# Statistical Speech and Language Processing

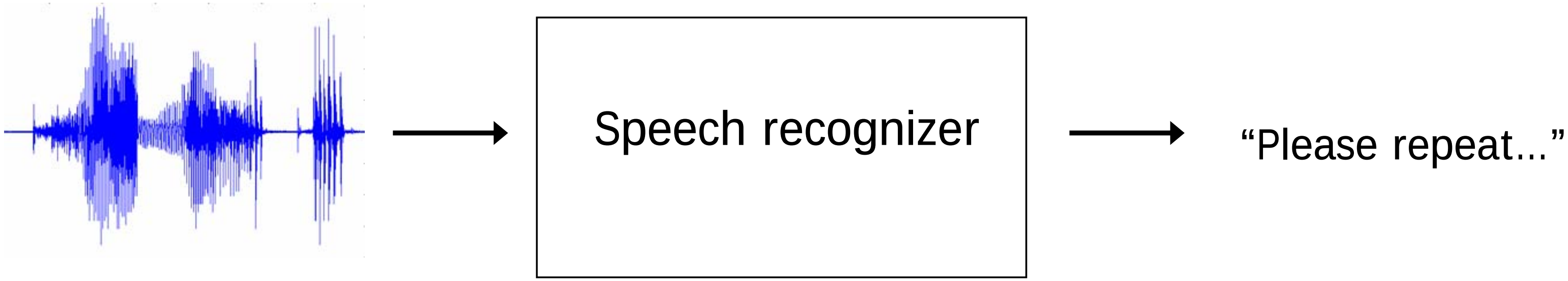
Katrin Kirchhoff

Signal, Speech and Language Interpretation Laboratory (SSLI Lab)

UW EE Department

# Speech Processing: ASR

- Automatic speech recognition (ASR): given an acoustic signal, determine the sequence of words spoken



# ASR Applications

- Dictation (e.g. IBM ViaVoice, Dragon Dictate,…)
- Voicemail transcription
- Automated Dialogue Systems (flight status information, stock quotes, home shopping channel, etc.)
- Call centers (AT&T: How may I help you?™)
- Transcription of phone conversations
- Transcription of broadcast news/TV shows (archiving)

# ASR Applications

- Command-and-control: controlling devices in hands-free scenarios, e.g. in cars
- Household appliances with speech interfaces
- Assistive devices (e.g. users with mobility impairments)
- Audio search: searching in large archives of audio documents

# Speech Processing: Speaker Recognition

- Speaker Recognition: given an acoustic signal, determine the speaker
  - Closed system/*1-of-N* identification: speaker is one of N speakers represented in the system
  - Open system/verification: given a hypothesis about true speaker, determine whether speaker in the signal is the hypothesized speaker or an impostor
- Applications: security applications, biometrics

# Speech Processing: Language Recognition

- Language Recognition/Identification: given an acoustic signal, determine the language
- Set-up similar to speaker recognition
- Applications: automatic call routing, front-end to speech recognition modules, automated dialogue systems, wiretapping

# Challenges in Speech Processing

- Acoustic variability: signal characteristics change depending on
  - Speaker: vocal tract length, mental/physical condition, age
  - Speaker group: dialects/accents
  - Channel: microphone type
  - Recording conditions: background noise, room reverberation…

# Challenges in Speech Processing

- Coarticulation: modification of sounds based on neighbouring sounds, compare:
  - /t/ in *tool* vs. *tan*
  - /n/ in *infamous* vs. *incomparable*
  - *Massive deletion/assimilation of sounds in conversational speech*
- Vocabulary size: small vocabulary (digits) vs. large or unlimited vocabulary
- Domain/topics changes: level of predictability of words

“so, I have, like, the same situation. Yeah, I just, like, listen to whatever’s on the radio. I just listen.

## Effect of accented speech on ASR in an automated dialogue system

U: I want to go to Paris

R: ++uh++ on to go to paris

S: Where are you departing from?

U: Seattle.

R: seattle

S: What day do you want to leave Seattle?

U: October seventeenth.

R: october seventeenth

S: Do you have a preferred airline?

U: United Airlines.

R: tonight to dallas

S: I have you going from Seattle to Dallas, is that correct?

U: No, I'm going to Paris.

R: no, going to paris

S: Do you have a preferred airline?

U: I'd prefer United Airlines.

R: i'd prefer tonight to dallas

S: Do you have a preferred airline?

U: start over

U: user

R: recognition output

S: system prompt

# Natural Language Processing: Applications

- Document sorting (eg Spam filtering)
- Question Answering, Information Extraction
- Machine Translation
- Document summarization
- Etc.

