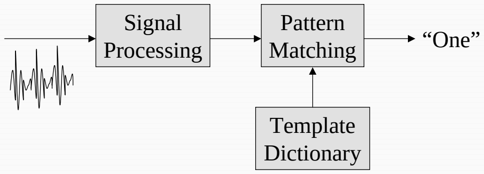
Overview of Speech Recognition

SAIL Speech Analysis and Interpretation Lab

- ASR:
  - Front End
    - Signal Processing
    - Features & motivation from biology
  - Pattern Matching
    - Basics of Probability Theory
    - Markov Models
    - Hidden Markov Models in speech
  - Creating the database of patterns
  - More probabilities & more Bayes:
    - Language Modeling (the holly grail of DM?)
- The SAIL Lab
  - Examples of where DM is needed!

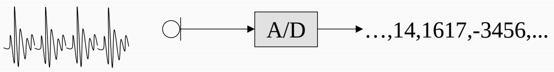
Speech Recognition Basics



Signal Processing

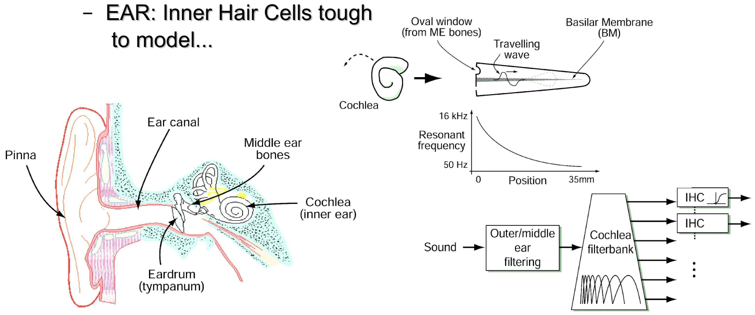
Signal Acquisition

- Speech is captured by a microphone.
- The analog signal is converted to a digital signal by sampling it 16000 times a second.
- Each value is quantized to 16 bits, a number between -32768 and 32767.



Front End

- Need to find the **features** of the speech sound
- What are the features:
  - derived from knowledge of the workings of the ear=perception or the production or modeling of resonances etc
  - EAR: Inner Hair Cells tough to model...



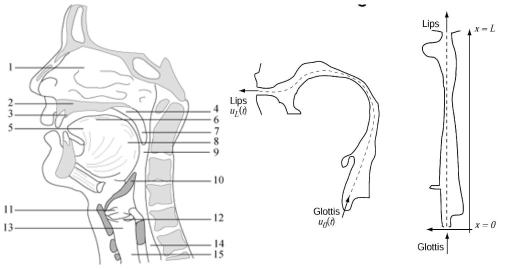
Front End

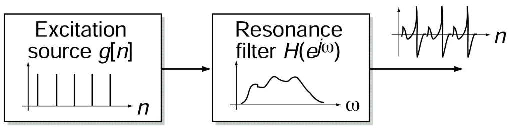
- The ear is divided into three parts: the external ear, the middle ear and the inner ear. The external ear collects sound, the middle ear mechanism transforms the sound and the inner ear receives and transmits the sound.
- Sound vibrations enter the ear canal and cause the eardrum to vibrate. Movements of the eardrum are transmitted across the middle ear to the inner ear fluids by three small ear bones. These middle ear bones (hammer or malleus, anvil or incus and stirrup or stapes) act as a transformer changing sound vibrations in air into fluid waves in the inner ear. The fluid waves stimulate delicate nerve endings in the hearing canals. Electrical impulses are transmitted on the nerve to the brain where they are interpreted as understandable sound.

Production

- Tube model:

