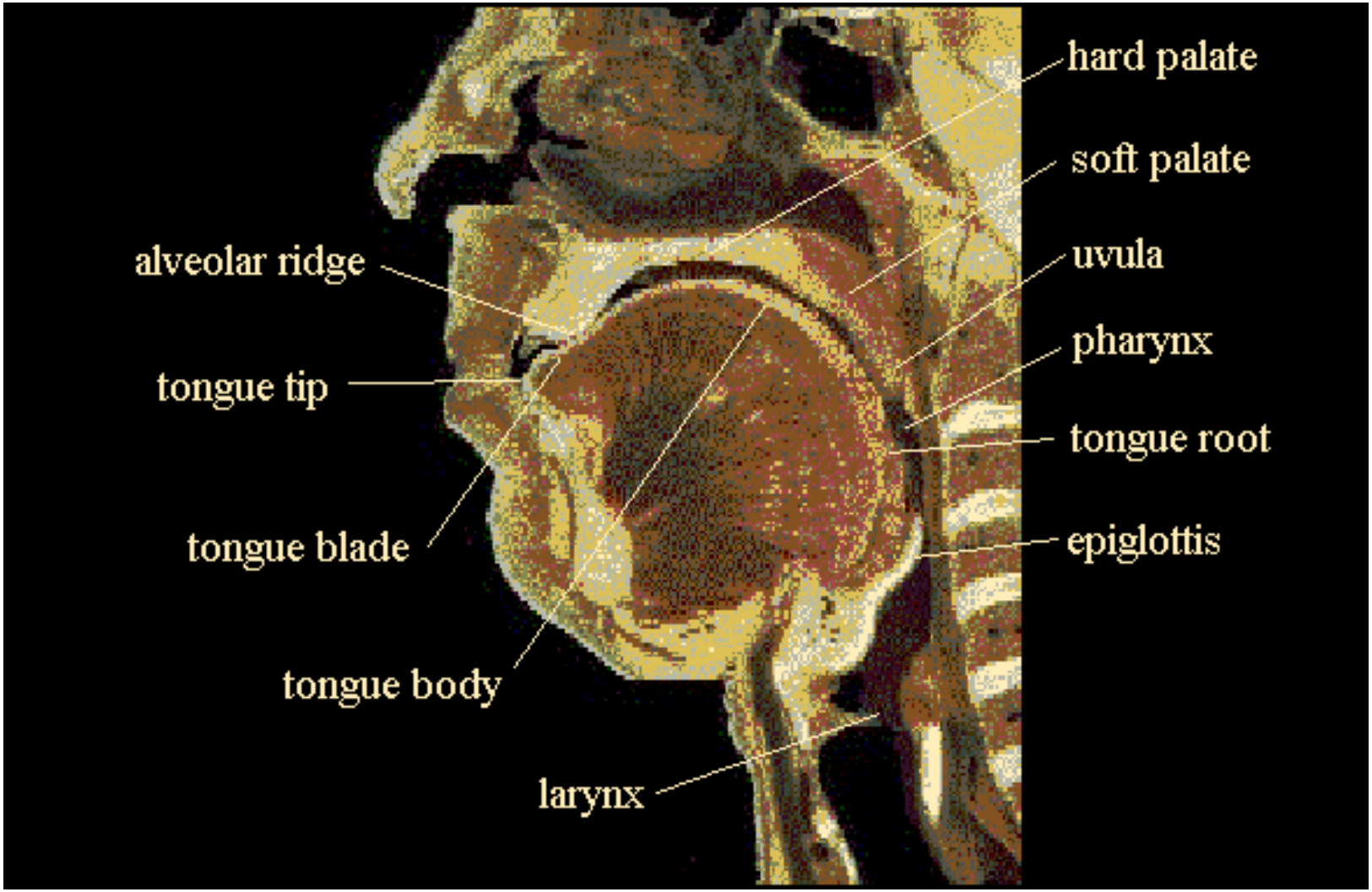
Speech Processing

Using Speech with Computers

Overview

- Speech vs Text
  - Same but different
- Core Speech Technologies
  - Speech Recognition
  - Speech Synthesis
  - Dialog Systems
  - Other Speech Processing