## Challenges: Ambiguities

- Iraqi head seeks arms

  *[word sense disambiguation]*
- Enraged cow injures farmer with axe

  *[parsing: PP attachment]*
- Eye drops off shelf

  *[parsing]*
- Include your children when baking cookies

  *[pragmatic interpretation]*

# More examples...just for fun

- Bush wins budget – more lies ahead
- Squad helps dog bite victim
- Tiger Woods is playing with own balls, Nike says
- Drunks get nine years in violin case

# Natural Language Processing

- Language modeling
- Parsing
- Tagging
- Word sense disambiguation
- Coreference resolution
- Machine translation
- Etc.

# Statistical Approach

- Earlier approaches to speech/NLP used rule-based paradigm
- Predominant paradigm today: statistical pattern recognition
- Develop probabilistic model for problem at hand
- Train model parameters from large amounts of data

# Noisy Channel Model

- Data (eg words) are generated, passed through a noisy channel, observed only at output
- To recover original word string, compute

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

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

$$\propto P(O \mid W)P(W)$$

# Noisy Channel Model

- Widely used in:
  - ASR
  - Machine translation
  - Automatic summarization
  - Spelling correction
  - Text compression…
  - As well as other non-text applications
- For each application, need to determine W, O, and how to compute P(O|W) and P(W)

# Noisy Channel Model in ASR

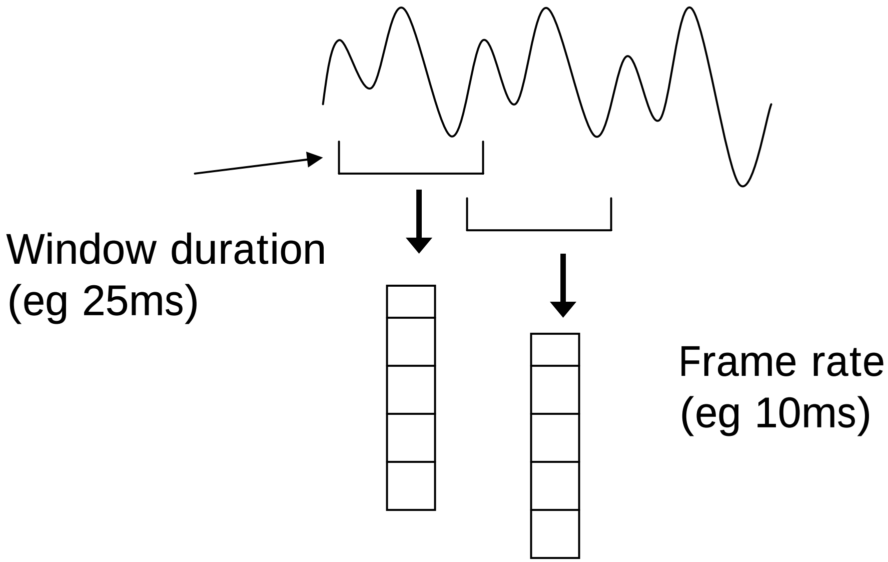
- **O** = $o_1,\ldots,o_T$ sequence of acoustic feature vectors
- **W** = $w_1,\ldots,w_N$: word sequence

$$P(W|O) \cong \underbrace{P(O|W)}_{\text{“acoustic model”}}\ \underbrace{P(W)}_{\text{“language model”}}$$

- acoustic model: determines how well acoustic signal matches hypothesized word sequence
- language model: determines prior probability of word sequence

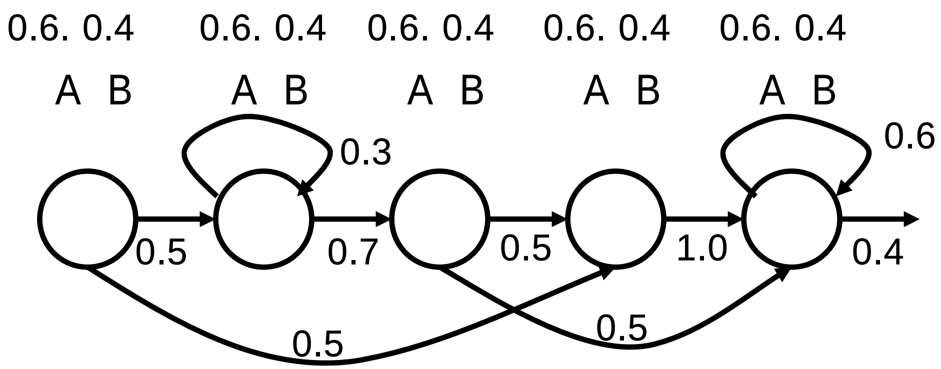
# Preprocessing

- **O**: sequence of d-dimensional feature vectors, obtained by:
  - Digitization of speech signal (sampling, quantization)
  - Windowing
  - Extraction of speech features representing time-frequency characteristics
  - Common analysis techniques
    - LPC
    - MFCC
    - PLP…



# Acoustic Model

- Cannot compute P(**O**|**W**) directly (variable number of words, variable length)
- Need flexible temporal model for each w (or subunit of w)
- Hidden Markov Model (HMM): stochastic finite-state automaton




- Consider observation sequence “AABBBBAAA”
- Which hidden state sequence generated it?

