Speech and Language Processing

Chapter 8 of SLP
Speech Synthesis / Waveform synthesis

Waveform Synthesis

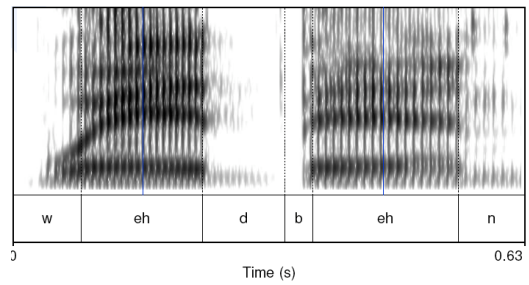
- Given:
  - String of phones
  - Prosody
    - Desired F0 for entire utterance
    - Duration for each phone
    - Stress value for each phone, possibly accent value
- Generate:
  - Waveforms

Diphone TTS architecture

- Training:
  - Choose units (kinds of diphones)
  - Record 1 speaker saying 1 example of each diphone
  - Mark the boundaries of each diphones,
    - cut each diphone out and create a diphone database
- Synthesizing an utterance,
  - grab relevant sequence of diphones from database
  - Concatenate the diphones, doing slight signal processing at boundaries
  - use signal processing to change the prosody (F0, energy, duration) of selected sequence of diphones

Diphones

- Mid-phone is more stable than edge:



Diphones

- mid-phone is more stable than edge
- Need O(phone^2) number of units
  - Some combinations don't exist (hopefully)
  - ATT (Olive et al. 1998) system had 43 phones
    - 1849 possible diphones
    - Phonotactics ([h] only occurs before vowels), don't need to keep diphones across silence
    - Only 1172 actual diphones
  - May include stress, consonant clusters
    - So could have more
  - Lots of phonetic knowledge in design
- Database relatively small (by today's standards)
  - Around 8 megabytes for English (16 KHz 16 bit)

Voice

- Speaker
  - Called a **voice talent**
- Diphone database
  - Called a **voice**

MBROLA

- Difoon synthese systeem (open source) (Thierry Dutoit, Mons, België)

Als ingrediënten opgeven, voor elke klank:

- Foneem
- Toonhoogte
- Duur

MBROLA procedure

Nodig:

- MBROLA difoonset
- Stuurgegevens in .pho fil
  fonemen, toonhoogtes, duren

MBROLA maakt .wav file

$mbrola mbrola/nl2/nl2 woord.pho woord.wav

MBROLA synthese

– **duur (ms)** – **toonhoogte (Hz)** – **%**

```
; Utterance: "Hallo!"
_  100  100 120
h   96
A   48
l   76    5 100   75 120
o  224   25  85
_  100   40  70
```

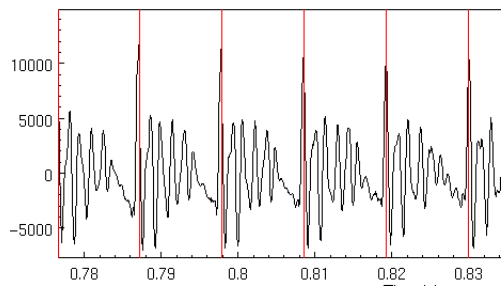
percentages

Prosodic Modification

- Modifying pitch and duration *independently*
- Changing sample rate modifies both:
  - Chipmunk speech
- Duration: duplicate/remove parts of the signal
- Pitch: resample to change pitch

Text from Alan Black

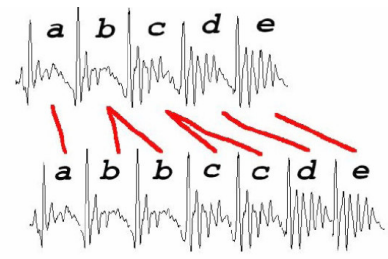
Speech as Short Term signals



Alan Black

Duration modification

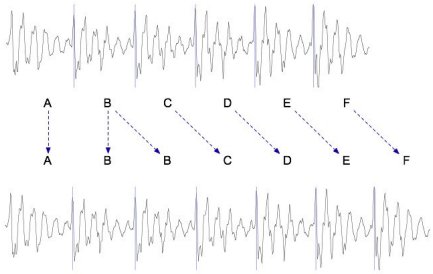
- Duplicate/remove short term signals



Slide from Richard Sproat

Duration modification

- Duplicate/remove short term signals



Pitch Modification

- Move short-term signals closer together/further apart

Slide from Richard Sproat

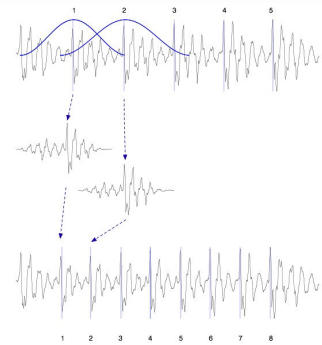
TD-PSOLA ™

- Time-Domain Pitch Synchronous Overlap and Add
- Patented by France Telecom (CNET)
- Very efficient
  - No FFT (or inverse FFT) required
- Can modify Hz up to two times or by half