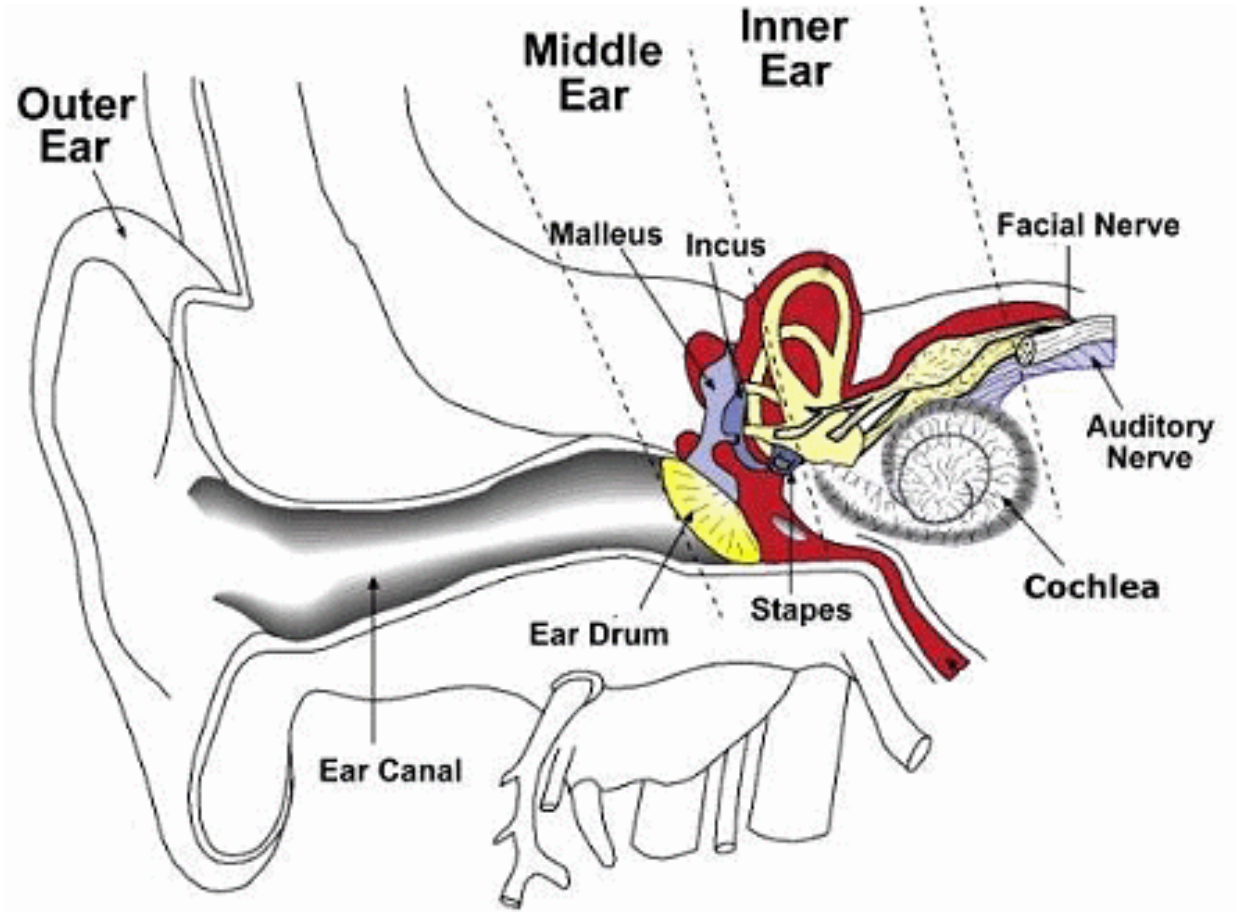
The vocal tract



From meat to voice

- *Blow air through lungs*
  - *Vibrate larynx*
  - *Vocal tract shape defines resonance*
  - *Obstructions modify sound*
    - → *Tongue, teeth, lips, velum (nasal passage)*

