Back-End Synthesis and Evaluation

Julia Hirschberg
CS 4706
(*Thanks to Dan and Jim)

Outline

- Waveform Generation
  - Diphones
  - Unit Selection
  - HMM Synthesis
- TTS Evaluation
  - Objective Measures
  - Subjective Measures

Modern TTS systems

- 1960's first full TTS: Umeda et al (1968)
- 1970's
  - Joe Olive 1977 concatenation of linear-prediction diphones
  - Speak and Spell
- 1980's
  - 1979 MIT MITalk (Allen, Hunnicut, Klatt)
- 1990's-present
  - Diphone synthesis
  - Unit selection synthesis
  - HMM synthesis

Architectures of Modern Synthesis

- Articulatory Synthesis:
  - Model movements of articulators and acoustics of vocal tract
- Formant Synthesis:
  - Start with acoustics, create rules/filters to create each formant
- Concatenative Synthesis:
  - Use databases of stored speech to assemble new utterances.
- HMM Synthesis

Text from Richard Sproat slides

Formant Synthesis

- Were the most common commercial systems while computers were relatively underpowered.
- 1979 MIT MITalk (Allen, Hunnicut, Klatt)
- 1983 DECtalk system
- The voice of Stephen Hawking

Concatenative Synthesis

- All current commercial systems.
- Diphone Synthesis
  - Units are diphones; middle of one phone to middle of next.
  - Why? Middle of phone is steady state.
  - Record 1 speaker saying each diphone
- Unit Selection Synthesis
  - Larger units
  - Record 10 hours or more, so have multiple copies of each unit
  - Use search to find best sequence of units

TTS Demos (all are Unit-Selection)

- Festival
  - http://www-2.cs.cmu.edu/~awb/festival_demos/index.html
- Cepstral
  - http://www.cepstral.com/cgi-bin/demos/general
- IBM
  - http://www-306.ibm.com/software/pervasive/tech/demos/tts.shtml

How do we get from Text to Speech?

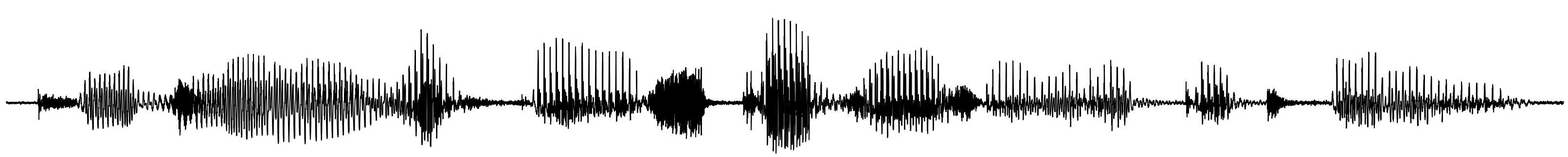
- TTS "Backend" only covers the segments+f0+duration to waveform part
- A full system needs to go all the way from random text to sound

