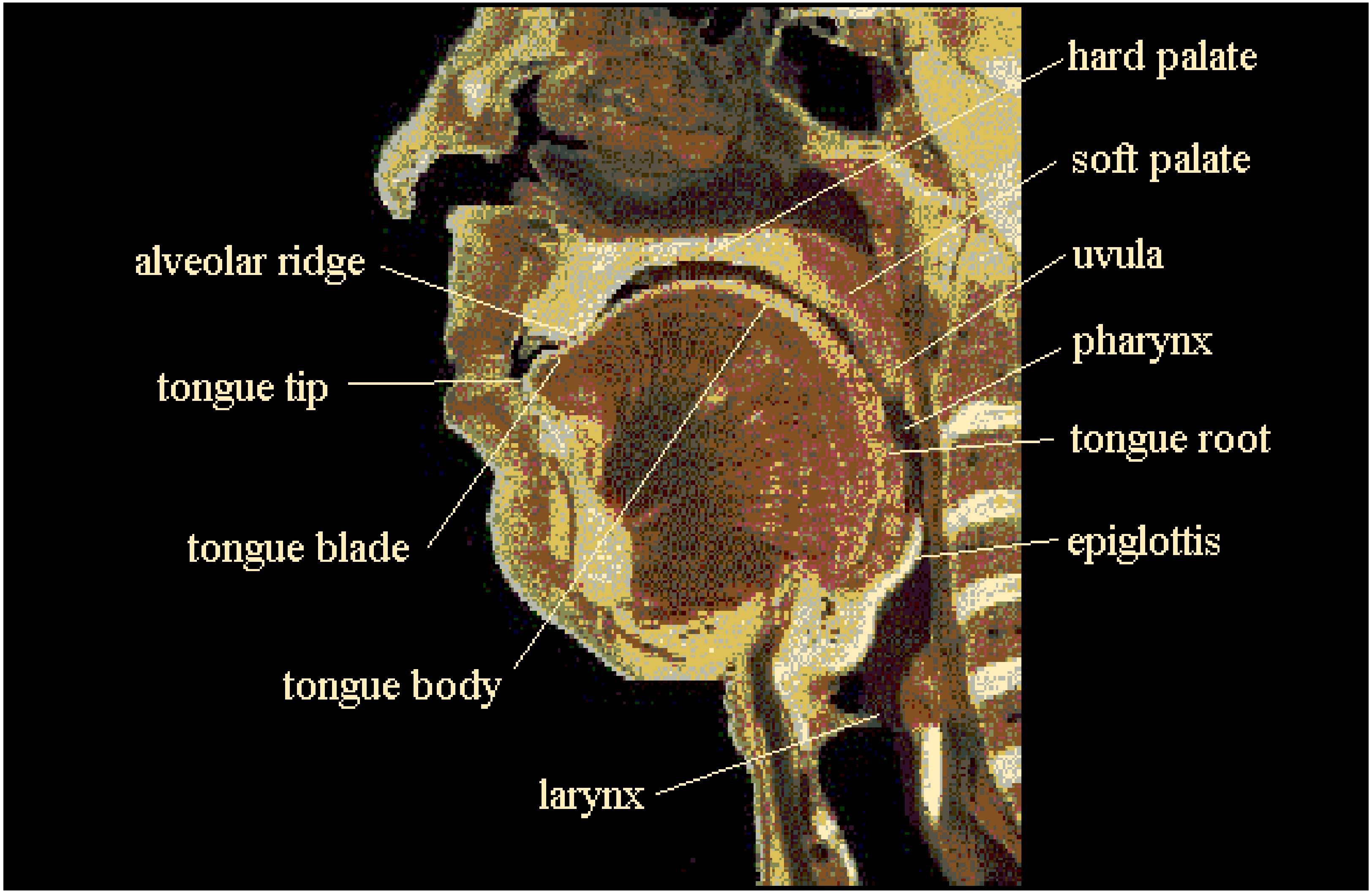
Speech Processing

Using Speech with Computers

Overview

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

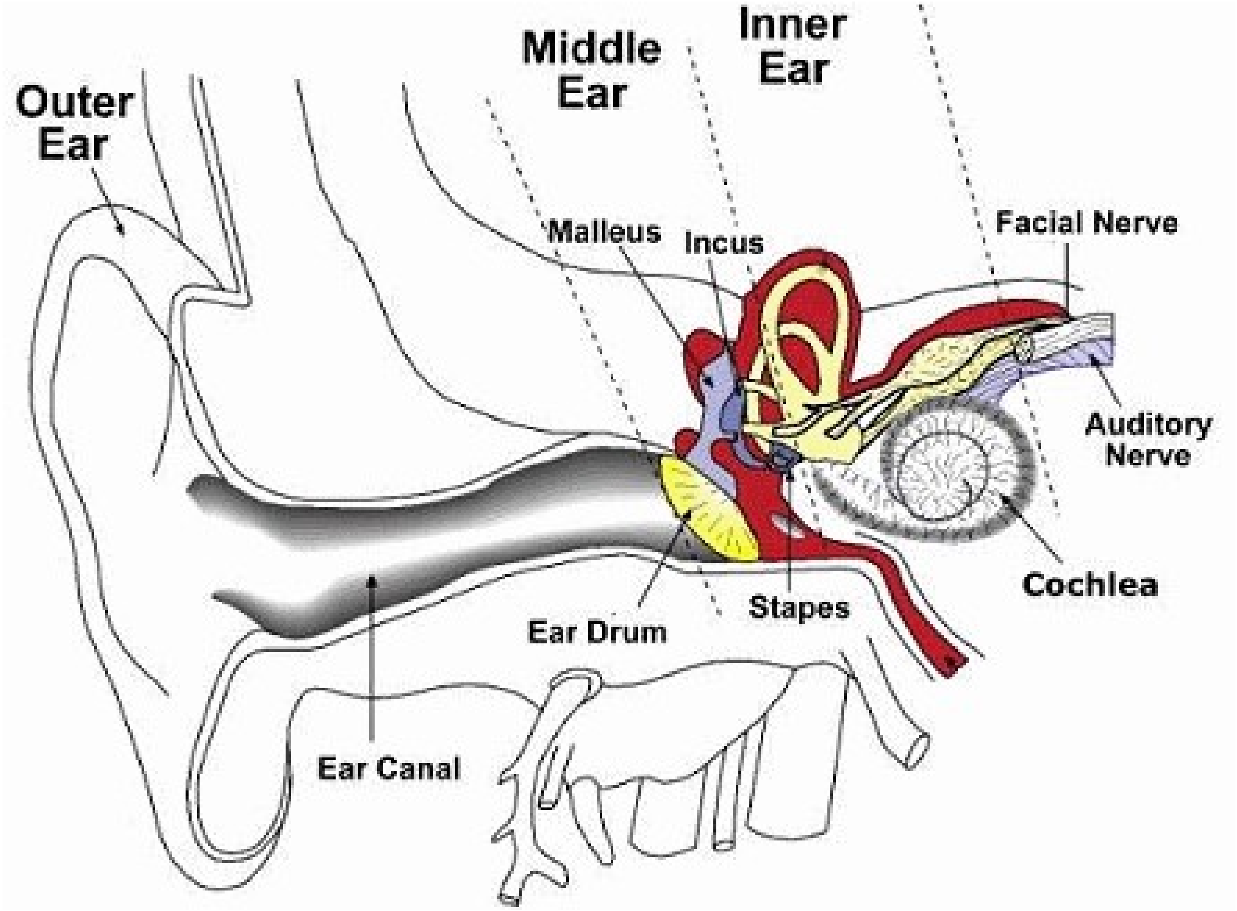