The ear


Outer Ear
Middle Ear
Inner Ear
Malleus
Incus
Facial Nerve
Auditory Nerve
Cochlea
Stapes
Ear Drum
Ear Canal

From sound to brain waves

- *Sound waves*
  - *Vibrate ear drum*
  - *Cause fluid in cochlear to vibrate*
  - *Spiral cochlear*
    - → *Vibrate hairs inside cochlear*
    - → *Different frequencies vibrate different hairs*
    - → *Converts time domain to frequency domainS*

Phonemes

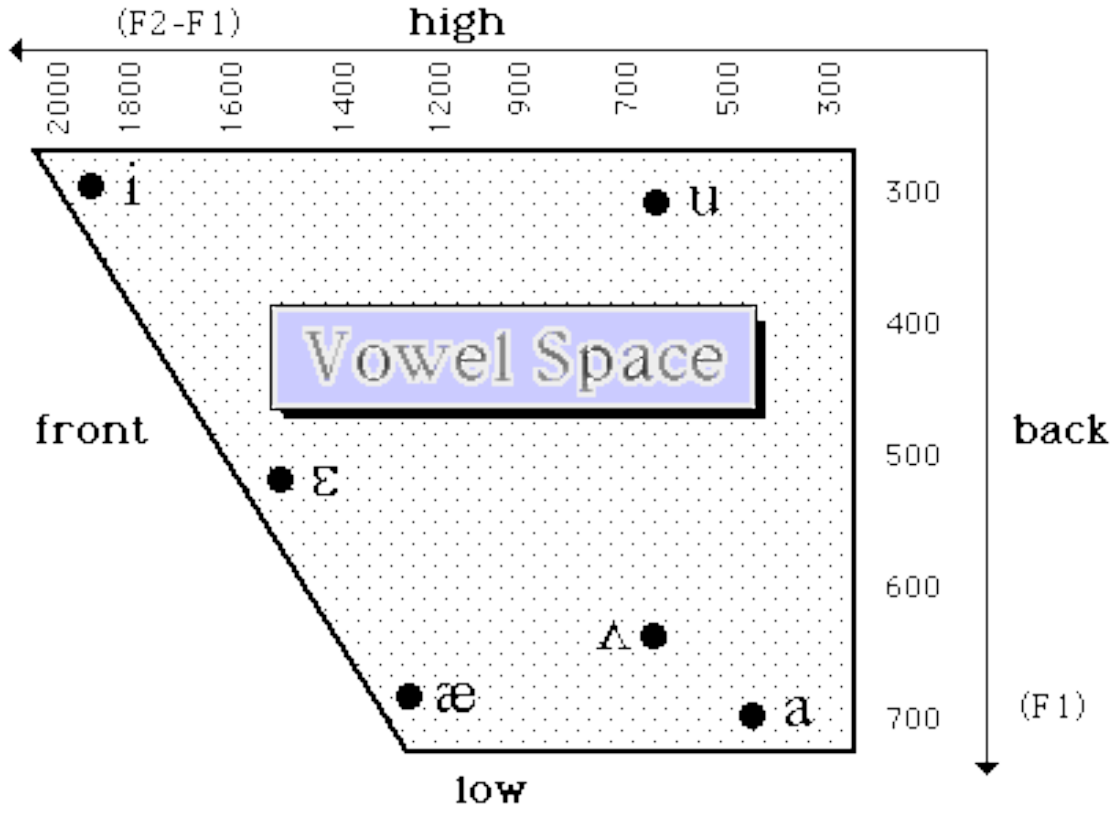
- *Defined as fundamental units of speech*
  - *If you change it, it (can) change the meaning*

“pat” to “bat”

“pat” to “pam”

Vowel Space

- One or two banded frequencies (formants)



English (US) Vowels

AA	wAshington	AE	fAt, bAd
AH	bUt, hUsh	AO	lAWn, mAll
AW	hOW, sOUth	AX	About, cAnoe
AY	hIde, bUY	EH	gEt, fEAther
ER	makER, sEARch	EY	gAte, EIght
IH	bIt, shIp	IY	bEAt, shEEp
OW	lOne, nOse	OY	tOY, OYster
UH	fUll	UW	fOOl