Two steps

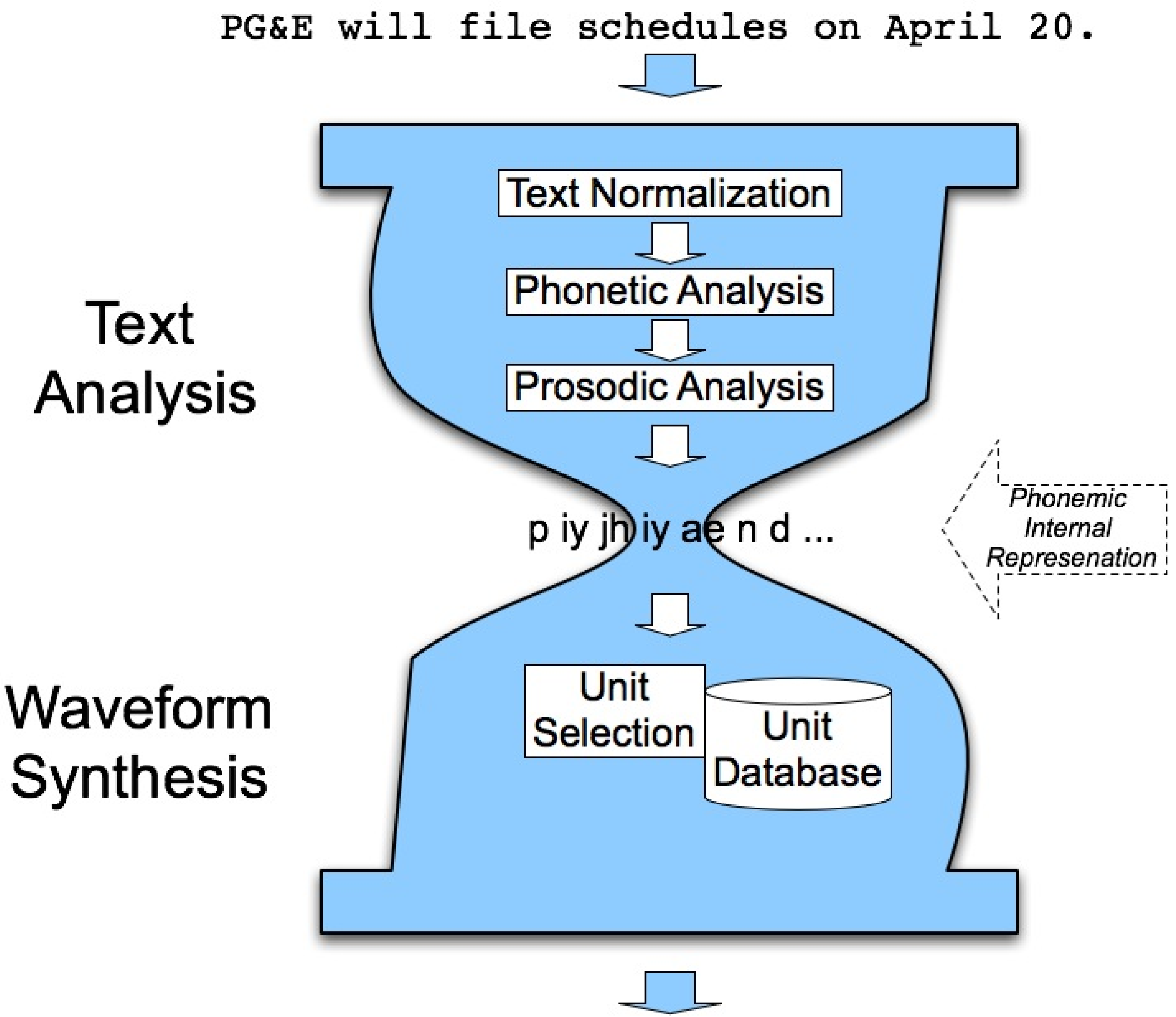
- `PG&E will file schedules on April 20.`
- TEXT ANALYSIS: Text into intermediate representation:

			*			*			* L-L%
P	G	AND	E	WILL	FILE	SCHEDULES	ON	APRIL	TWENTIETH
p iy	jh iy	ae n d	iy	w ih l	f ay l	s k eh jh ax l z	aa n	ey p r ih l	t w eh n t iy ax th