The vocal tract



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

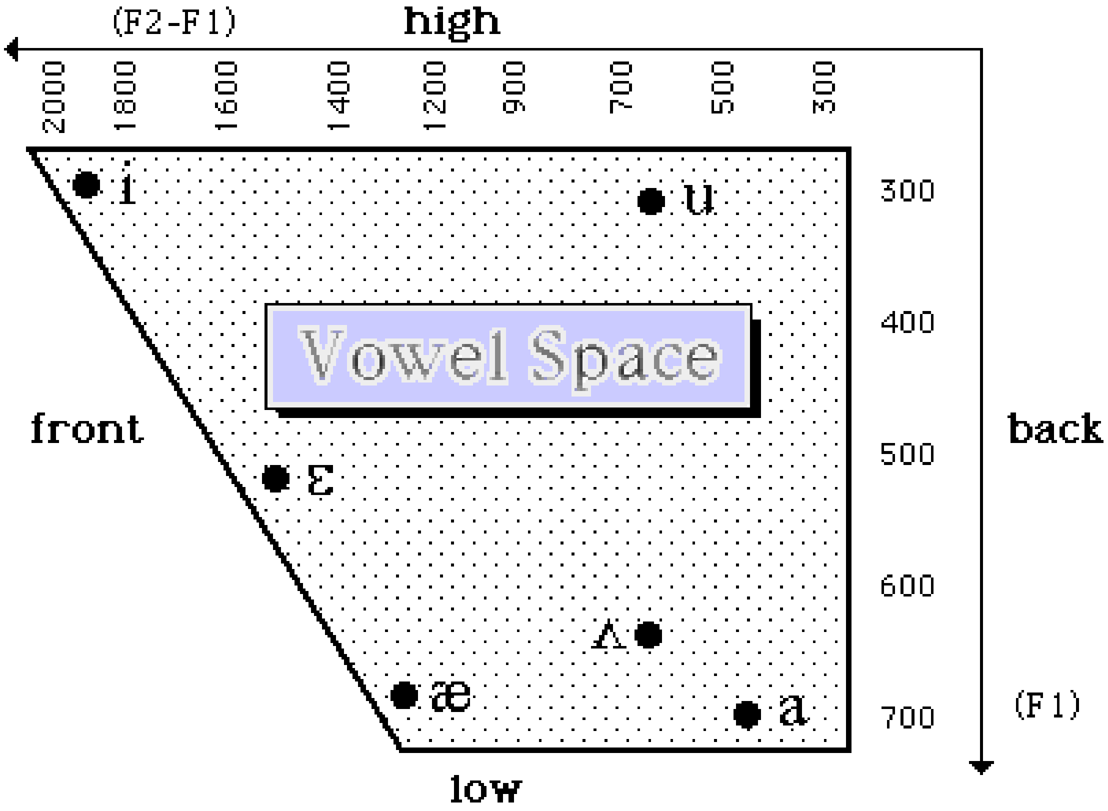
- *Defined as fundamental units of speech*
  - *If you change it, it (can) change the meaning*