English Consonants

- Stops: P, B, T, D, K, G
- Fricatives: F, V, HH, S, Z, SH, ZH
- Affricatives: CH, JH
- Nasals: N, M, NG
- Glides: L, R, Y, W

- Note: voiced vs unvoiced:
  - P vs B, F vs V

Not all variation is Phonetic

- *Phonology: linguistically discrete units*
  - *May be a number of different ways to say them*
  - */r/ trill (Scottish or Spanish) vs US way*
- *Phonetics vs Phonemics*
  - *Phonetics: discrete units*
  - *Phonemics: all sounds*
- */t/ in US English: becomes "flap"*
  - *"water" / w ao t er /*
  - *"water" / w ao dx er /*

Dialect and Idiolect

- *Variation within language (and speakers)*
- *Phonetic*
  - *“Don” vs “Dawn”, “Cot” vs “Caught”*
  - *R deletion (Haavaad vs Harvard)*
- *Word choice:*
  - *Y’all, Yins*
  - *Politeness levels*

Not all languages are the same

- *Asperated stops (Korean, Hindi)*
  - *P vs PH*
  - *English uses both, but doesn't care*
  - *Pot vs sPot (place hand over mouth)*
- *L-R in Japanese not phonological*
- *US English dialects:*
  - *Mary, Merry, Marry*
- *Scottish English vs US English*
  - *No distinction between "pull" and "pool"*
  - *Distinction between: "for" and "four"*

Different language dimensions

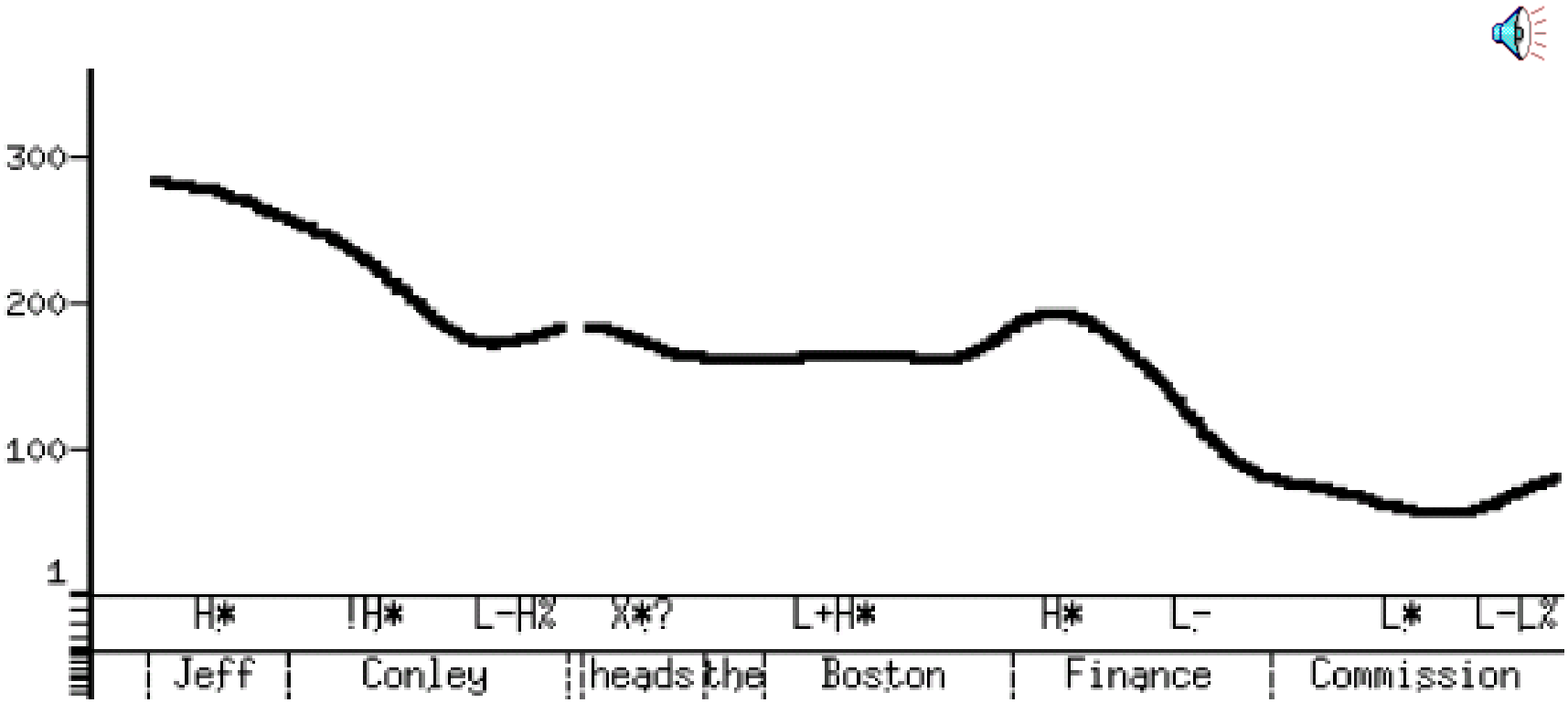
- *Vowel length*
  - *Bit vs beat*
  - *Japanese: shujin (husband) vs shuujin (prisoner)*
- *Tones*
  - *F0 (tune) used phonetically*
  - *Chinese, Thai, Burmese*
- *Clicks*
  - *Xhosa*