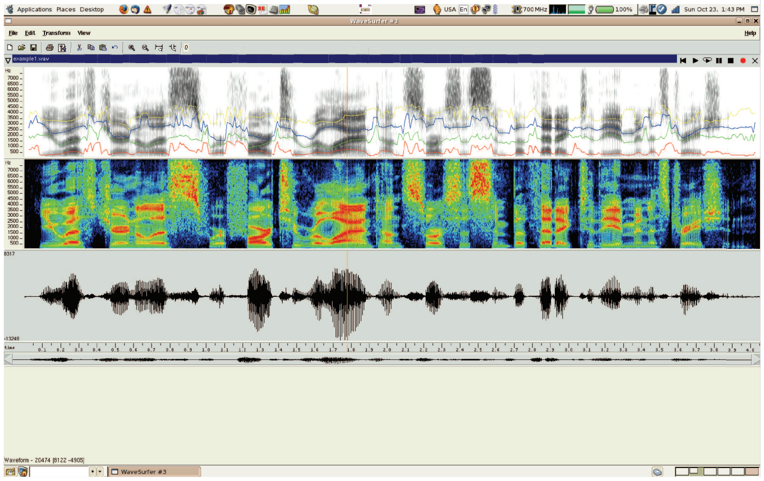
(1) Nasal cavity,
(2) Hard palate,
(3) Alveoral ridge,
(4) Soft palate (Velum),
(5) Tip of the tongue (Apex),
(6) Dorsum,
(7) Uvula,
(8) Radix,
(9) Pharynx,
(10) Epiglottis,
(11) False vocal cords,
(12) Vocal cords,
(13) Larynx,
(14) Esophagus, and
(15) Trachea.




Excitation source $g[n]$

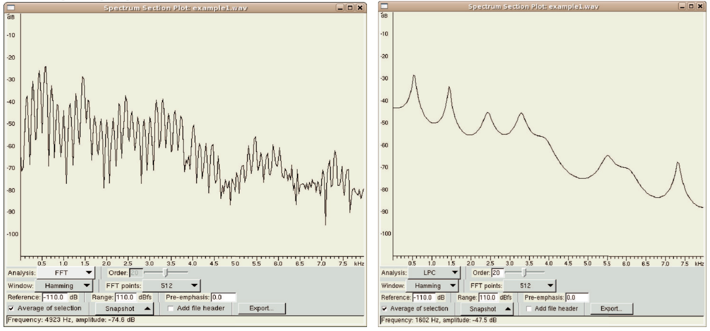
Resonance filter $H(e^{j\omega})$

Front End



Front End

- Remove data keeping information intact
- Periodic excitation (air through the lungs) contains little lexical information


Spectrum Section Plot: example1.wav

Analysis: FFT
Order: 20
Window: Hamming
FFT points: 512
Reference: -110.0 dB
Range: 110.0 dBfs
Pre-emphasis: 0.0
☑ Average of selection
Snapshot
□ Add file header
Export...
Frequency: 4923 Hz, amplitude: -74.6 dB

Spectrum Section Plot: example1.wav

Analysis: LPC
Order: 20
Window: Hamming
FFT points: 512
Reference: -110.0 dB
Range: 110.0 dBfs
Pre-emphasis: 0.0
☑ Average of selection
Snapshot
□ Add file header
Export...
Frequency: 1602 Hz, amplitude: -47.5 dB

Spectral Analysis

- Take a **window** of 20-30 ms (320 to 480 samples).
- **FFT** (Fast Fourier Transform) used to compute energy in several frequency bands, typically 20 to 40.
- **Cepstrum** computed, 12 or so coefficients.

Spectral Analysis

- A **frame**, a snapshot of the input's spectrum, is represented by a vector of 12 cepstrum coefficients.
- To describe the spectral change with time, one frame is computed every 10 ms or so, *i.e.* 100 frames per second.
- A word that lasts 1 second is described by 1200 numbers.

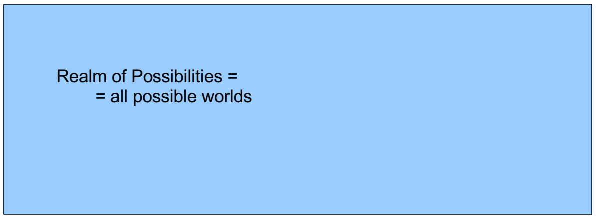
Probability

Discrete Random Variables

- A is a Boolean-valued random variable if A denotes an event, and there is some degree of uncertainty as to whether A occurs.
- Examples
  - A = The US president in 2023 will be male
  - A = You wake up tomorrow with a headache
  - A = You are sick

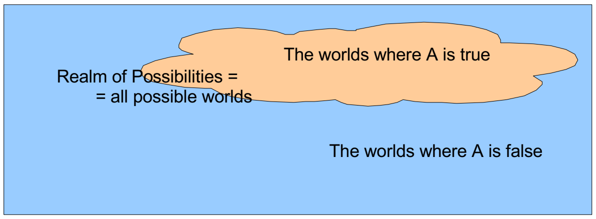
Probabilities

- Probability of A = P(A) is “the fraction of possible worlds in which A is true”