“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

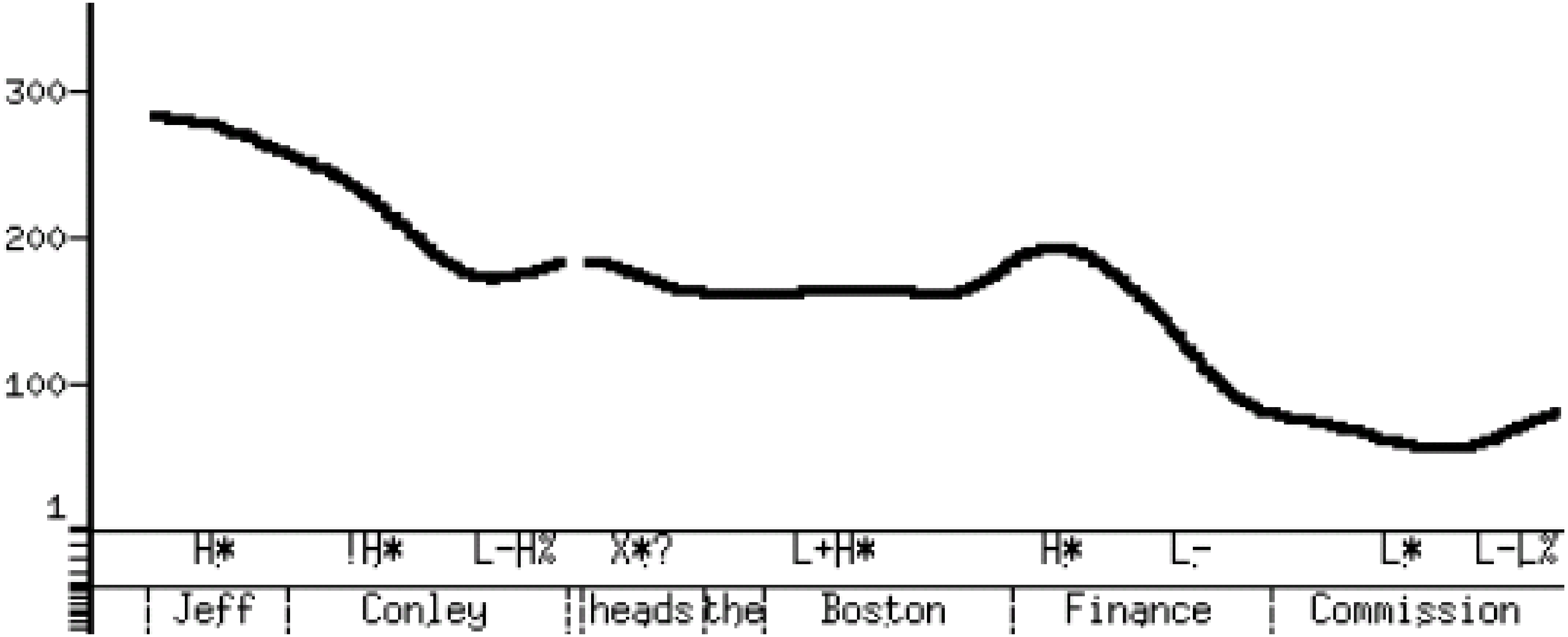
- *Vowel length*
  - *Bit vs beat*
  - *Japanese: shujin (husband) vs shuujin (prisoner)*
- *Tones*
  - *F0 (tune) used phonetically*
  - *Chinese, Thai, Burmese*
- *Clicks*
  - *Xhosa*