Prosody

- *Intonation*
  - *Tune*
- *Duration*
  - *How long/short of each phoneme*
- *Phrasing*
  - *Where the breaks are*

- *Used for:*
  - *Style, emphasis, confidence etc*

Intonation Contour



Intonation Information

- *Large pitch range (female)*
- *Authoritive since goes down at the end*
  - *News reader*
- *Emphasis for Finance H**
- *Final has a raise – more information to come*

- *Female American newsreader from WBUR*
- *(Boston University Radio)*

Words and Above

- *Words*
  - *The things with space around them (sort of)*
  - *Chinese, Thai, Japanese doesn't use spaces*
- *Words aren't always what they seem*
  - *Can you pass the salt?*
  - *Boston. Boston! Boston?*
  - *Yeah, right*
- *Multiple ways to say the same thing:*
  - *I want to go to Boston.*
  - *Yes*

Speech Recognition

- *Two major components*
  - *Acoustic Models*
  - *Language Models*
- *Accuracy various with*
  - *Speaker, language, dialect*
  - *Microphone type, environment*
  - *Speaking style:*
  - *Good Recognition:*
    - → *Head mounted mike, controlled language, careful speaker*
  - *Not so good recognition:*
    - → *Remote mike, chatting between friends, in open cafe*

But not just acoustics

- But not all phones are equi-probable
- Find word sequences that maximizes

$$P(W \mid O)$$

- Using Bayes' Law

$$\frac{P(W)P(O|W)}{P(O)}$$

- Combine models
  - Us HMMs to provide $P(O \mid W)$
  - Use language model to provide $P(W)$

Speech Synthesis

- Three Levels
  - Text analysis
    - → From characters to words
  - Prosody and Pronunciation
    - → From words to phonemes and intonation
  - Waveform generation
    - → From phonemes to waveforms

Text Analysis

- *This is a pen.*
- *My cat who lives dangerously has nine lives.*
- *He stole $100 from the bank.*
- *He stole 1996 cattle on 25 Nov 1996.*
- *He stole $100 million from the bank.*
- *It's 13 St. Andrew St. near the bank.*
- *Its a PIII 1.5Ghz, 512MB RAM, 160Gb SATA, (no IDE) 24x cdrom and 19" LCD.*
- *My home pgae is http://www.geocities.com/awb/.*

Waveform Generation

- *Formant synthesis* 
- *Random word/phrase concatenation* 
- *Phone concatenation* 
- *Diphone concatenation* 
- *Sub-word unit selection* 
- *Cluster based unit selection* 
- *Statistical Parametric Synthesis* 
- *Wavenet Neural Synthesis* 