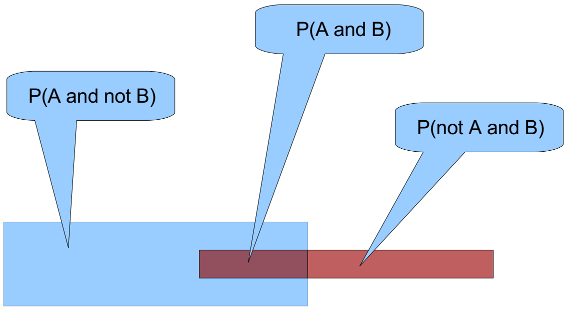


Probabilities

- Probability of A = P(A) is “the fraction of possible worlds in which A is true”



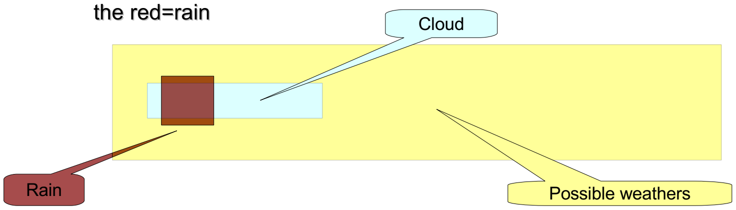
Axioms visually

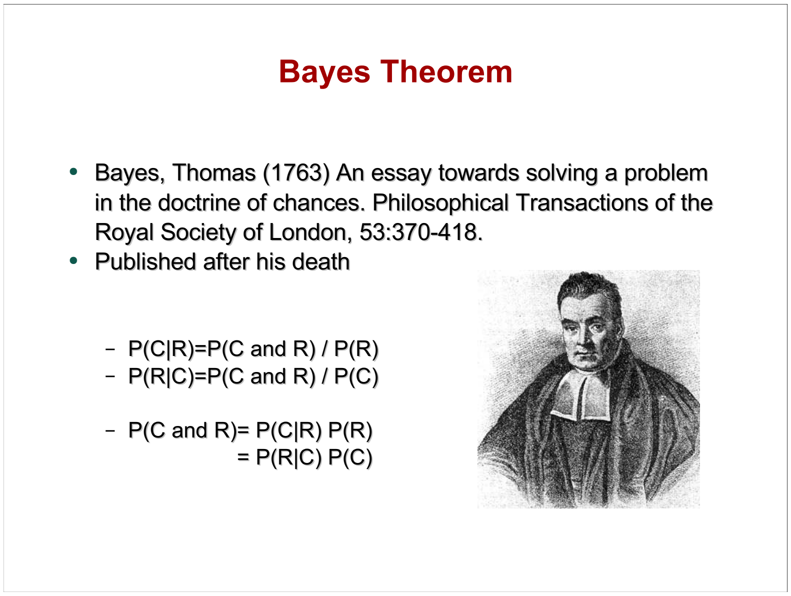


- P(A or B) = P(A) + P(B) – P(A and B)

Conditionals

- Example: Rain & Cloudy in Los Angeles
  - There is a small chance of Rain and a slightly larger chance of Cloudy
  - The chances of Cloudy when it is Raining are very large
  - P(C|R)=P(C and R) / P(R)
  - Probability we are in the cyan=cloud given we are in the red=rain is the area of the overlap normalized by the area of the red=rain





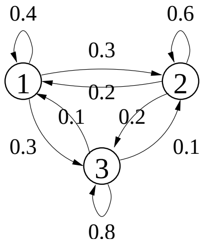
Bayes Theorem

- Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, 53:370-418.
- Published after his death
  - P(C|R)=P(C and R) / P(R)
  - P(R|C)=P(C and R) / P(C)
  - P(C and R)= P(C|R) P(R)
    = P(R|C) P(C)

Markov Models

Weather predictor example of a Markov model

State 1: rain
State 2: cloud
State 3: sun



State-transition probabilities,

$$A = \left\{a_{ij}\right\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \qquad (12)$$

Weather predictor calculation

Given today is sunny (i.e., $x_1 = 3$), what is the probability of "sun-sun-rain-cloud-cloud-sun" with model $\mathcal{M}$?