# Acoustic Model

HMM $\lambda := <\pi, Q, O, A, B>$

Q: set of states, $q_1, ..., q_N$

$O$ : set of observation symbols, $o_1, , , o_M$

A: transition probability matrix for $p(q_j \mid q_i)$

$B$ : observation probability matrix for $p(o_m \mid q_i)$

$\pi$ : vector of start probabilities

Compute likelihood of observation sequence **O** given HMM λ

$$P(O \mid \lambda) = \sum_{Q} \pi_{q(1)} \prod_{t=2}^{T} P(o(t) \mid q(t)) P(q(t) \mid q(t-1))$$

$$\text{for all } Q = q(1), q(2), ..., q(T)$$

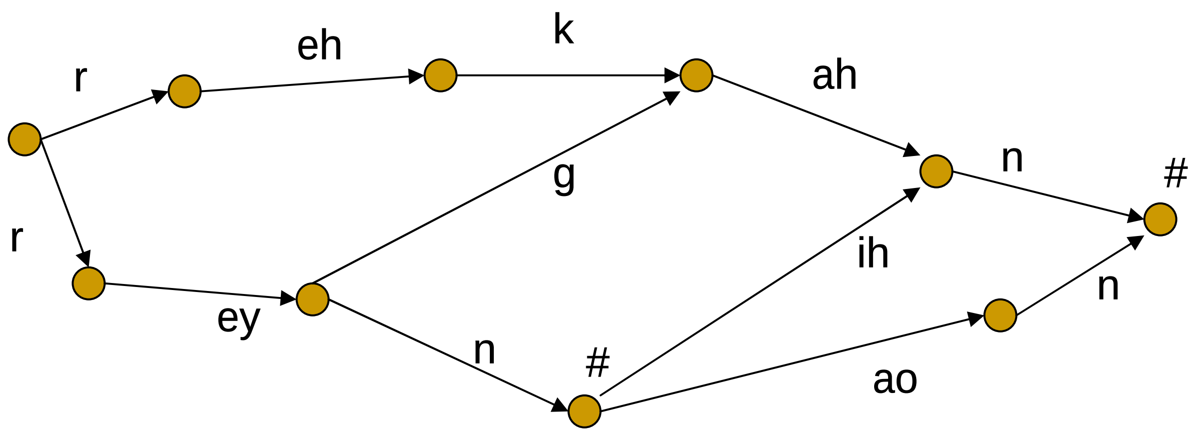
# Acoustic Model

- Training of acoustic model probabilities: Expectation-Maximization (EM)
- Requires annotated training data: acoustic signals and transcriptions (true word sequences)
- Dozens of hours or hundreds of hours
- But: word sequences need not be perfectly accurate (eg close captions can be used)

# Decoding

- Word-HMMs only used for very small vocabularies
- One HMM for each subword unit (phone)
- Requires pronunciation dictionary mapping words to phone sequences
- Recognition network consisting of possible word sequences, each word mapped to phone sequence, each phone mapped to HMM:
- Large network of HMM states; best state sequence implicitly defines best word sequence

# Decoding




- Stack decoder
- Large search space, needs to be constrained
- P(W)!

# Language Modeling

- Problem: compute probability of W = $w_1,\ldots,w_N$
- Chain rule:

$$P(w_1,\ldots,w_N) = P(w_1)\prod_{i=2}^{N} P(w_i \mid w_{i-1},\ldots,w_1)$$

- Cannot condition on full history: too many parameters, too little data!
- History is truncated to small value (2 or 3 words), defining equivalence classes of histories

$$P(w_1,\ldots,w_N) = \prod_{i=n}^{N} P(w_i \mid w_{i-1},\ldots,w_{1-n+1})$$

- "n-gram": n=2: bigram, n=3: trigram

# Language Modeling

- LM probabilities need to be estimated from data
- Large number of parameters even for small *n*: $|V|^2$ or $|V|^3$
- For $|V| = 20k$, bigram has ~400M parameters, trigram has $\sim 10^{12}$
- Many words/word combinations in test data that were not observed in training data
- Need to prevent zero probabilities!
- $\Rightarrow$ “Smoothing”

# Smoothing

- Discount relative frequency estimates and assign non-zero probabilities
- One simple way: additive smoothing: add 1 to every count, including zero counts
- Other methods:
  - Good-Turing
  - Witten-Bell
  - Kneser-Ney
- Backoff models: use higher-order n-gram probability estimates if there are enough training samples in the training data, else use lower-order n-gram probabilities