Pronunciation Lexicon

- List of words and their pronunciation
  - ("pencil" n (p eh1 n s ih l))
  - ("table" n (t ey1 b ax l))
- Need the right phoneme set
- Need other information
  - Part of speech
  - Lexical stress
  - Other information (Tone, Lexical accent …)
  - Syllable boundaries

Homograph Representation

- *Must distinguish different pronunciations*
  - *(“project” n (p r aa1 jh eh k t))*
  - *(“project” v (p r ax jh eh1 k t))*
  - *(“bass” n_music (b ey1 s))*
  - *(“bass” n_fish (b ae1 s))*
- *ASR multiple pronunciations*
  - *(“route” n (r uw t))*
  - *(“route(2)” n (r aw t))*

Pronunciation of Unknown Words

- *How do you pronounce new words*
- *4% of tokens (in news) are new*
- *You can't synthesis them without pronunciations*
- *You can't recognize them without pronunciations*
- *Letter-to-Sounds rules*
- *Grapheme-to-Phoneme rules*

LTS: Hand written

- *Hand written rules*
  - *[LeftContext] X [RightContext] -> Y*
  - *e.g.*
  - *c [h r] -> k*
  - *c [h] -> ch*
  - *c [i] -> s*
  - *c -> k*

LTS: Machine Learning Techniques

- Need an existing lexicon
  - Pronunciations: words and phones
  - But different number of letters and phones
- Need an alignment
  - Between letters and phones
  - checked -> ch eh k t

LTS: alignment

- checked -> ch eh k t

c	h	e	c	k	e	d
ch	_	eh	k	_	_	t

- Some letters go to nothing
- Some letters go to two phones
  - box -> b aa k-s
  - table -> t ey b ax-l -

Find alignment automatically

- *Epsilon scattering*
  - *Find all possible alignments*
  - *Estimate p(L,P) on each alignment*
  - *Find most probable alignment*
- *Hand seed*
  - *Hand specify allowable pairs*
  - *Estimate p(L,P) on each possible alignment*
  - *Find most probable alignment*
- *Statistical Machine Translation (IBM model 1)*
  - *Estimate p(L,P) on each possible alignment*
  - *Find most probable alignment*

Not everything aligns

- *0, 1, and 2 letter cases*
  - *e -> epsilon "moved"*
  - *x -> k-s, g-z "box" "example"*
  - *e -> y-uw "askew"*
- *Some alignments aren't sensible*
  - *dept -> d ih p aa r t m ax n t*
  - *cmu -> s iy eh m y uw*

Training LTS models

- *Use CART trees*
  - *One model for each letter*
- *Predict phone (epsilon, phone, dual phone)*
  - *From letter 3-context (and POS)*
- *# # # c h e c -> ch*
- *# # c h e c k -> _*
- *# c h e c k e -> eh*
- *c h e c k e d -> k*

LTS results

- Split lexicon into train/test 90%/10%
  - i.e. every tenth entry is extracted for testing

Lexicon	Letter Acc	Word Acc
OALD	95.80%	75.56%
CMUDICT	91.99%	57.80%
BRULEX	99.00%	93.03%
DE-CELEX	98.79%	89.38%
Thai	95.60%	68.76%

Example Tree

```
For letter V:
if (n.name is v)
    return _
    if (n.name is #)
            if (p.p.name is t)
                return f
                return v
            if (n.name is s)
                if (p.p.p.name is n)
                    return f
                    return v
                return v
```

But we need more than phones

- What about lexical stress
  - p r aa1 j eh k t -> p r aa j eh1 k t
- Two possibilities
  - A separate prediction model
  - Join model – introduce eh/eh1 (BETTER)

	LTP+S	LTPS
L no S	96.36%	96.27%
Letter	---	95.80%
W no S	76.92%	74.69%
Word	63.68%	74.56%

Does it really work

- 40K words from Time Magazine
  - 1775 (4.6%) not in OALD
  - LTS gets 70% correct (test set was 74%)