$$\begin{aligned} P(X|\mathcal{M}) &= P(X = \{3, 3, 1, 2, 2, 3\}|\mathcal{M}) \\ &= P(x_1 = 3)\, P(x_2 = 3|x_1 = 3) \\ &\quad P(x_3 = 1|x_2 = 3)\, P(x_4 = 2|x_3 = 1) \\ &\quad P(x_5 = 2|x_4 = 2)\, P(x_6 = 3|x_5 = 2) \\ &= \pi_3\, a_{33}\, a_{31}\, a_{12}\, a_{22}\, a_{23} \\ &= 1 \cdot (0.8)(0.1)(0.3)(0.6)(0.2) \\ &= 0.00288 \end{aligned}$$

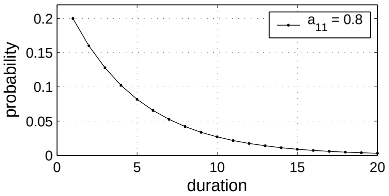
where the initial state probability for state i is

$$\pi_i = P(x_1 = i). \qquad (13)$$

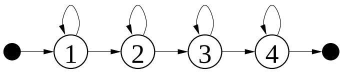
State duration probability

As a consequence of the first-order Markov model, the probability of occupying a state for a given duration, τ, is exponential:

$$p(X|\mathcal{M}, x_1 = i) = (a_{ii})^{\tau-1}\,(1 - a_{ii})\,. \qquad (14)$$



Summary of Markov models



Transition probabilities:

$$A = \left\{a_{ij}\right\} = \begin{bmatrix} 0.6 & 0.4 & 0 & 0 \\ 0 & 0.9 & 0.1 & 0 \\ 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}$$

and $\qquad \pi = \{\pi_i\} = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}.$

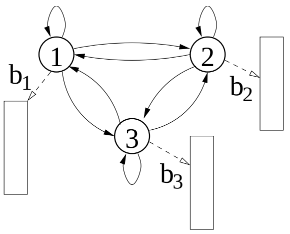
Probability of a given state sequence X:

$$\begin{aligned} P(X|\mathcal{M}) &= \pi_{x_1}\, a_{x_1x_2}\, a_{x_2x_3}\, a_{x_3x_4} \ldots \\ &= \pi_{x_1} \prod_{t=2}^{T} a_{x_{t-1}x_t}. \end{aligned} \qquad (15)$$

Hidden Markov Models

Hidden Markov Models

Urns and balls example (Ferguson)



Probability of state i producing an observation o_t is:

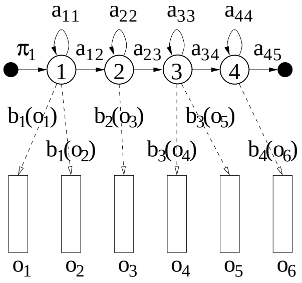
$$b_i(o_t) = P(o_t | x_t = i), \tag{16}$$

which can be *discrete* or *continuous* in o.

Elements of a discrete HMM, λ

1. Number of states N, $x \in \{1, \ldots, N\}$;
2. Number of events K, $k \in \{1, \ldots, K\}$;
3. Initial-state probabilities,
   $\pi = \{\pi_i\} = \{P(x_1 = i)\}$ for $1 \le i \le N$;
4. State-transition probabilities,
   $A = \{a_{ij}\} = \{P(x_t = j | x_{t-1} = i)\}$ for $1 \le i, j \le N$;
5. Discrete output probabilities,
   $B = \{b_i(k)\} = \{P(o_t = k | x_t = i)\}$ for $1 \le i \le N$ and $1 \le k \le K$.

Hidden Markov model example



with state sequence $X = \{1, 1, 2, 3, 3, 4\}$,

$$\begin{aligned} P(\mathcal{O}|X,\lambda) &= b_1(o_1)\, b_1(o_2)\, b_2(o_3)\, b_3(o_4)\, b_3(o_5)\, b_4(o_6) \\ P(X|\lambda) &= \pi_1\, a_{11}\, a_{12}\, a_{23}\, a_{33}\, a_{34} \end{aligned} \tag{17}$$

$$P(\mathcal{O}, X|\lambda) = \pi_1 b_1(o_1)\, a_{11} b_1(o_2)\, a_{12} b_2(o_3) \ldots \tag{18}$$

Isolated digit recognition

- 10 templates: one template M$_i$ per digit.
- Compare input **x** with all templates.
- Select the most similar template *j:*

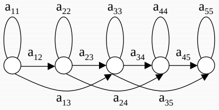
  $$j = \min\{f(\mathbf{x}, M_i)\}$$

where *f* is the comparison function.