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

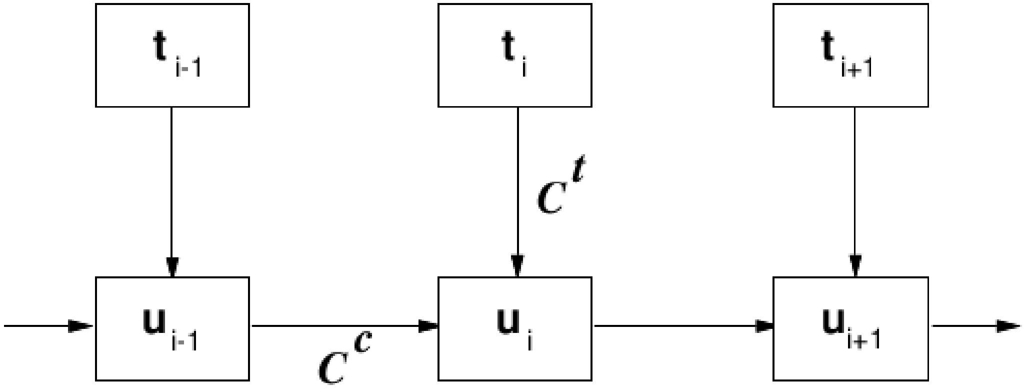
- *Formant synthesis*
- *Random word/phrase concatenation*
- *Phone concatenation*
- *Diphone concatenation*
- *Sub-word unit selection*
- *Cluster based unit selection*
- *Statistical Parametric Synthesis*
- *Wavenet Neural Synthesis*

Speech Synthesis Techniques

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

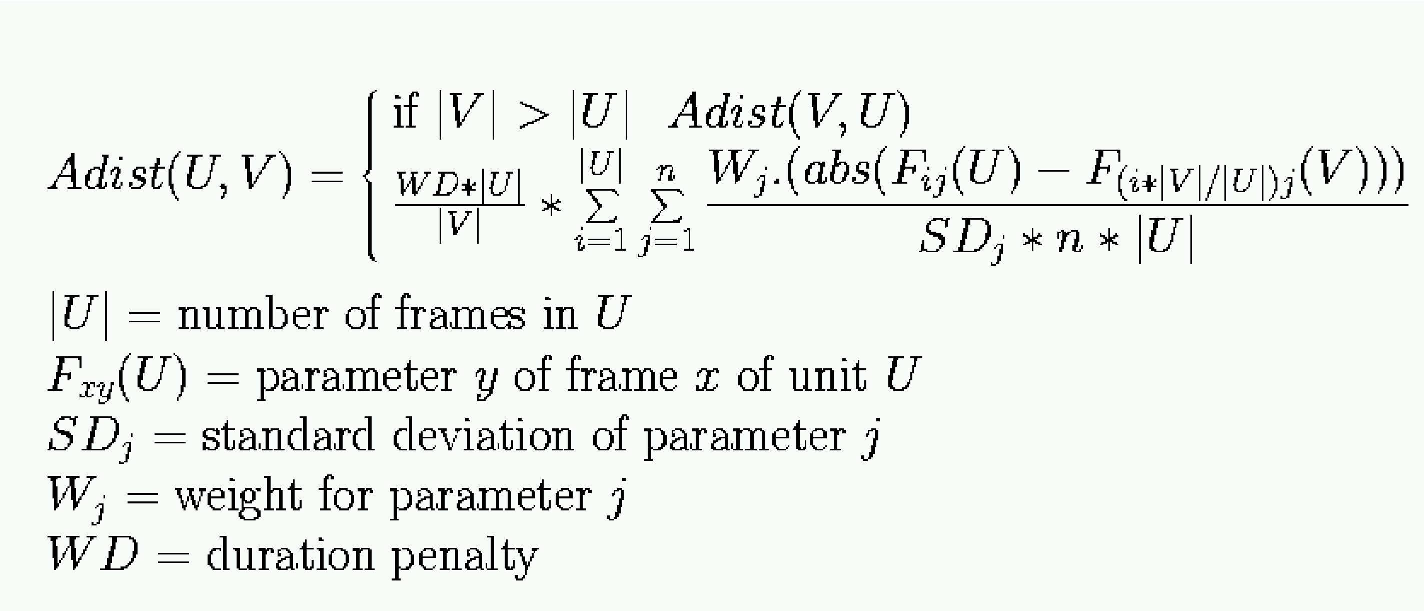
Unit Selection

- 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]



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

$|U| =$ 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