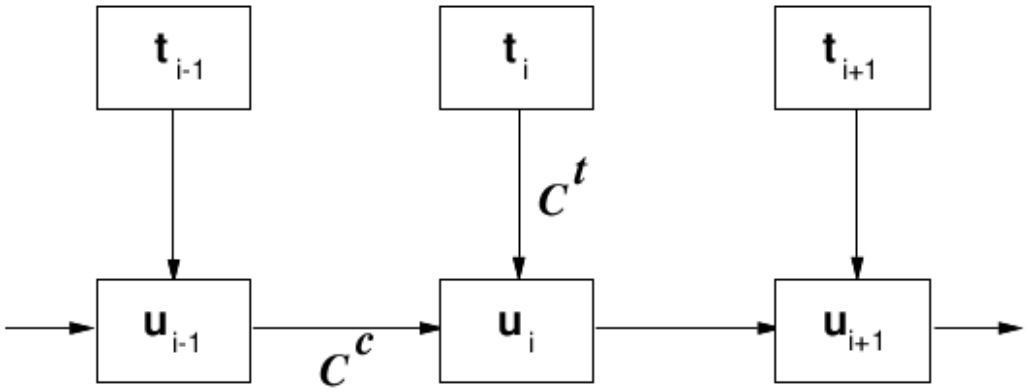
	Occurs	%
Names	1360	76.6
Unknown	351	19.8
US Spelling	57	3.2
Typos	7	0.4

Speech Synthesis Techniques

- *Unit selection*
- *Statistical parameter synthesis*
- *Automated voice building*
  - *Database design*
  - *Language portability*
- *Voice conversion*

Unit Selection

- Target cost and Join cost [Hunt and Black 96]
  - Target cost is distance from desired unit to actual unit in the databases
    - Based on phonetic, prosodic metrical context
  - Join cost is how well the selected units join



Clustering Units

- Cluster units [Donovan et al 96, Black et al 97]

$$Adist(U,V) = \begin{cases} \text{if } |V| > |U| \quad Adist(V,U) \\ \frac{WD*|U|}{|V|} * \sum_{i=1}^{|U|} \sum_{j=1}^{n} \frac{W_j.(abs(F_{ij}(U) - F_{(i*|V|/|U|)j}(V)))}{SD_j * n * |U|} \end{cases}$$

$|U|$ = number of frames in U
$F_{xy}(U)$ = parameter y of frame x of unit U
SD_j = standard deviation of parameter j
W_j = weight for parameter j
WD = duration penalty

Unit Selection Issues

- Cost metrics
  - Finding best weights, best techniques etc
- Database design
  - Best database coverage
- Automatic labeling accuracy
  - Finding errors/confidence
- Limited domain:
  - Target the databases to a particular application
  - Talking clocks
  - Targeted domain synthesis

Old vs New

Unit Selection: 

- large carefully labelled database
- quality good when good examples available
- quality will sometimes be bad
- no control of prosody

Parametric Synthesis: 

- smaller less carefully labelled database
- quality consistent
- resynthesis requires vocoder, (buzzy)
- can (must) control prosody
  - model size much smaller than Unit DB

Parametric Synthesis

- Probabilistic Models

$$argmax(P(O|W))$$

- Simplification

$$argmax(P(o_0|W), P(o_1|W), ..., P(o_n|W))$$

- Generative model
  - Predict acoustic frames from text

Spoken Dialog Systems

- *Information giving*
  - *Flights, buses, stocks weather*
  - *Driving directions*
  - *News*
- *Information navigators*
  - *Read your mail*
  - *Search the web*
  - *Answer questions*
- *Provide personalities*
  - *Game characters (NPC), toys, robots, chatbots*
- *Speech-to-speech translation*
  - *Cross-lingual interaction*