# Evaluation of ASR

- Compute string alignment between recognition output and reference:

| | | | | | |
|---|---|---|---|---|---|
| REF: | It is not easy to recognize | *** | *** | speech |
| HYP: | It is not easy to wreck | a | nice | beach |

- Compute word error rate

$$WER = 100 \frac{\text{insertions} + \text{deletions} + \text{substitutions}}{\text{total \# words in reference}}$$

# State of the Art in ASR

- Current word error rates:
  - Digit recognition < 2%
  - Isolated word recognition (600 words): 3%
  - Broadcast news recognition: 13-16%
  - Conversational telephone speech recognition: 15-20%
  - Recognition of meeting speech: 35-45%

# State of the Art in ASR

- Error examples:

REF: I really like to see other people ON HALLOWEEN
HYP: I really like to see other people IN HOLY

REF: this past year I went AS A catholic PRIEST AND my girlfriend WENT AS a nun
HYP: this past year I went TO THE catholic PRIESTS IN my girlfriend ONE IT'S a nun

REF: that was a little bit ** IRREVERENT we WERE both raised catholic
HYP: that was a little bit OF REFERENCE we ***** both raised catholic

- Confusions with acoustically similar words (affects mostly word endings in content words: *priest - priests*)
- Confusion of function words
- Unknown words in the vocabulary (*irreverent*, *Halloween*?)

# Noisy Channel Model in Machine Translation

- we are interested in English
- by some accident, data generated in English actually comes out as French

Please repeat → Noisy channel → Répétez s'il vous plaît

e f

- O=f, W=e:

$$e^* = \arg\max_e P(e \mid f)$$

$$P(e \mid f) \propto \underbrace{P(f \mid e)}_{\text{“translation model”}} \underbrace{P(e)}_{\text{“language model”}}$$

# Translation Model

- 5 classical translation models: IBM Models 1-5
- Model 1: suppose the alignment *a* of words in a bilingual corpus is known

f: Il m'a acheté un bouquet de fleurs rouges.

e: He bought me a bunch of red flowers.

$$P(f \mid e) = \sum_{a} P(f, a \mid e)$$

$$P(f, a \mid e) = P(m \mid l) \prod_{j=1}^{m} P(f_j \mid e_{a_j})$$

m: length of f

l: length of e

# Translation Model

- Model 2: alignment is a hidden variable, dependent on alignment of previous word and lengths of *e* and *f*

$$P(f,a|e) = P(m|l)\prod_{j=1}^{m} P(f_j|e_{a_j})P(a_j|a_{j-1},m,l)$$

- Model 3*:* introduces fertility (number of words a source language word can generate) and distortion model

$$P(f,a|e) = P(m)\prod_{j=1}^{m} P(f_j|e_{a_j})P(j|a_j,l,m)\prod_{i=1}^{l} p(\phi_i|e_i)$$

(somewhat simplified)

# Machine translation

- Training of translation models from aligned (sentence-aligned, paragraph-aligned) bilingual data using Expectation-Maximization
- Current systems use phrase-based models: mappings between chunks of words in both languages
- Language model: same as in ASR
- Decoding: stack decoder

# State of the Art in MT

- Evaluation metrics: see slides from previous lecture
- Good example:

INPUT: deseo felicitarme ante todo del trabajo realizado por el parlamento europeo en forma de informe por separado , dedicado a cada uno de los doce países candidatos que han iniciado negociaciones

TRANS: i wish to congratulate above all of the work done by the european parliament in the form of a separate report on each of the twelve candidate countries that have begun negotiations

REF: i wish to congratulate the european parliament on the work accomplished in the reform of a report devoted to each of the twelve candidate countries which have entered into negotiations

# State of the Art in MT

- Bad example:

INPUT: durante nueve noches los fieles acuden al templo y a las 7 de la noche se lleva a cabo una misa y una reflexión inspirada en el ave maría que los líderes de la sociedad guadalupana han preparado previamente

TRANS: for nine nights the faithful come to the temple and at 7 a.m. this evening is carried out a hounds and a reflection inspired by the poultry maría that the leaders of society guadalupana have prepared in advance .

REF: for nine nights the faithful have been going to the temple and at seven in the evening a mass is held with a reflection inspired by the ave maria that the leaders of the guadalupan society had prepared beforehand .

# Noisy Channel Model in Summarization

- Summarization: find a shorter, compressed form of a document that contains its most relevant information


- W = S, O = D
- P(S): language model or probabilistic grammar
- P(D|S): product of probabilities of operations that expand S (e.g. insertion of syntactic constituents)

# Further Reading

- D. Jurafsky and J.H. Martin: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition,* Prentice Hall, 2000

katrin@ee.washington.edu