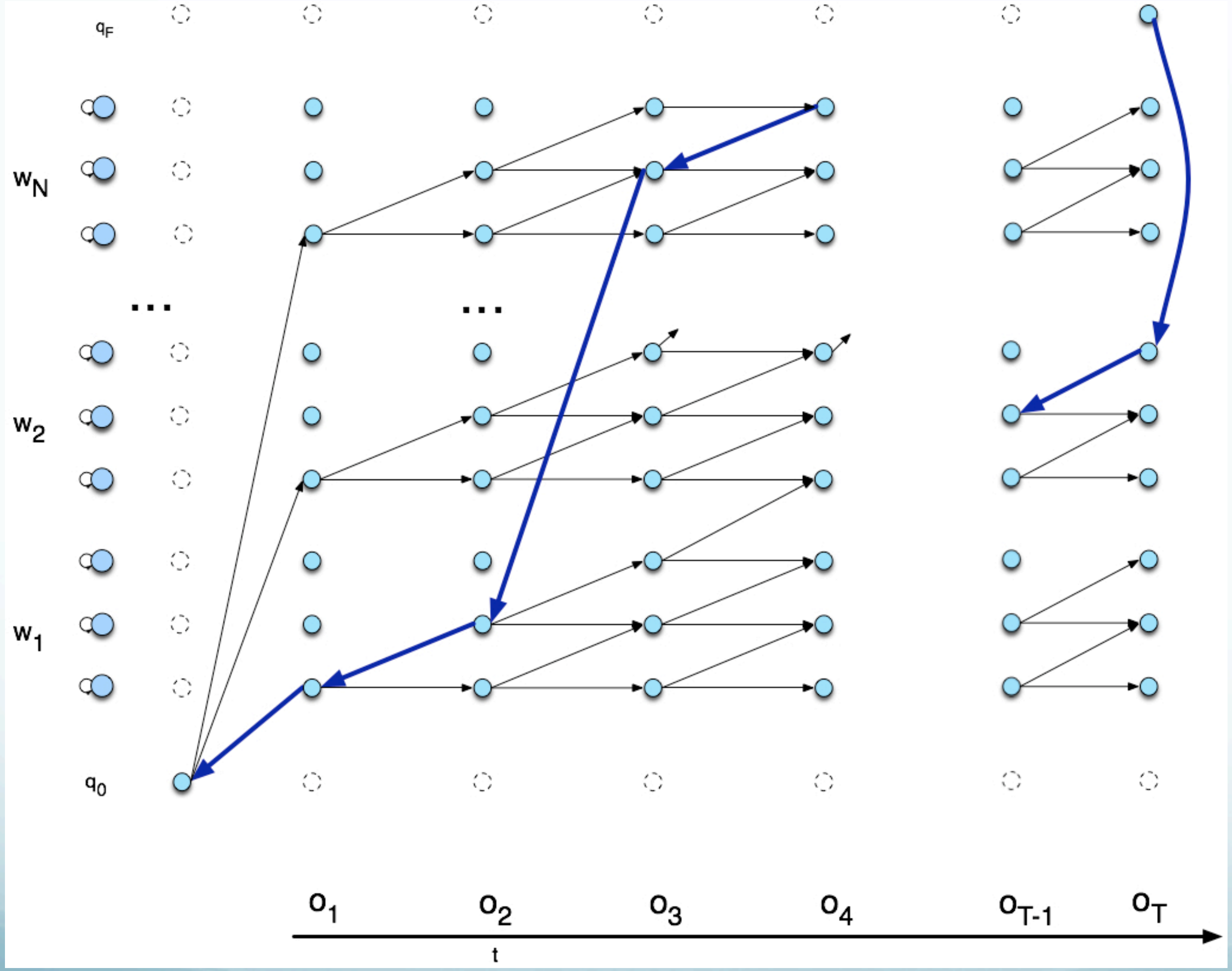
Search space with bigrams



Viterbi trellis

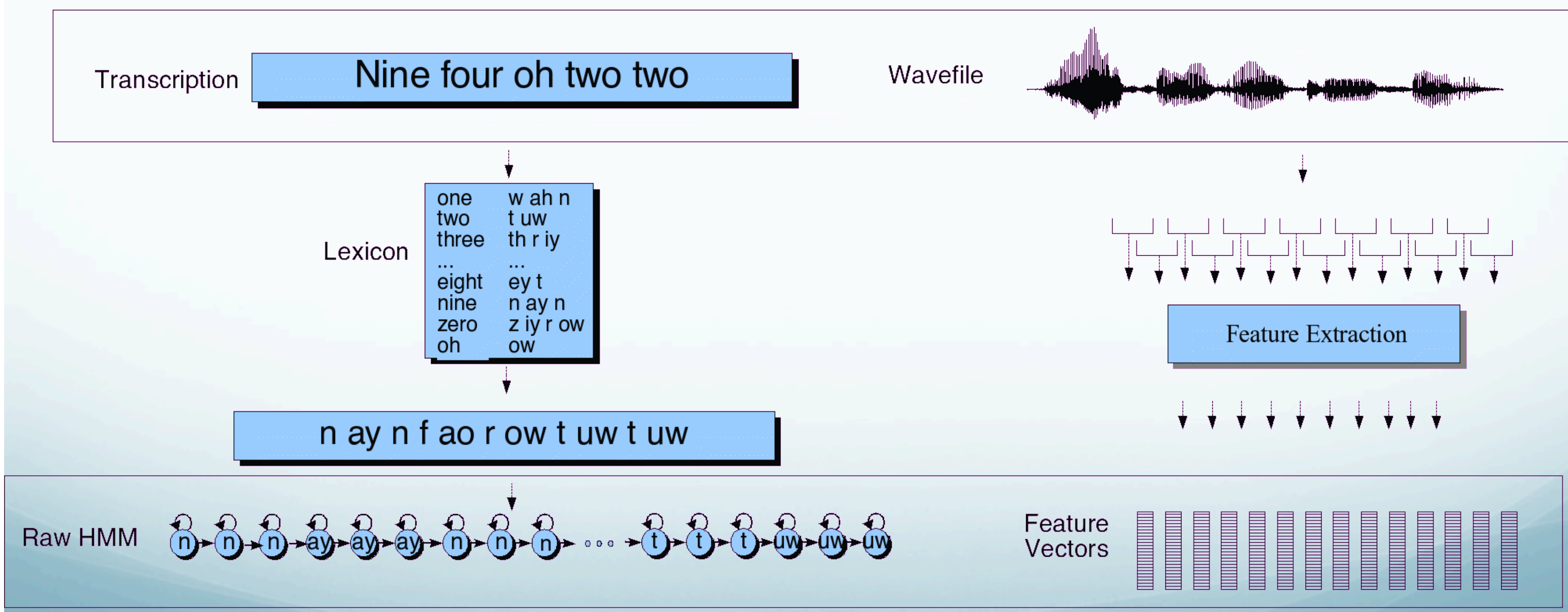


Viterbi backtrace



Training

- Trained using Baum-Welch algorithm



Summary: ASR Architecture

- Five easy pieces: ASR Noisy Channel architecture
  1) Feature Extraction:
     39 "MFCC" features
  2) Acoustic Model:
     Gaussians for computing $p(o|q)$
  3) Lexicon/Pronunciation Model
     - HMM: what phones can follow each other
  4) Language Model
     - N-grams for computing $p(w_i|w_{i-1})$
  5) Decoder
     - Viterbi algorithm: dynamic programming for combining all these to get word sequence from speech!

HW #1

- Automatic Speech Recognition
- Goals:
  - Gain familiarity with the Kaldi ASR system
  - Build a basic digit recognizer
  - Investigate training/tuning conditions
  - Evaluate a system

Tasks

- Create a kaldi working directory/environment
- Train and run system under different conditions
- Write short report to analyze and compare results
- Due Tuesday 4/11

Specialized Topics

- Everyone will lead discussion of a special topic
  - 1-2 people per topic
  - Brief summary
  - Critique
  - Discussion
- Topics will be posted shortly
  - Based KWLA responses and field
- Reply to GoPost with preferred topic(s)