Dialog Types

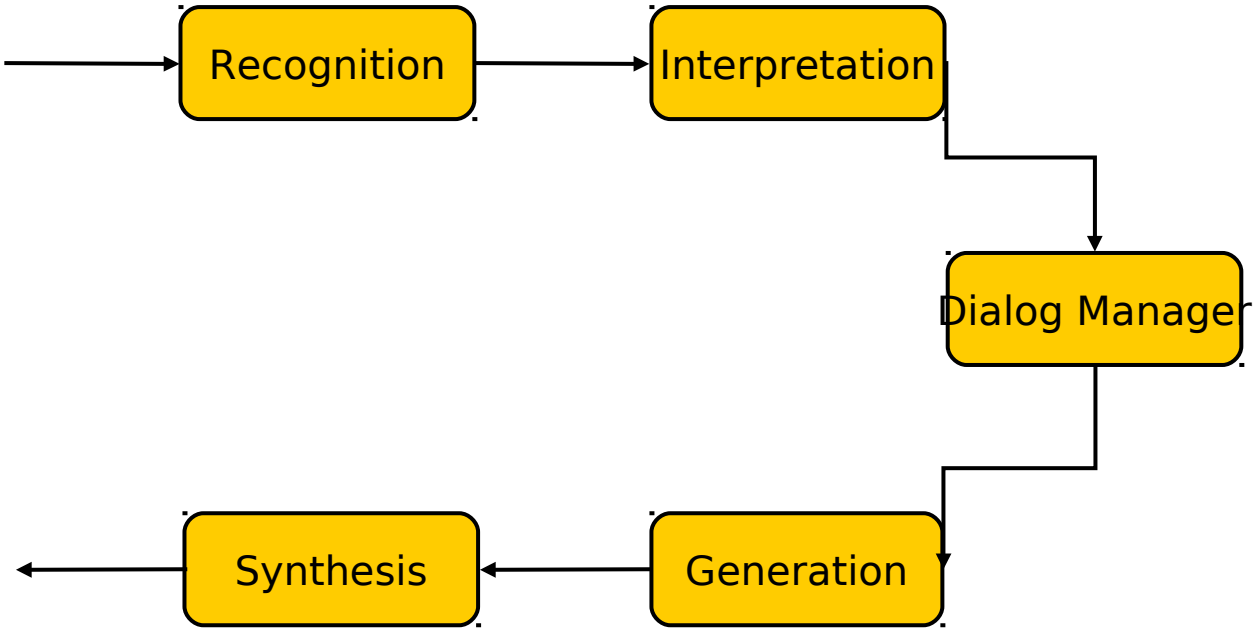
- System initiative
  - Form-filling paradigm
  - Can switch language models at each turn
  - Can "know" which is likely to be said
- Mixed initiative
  - Users can go where they like
  - System or user can lead the discussion
- Classifying:
  - Users can say what they like
  - But really only "N" operations possible
  - E.g. AT&T? "How may I help you?"
- Non-task oriented

System Initiative

- *Let's Go Bus Information*
  - *412 268 3526*
  - *Provides bus information for Pittsburgh*


- *Tell Me*
  - *Company getting others to build systems*
  - *Stocks, weather, entertainment*
  - *1 800 555 8355*

SDS Architecture



SDS Components

- *Interpretation*
  - *Parsing and Information Extraction*
  - *(Ignore politeness and find the departure stop)*
- *Generation*
  - *From SQL table output from DB*
  - *Generate “nice” text to say*

Siri-like Assistants

- *Advantages*
  - *Hard to type/select things on phone*
  - *Can use context (location, contacts, calendar)*
- *Target common tasks*
  - *Calling, sending messages, calendar*
  - *Fall back on google lookup*

SPDA: Scope

- *"Call John"*
- *"Call John, Bill and Mary and setup a meeting sometime next week about Plan B that's fits my schedule"*
- *"Make a reservation at a local Chinese restaurant for 4 at 8pm."*
- *"You should call your mom as its her birthday"*
- *"I have sent flowers to your mom as its her birthday"*

CALO (DARPA)

- Cognitive Assistant that Learns Online
  - DARPA project (2003-2008)
  - Led by SRI (involved many sites, including CMU)
- Personal Assistant that Learns (Pal)
  - Answers questions
  - Learn from experience
  - Take initiative
- Spin-off company -> SIRI
  - Aquired by Apple in April 2010