Slide from Richard Sproat

TD-PSOLA ™

- Time-Domain Pitch Synchronous Overlap and Add
- Patented by France Telecom (CNET)
- Windowed
- Pitch-synchronous
- Overlap-and-add
- Very efficient
- Can modify Hz up to two times or by half



Unit Selection Synthesis

- Generalization of the diphone intuition
  - Larger units
    - From diphones to sentences
  - Many many copies of each unit
    - 10 hours of speech instead of 1500 diphones (a few minutes of speech)

Unit Selection Intuition

- Given a big database
- Find the unit in the database that is the *best* to synthesize some target segment
- What does "best" mean?
  - "Target cost": Closest match to the target description, in terms of
    - Phonetic context
    - F0, stress, phrase position
  - "Join cost": Best join with neighboring units
    - Matching formants + other spectral characteristics
    - Matching energy
    - Matching F0

Targets and Target Costs

- Target cost $T(u_t,s_t)$: How well the target specification s_t matches the potential unit in the database u_t
- Features, costs, and weights
- Examples:
  - /ih-t/ +stress, phrase internal, high F0, content word
  - /n-t/ -stress, phrase final, high F0, function word
  - /dh-ax/ -stress, phrase initial, low F0, word "the"

Target Costs

- Comprised of k subcosts
  - Stress
  - Phrase position
  - F0
  - Phone duration
  - Lexical identity
- Target cost for a unit:

$$C^t(t_i,u_i)=\sum_{k=1}^{p} w_k^t C_k^t(t_i,u_i)$$

difoonaansluiting(skosten)

- pa
- ka
- ta
- ap
- ak
- at

Dit zijn meestal zes verschillende opnamen, maar dat geeft spectrale verschillen bij de aansluiting:

pa – ap, pa – ak, pa – at

ka – ap, ka – ak, ka – at

ta – ap, ta – ak, ta – at

Join (Concatenation) Cost

- Measure of smoothness of join
- Measured between two database units (target is irrelevant)
- Features, costs, and weights
- Comprised of k subcosts:
  - Spectral features
  - F0
  - Energy
- Join cost:

$$C^j(u_{i-1},u_i)=\sum_{k=1}^{p} w_k^j C_k^j(u_{i-1},u_i)$$

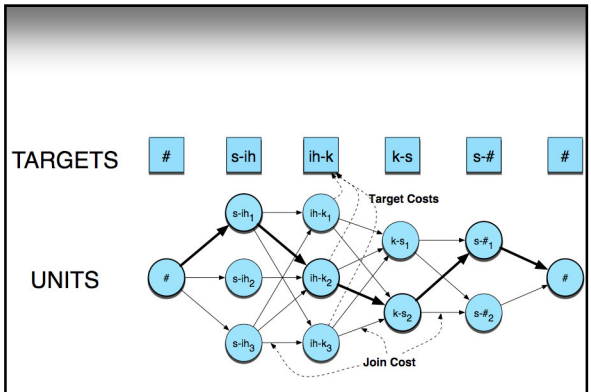
Total Costs

- Hunt and Black 1996
- We now have weights (per phone type) for features set between target and database units
- Find best path of units through database that minimize:

$$C(t_1^n,u_1^n)=\sum_{i=1}^{n} C^{target}(t_i,u_i)+\sum_{i=2}^{n} C^{join}(u_{i-1},u_i)$$

$$\hat{u}_1^n=\underset{u_1,\ldots,u_n}{\operatorname{argmin}} C(t_1^n,u_1^n)$$

- Standard problem solvable with Viterbi search with beam width constraint for pruning



Unit Selection Summary

- Advantages
  - Quality is far superior to diphones
  - Natural prosody selection sounds better
- Disadvantages:
  - Quality can be very bad in places
    - HCI problem: mix of very good and very bad is quite annoying
  - Synthesis is computationally expensive
  - Can't synthesize everything you want:
    - Diphone technique can move emphasis
    - Unit selection gives good (but possibly incorrect) result