Hidden Markov Models (HMM)

HMM defined by:

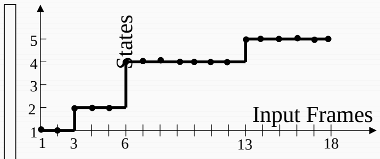
- Number of states
- Transition probabilities a$_{ij}$
- Output probabilities $b_i(x)$:

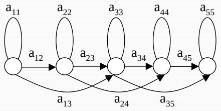


$$b_i(x) = \sum_{k=0}^{K-1} c_{ik} N(x, \mu_{ik}, \sigma_{ik})$$

$$N(x, \mu, \sigma) = \frac{1}{(2\pi)^{p/2}} \exp\{-\sum_{m=0}^{p-1} \frac{(x_m - \mu_m)^2}{2\sigma_m^2}\}$$

Probability of a path





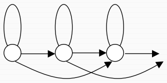
$$P = b_1(x_1)a_{11}b_1(x_2)a_{12}b_2(x_3)a_{22}b_2(x_4)a_{22}b_2(x_5)a_{24}b_4(x_6)$$
$$a_{44}b_4(x_7)a_{44}b_4(x_8)a_{44}b_4(x_9)a_{44}b_4(x_{10})a_{44}b_4(x_{11})a_{44}b_4(x_{12})$$
$$a_{45}b_5(x_{13})a_{55}b_5(x_{14})a_{55}b_5(x_{15})a_{55}b_5(x_{16})a_{55}b_5(x_{17})a_{55}b_5(x_{18})$$

Probability of a model

- Is the probability of the best path.
- Problem: We need to evaluate all possible paths, and there grow exponential with number of states and input frames.
- Solution: The Viterbi algorithm has complexity that grows linearly with number of states and input frames.

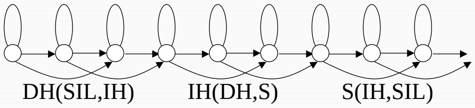
Large vocabulary

- Problem:
  - Huge number of models: 60,000 models
  - Models are hard to “train”
  - Cannot easily add new words.
- Solution: Use phoneme models



Context Dependent Models

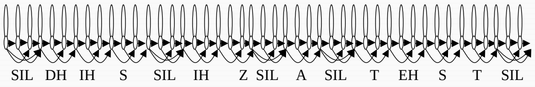
- To improve accuracy, phone models are context-dependent.
- Example: THIS = DH IH S



- 50^3=125,000 possible models are clustered to about 8,000 *generalized triphones*

Continuous speech

- Example: THIS IS A TEST
- SIL DH(SIL,IH) IH(DH,S) S(IH,SIL) SIL IH(SIL,Z) Z(IH,SIL) SIL A(SIL,SIL) SIL T(SIL,EH) EH(T,S) S(EH,T) T(S,SIL) SIL
- SIL DH(SIL,IH) IH(DH,S) S(IH,IH) IH(S,Z) Z(IH,A) A(Z,T) T(A,EH) EH(T,S) S(EH,T) T(S,SIL) SIL

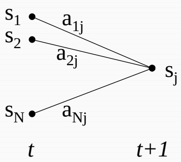


Baum-Welch Algorithm

$$\alpha_t(i) = P(x_1, x_2 \cdots x_t, s_t = i \mid \lambda)$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(x_{t+1})$$

$$P(X \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$



Exact solution.

Speeding Search: Pruning

- Viterbi algorithm: replace sum with max.
- Speed-Accuracy Tradeoff: Computations can be reduced by eliminating paths that are “not promising”, at the expense of having a chance of eliminating the “best” path.
- Typically, computations can be reduced by more than a factor of 10, without affecting the error rate significantly.

