



Speech Processing 15-492/18-492

Speech Synthesis
Waveform generation

Speech Synthesis

- ◆ *Text Analysis*
 - *Chunking, tokenization, token expansion*
- ◆ *Linguistic Analysis*
 - *Pronunciations*
 - *Prosody*
- ◆ *Waveform generation*
 - *From phones and prosody to waveforms*

Physical Models

- Blowing air through tubes...

- von Kempelen's synthesizer 1791



- Synthesis by physical models








- Homer Dudley's Voder. 1939



More Computation – More Data

- ◆ *Formant synthesis (60s-80s)*
 - *Waveform construction from components*
- ◆ *Diphone synthesis (80s-90s)*
 - *Waveform by concatenation of small number of instances of speech*
- ◆ *Unit selection (90s-00s)*
 - *Waveform by concatenation of very large number of instances of speech*
- ◆ *Statistical Parametric Synthesis (00s-..)*
 - *Waveform construction from parametric models*

Waveform Generation

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

Concatenative Synthesis

- ◆ *Use human speech*
- ◆ *Need to design database*
- ◆ *Need to carefully label it*
- ◆ *Need to impose prosody on selections*
- ◆ *Results depend of db contents*
 - *You get good synthesis*
 - *But style is like the databases*

Diphone Synthesis

- ◆ *Use databases of natural speech*
- ◆ *From mid-phone to mid-phone*
 - *Requires phones squared diphones*
- ◆ *Needs very good definition of phoneset*
 - *Dialect of speaker becomes important*

Diphone Databases

◆ *Collect nonsense carrier words*

- *t aa b aa b aa*
- *t aa m aa m aa*
- *t iy b iy b iy*
- *Good for coverage, consistent*

◆ *Collect from “natural” words*

- *Quebecois arguments (19)*
- *Arkansas arranging (11)*
- *Good for naturalness, but maybe not consistent*

Recording Databases

- ◆ *Recording in best conditions possible*
 - *Recording studio*
 - *Head mounted mike*
 - *Repeatable conditions*
- ◆ *Explain to voice talent*
 - *Get *signed* permission*
 - *You are going to steal their voice!*

Diphone Limitations

- ◆ *Only get fixed inventory*
 - *Need more than phone-phone*
 - *Need stressed, positional examples*
 - *What about consonant clusters*
- ◆ *Get more representative samples*
 - *Larger databases*
 - *More natural*
 - *Harder to ensure it is correct*

Database Design

◆ *Require:*

- *Good phonetic coverage*
- *Good prosodic coverage*
- *Easy to read sentences (few mistakes)*
- *Consistent delivery*

Database design

- ◆ *From large databases of text*
 - *E.g. out-of-copyright books*
- ◆ *Find “nice” sentences*
 - *Contain only high frequency words*
 - *5-15 words long*
 - *No homographs*
- ◆ *Greedily select “nice” sentences with*
 - *Best phone/diphone/triphone coverage*
 - *Best characters/dicharacter/tricharacter coverage*
- ◆ *Consider multiple genres*
 - *Novels, news, bus stops (domain dependent)*

CMU ARCTIC Databases

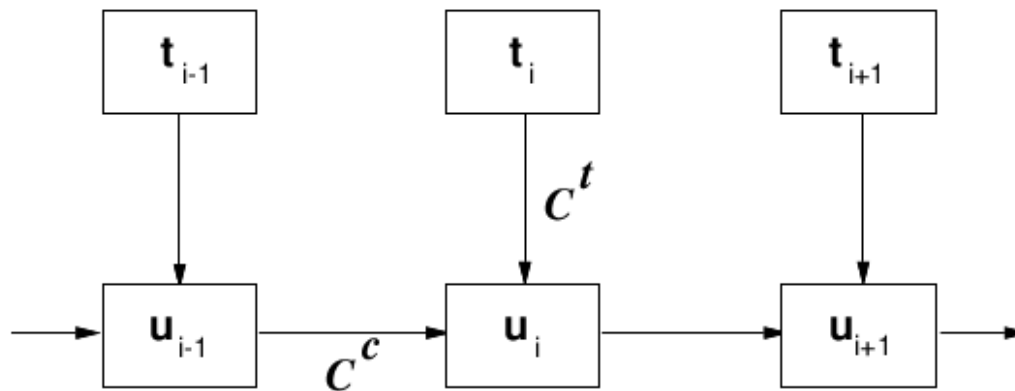
- ◆ *1132 sentences (about an hour of speech)*
 - *Author of the danger trail, Phillip Steels etc.*
 - *...*
- ◆ *9 different speakers*
 - *Different English accents*
- ◆ *Technique used for other languages*

Unit Selection Speech Synthesis

- ◆ *Select appropriate sub-word units from databases of natural speech.*
- ◆ *Not simply word concatenation*
- ◆ *Not simply longest phrase*
- ◆ *Balance*
 - *Appropriate unit*
 - *Good join costs*

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]

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

$|U|$ = 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 