Training

HMM Training

- A **lot** of speech is needed to *train* the models. Done with an iterative algorithm:
  1. Take initial model (*i.e.* uniform probabilities).
  2. Segment Database (Run Viterbi algorithm).
  3. Update models.
  $$\hat{\mu}_j = \frac{\sum_{i=0}^{N-1} x_{ij}}{N} \qquad \hat{\sigma}_j^2 = \frac{\sum_{i=0}^{N-1} (x_{ij} - \hat{\mu}_j)^2}{N}$$
  4. If Converged stop, otherwise go to 2.

HMM Training

- Rule of thumb: Each state has to appear in the training database at least 10 times.
- *Speaker-Independent* systems have lots of data (models well trained), but contain high variability.
- *Speaker-Dependent* systems do not have as much data (models not so well trained) but they are more consistent.

Language Modeling

Bayes Rule in ASR

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

$$\hat{W} = \max P(W \mid A) = \max P(A|W)P(W)$$

where A is the Acoustics and W the sequence of words. *P(W)* is the language model.

Statistical Language Model

$$P(W) = P(W_1 W_2 W_3 \cdots W_N) = P(W_1) P(W_2 \mid W_1)$$
$$P(W_3 \mid W_2 W_1) \cdots P(W_N \mid W_{N-1} \cdots W_2 W_1)$$

$$\approx P(W_1) P(W_2) P(W_3) \cdots P(W_N) \text{ Unigram}$$

$$\approx P(W_1) P(W_2 \mid W_1) P(W_3 \mid W_2) \cdots P(W_N \mid W_{N-1})$$
Bigram

$$\approx P(W_1) P(W_2 \mid W_1) P(W_3 \mid W_2 W_1) \cdots P(W_N \mid W_{N-1} W_{N-2})$$
Trigram

Trigrams

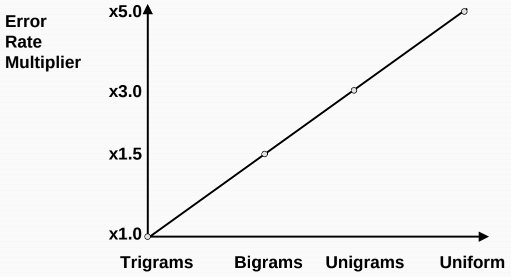
P(THIS IS A TEST) =

P(THIS) *P*(IS|THIS) *P*(A|THIS IS) *P*(TEST|IS A)

- These statistical language models **predict** the probability of the current word given the past history.
- While simplistic, they contain a lot of information

P(IS|THIS) >> *P*(IS|HAVE)

N-Gram Performance



Perplexity

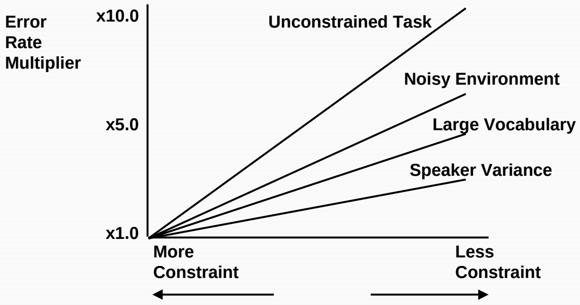
- Perplexity measures the average branching of a text when presented to a language model. $PP = 2^{LP} = \hat{P}(W_1, W_2, \cdots W_n)^{-1/n}$
- An empirical observation $E \approx k\sqrt{P}$ where *E* is the error rate and *P* the perplexity.
- A language model with low perplexity is good for ASR.
- Tradeoff between perplexity and coverage.

Context Free Grammar

- Advantages:
  - Low error rate (because of low perplexity).
  - Compact.
  - Easy for application developers.
- Disadvantages:
  - Poor coverage (out of language sentences).

Applications

ASR problem Space



ASR Accuracy in the Lab

- Command &Control (300 words + CFG)
  - 1% word error rate (speaker-independent)
- Discrete Dictation (60K words)
  - 3% word error rate (speaker-dependent)
- Continuous Dictation (60K words)
  - 7% word error rate (speaker-dependent)
- Telephone Digits
  - 0.3% word error rate (speaker-independent)

C & C Types of Errors

- Pronunciation errors make some words *unrecognizable*.
- User says something outside the vocabulary/grammar.
- User speaks while machine is talking.
- Background noise fires recognizer (false alarm). Rejection needs improvement.
- Spontaneous speech (uhms, ahms)

ASR in Real Applications

- New word addition
- Rejection
- Barge-in.
- Robustness to noise.
- ASR is just a part of an application, User Interface is critical.

Interface NLP/ASR

- Loose coupling: NLP selects correct input from a list of top N candidate sentences:
  - Easy to implement, not optimal
- Tight coupling. NLP provides the language probabilities for the search:
  - Optimal but hard to implement.
  - Can a NLP system reduce perplexity?

SAIL Lab:

Labs Mission

- Human communication involves a complex orchestration of cognitive, physiological, physical, social processes
  - Sophisticated human production-perception chain: can learn and adapt
  - Inherently multimodal: natural sensory communication involves speech, gestures, touch: Spoken language plays a key role
- Speech signal carries crucial information: intent, desires, emotions
  - Nonlinear progression and unfolding of events that can be marked with linguistic and extra-linguistic tags

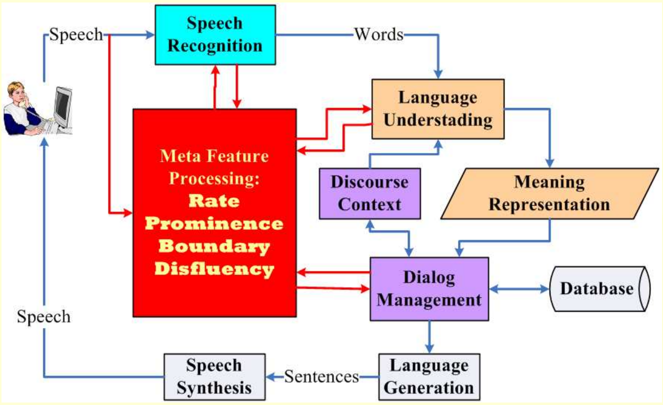
Labs Mission

- Information resides at multiple time scales, through multiple cues
  - Time-frequency modulation of speech signal not completely localized: interactions across multiple events, synchronous and asynchronous
  - Information contained at various levels of linguistic abstraction: neuro cognitive, articulatory, acoustic lexical, syntactic, semantic, discourse

Labs Mission

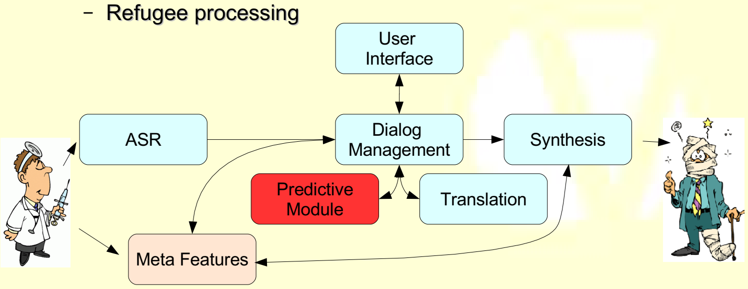
- Implementing the multi-level mapping: deciphering what, who and how
  - "what": speech content, the primary object of **speech** recognition
    - Robustly map speech signal to words
  - "who": speech source, the primary object of **speaker** recognition biometrics
    - Map speech and other cues to verify or determine speaker identity
  - "how": speech style and emotions, object of expressive speech processing (analysis, synthesis and recognition)
    - How – literary assessment
    - Emphasis, contrast of new, important information (linguistic)
      - Boundary information: utterance, phrase, word,
      - Speech rate information
      - Content word-function word distinction
      - Prominence detection
    - User state such as frustration, attention and emotions such as anger and sadness

Conversational Systems

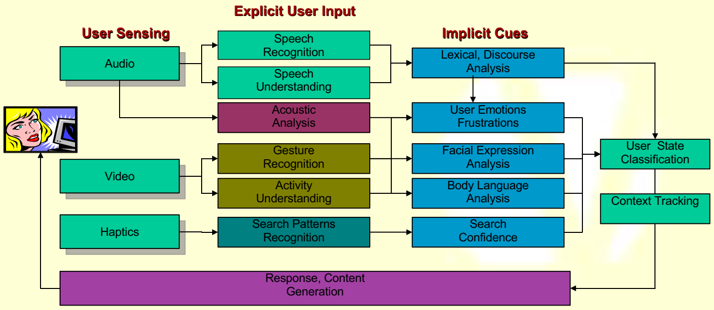


Conversational Systems – Translation

- Potential broad impact in
  - Hospitals: Especially Spanish for the Latino population in US
  - Refugee processing

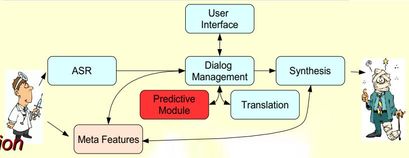


Multimodal Sensing



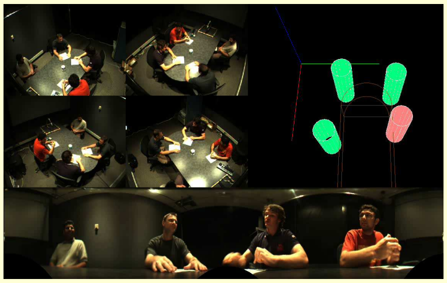
Transonics Program
Overall Focus

- ***Total Communication***
  - *Domain specific Spoken* language translation
  - *Cross-cultural communication*
  - *Rapid deployment* for:
    - *new language* pairs
    - *new domains*
  - *Co-evolution* of technology after deployment (inc. field adaptation & extension of functionality, and flow-back of data from field to development centers)
  - Exploit Domain Knowledge – Agent – for improved performance and aiding user
- *Employ Translation for Cultural & language training*

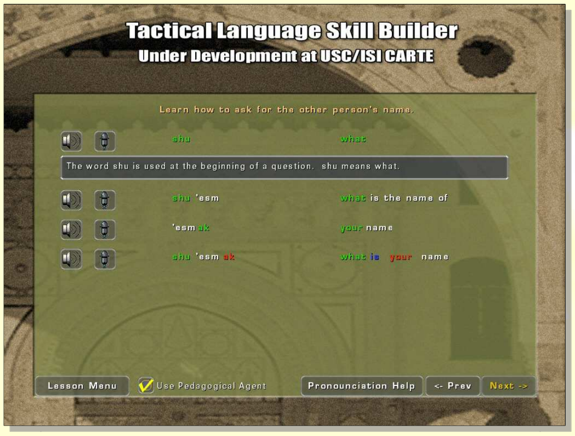


Smart-Room/Multimodal

- Scalable Immersive Environments (IMSC & SAIL)
  - Sense the user:
    - Identity of speaker
    - Location of speaker
    - Recognition of speech
    - Emotion recognition
    - Participant locations
  - and working on:
    - Gaze recognition
    - Engagement ....
  - Low latency, HD streaming
  - Summarization
  - ....



Language Training



Tactical Language Skill Builder

Under Development at USC/ISI CARTE

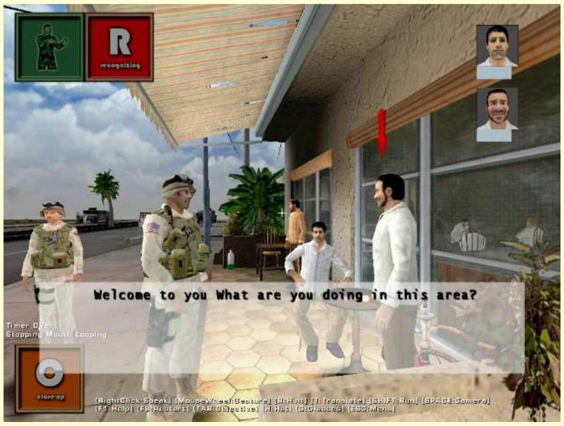
Learn how to ask for the other person's name.

shu	what

The word shu is used at the beginning of a question. shu means what.

shu 'esm	what is the name of
'esm ak	your name
shu 'esm ak	what is your name

Language Training



Welcome to you What are you doing in this area?

Timer Over
Stopping Mouth Looping

[RightClick:Speak] [MouseWheel:Gesture] [R:Hint] [T:Translate] [SHIFT:Run] [SPACE:Camera]
[F1:Help] [F5:Restart] [TAB:Objective] [H:Hat] [G:Glasses] [ESC:Menu]

Language Training

[RightClick:Speak] [MouseWheel:Gesture] [V:Aide] [SHIFT:Walk] [SPACE:Camera]
[F1:Help] [F8:Restart] [TAB:Objective] [H:Hat] [ESC:Menu]

Note: Lectures put together from own material as well as material from tutorials of Alex Acero & Philips Jackson