- WAVEFORM SYNTHESIS: From the intermediate representation into waveform



The Hourglass



Waveform Synthesis

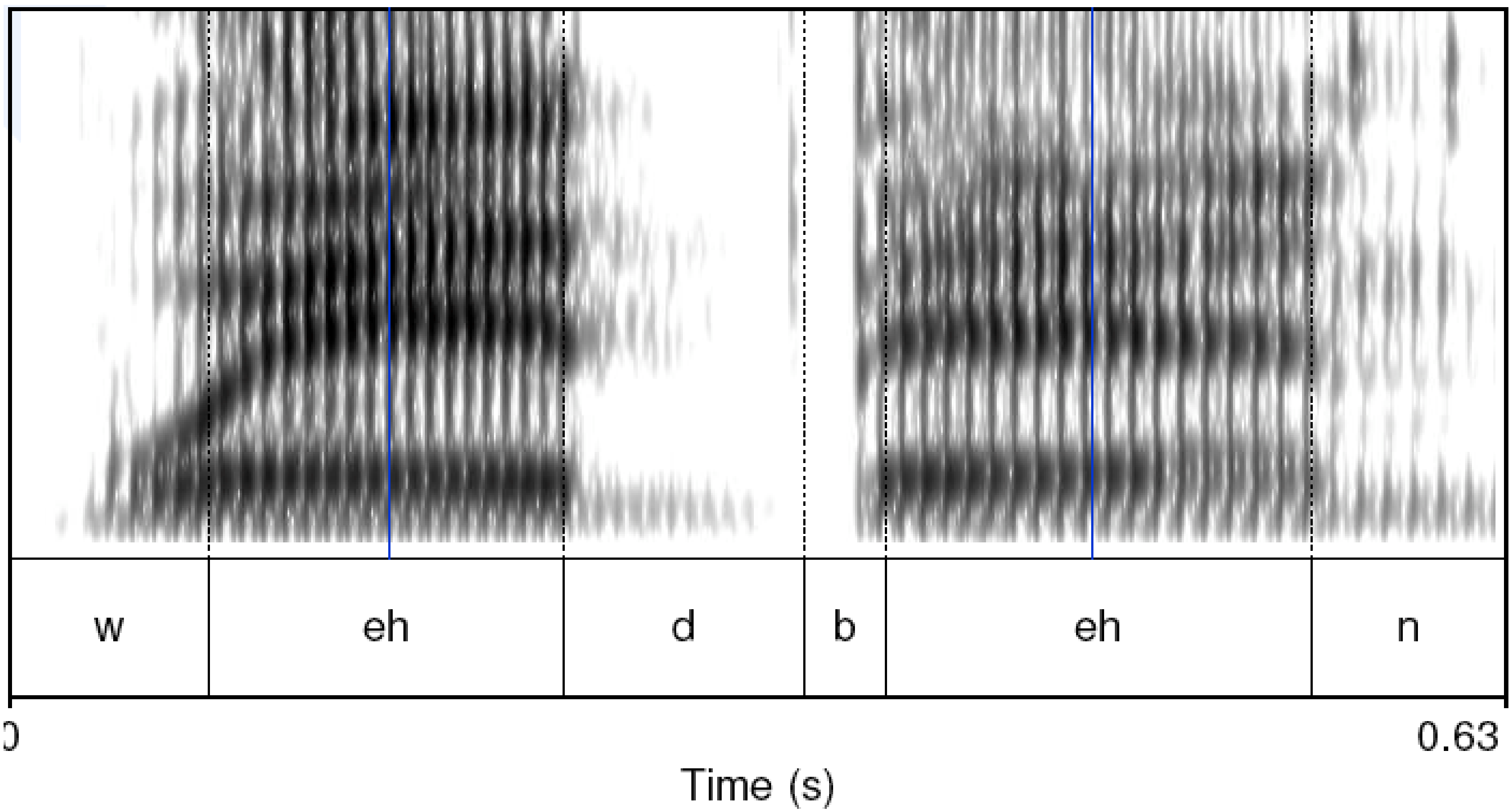
- Given:
  - String of phones
  - Prosody
    - Desired F0 for entire utterance
    - Duration for each phone
    - Stress value for each phone, possibly accent value
- Generate:
  - Waveforms

Diphone TTS Architecture

- Training:
  - Choose units (kinds of diphones)
  - Record 1 speaker saying 1 example of each diphone
  - Mark the boundaries of each diphone,
    - Cut each diphone out to create a diphone database
- Synthesizing an utterance,
  - Select relevant set of diphones from database
  - Concatenate them in order, doing minor signal processing at boundaries
  - Use signal processing techniques to change prosody (F0, energy, duration) of sequence

Diphones

- Where is the stable region?



Diphones

- Middle of phone more stable than edges
- Need O(phone^2) number of units
  - Some phone-phone sequences don't exist
  - ATT (Olive et al.'98) 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
  - But…may want to include stress or accent differences, consonant clusters, etc., so may need more
  - Requires much knowledge of phonetics in design
- Database relatively small (by today's standards)
  - Around 8 megabytes for English (16 KHz 16 bit)

Voice

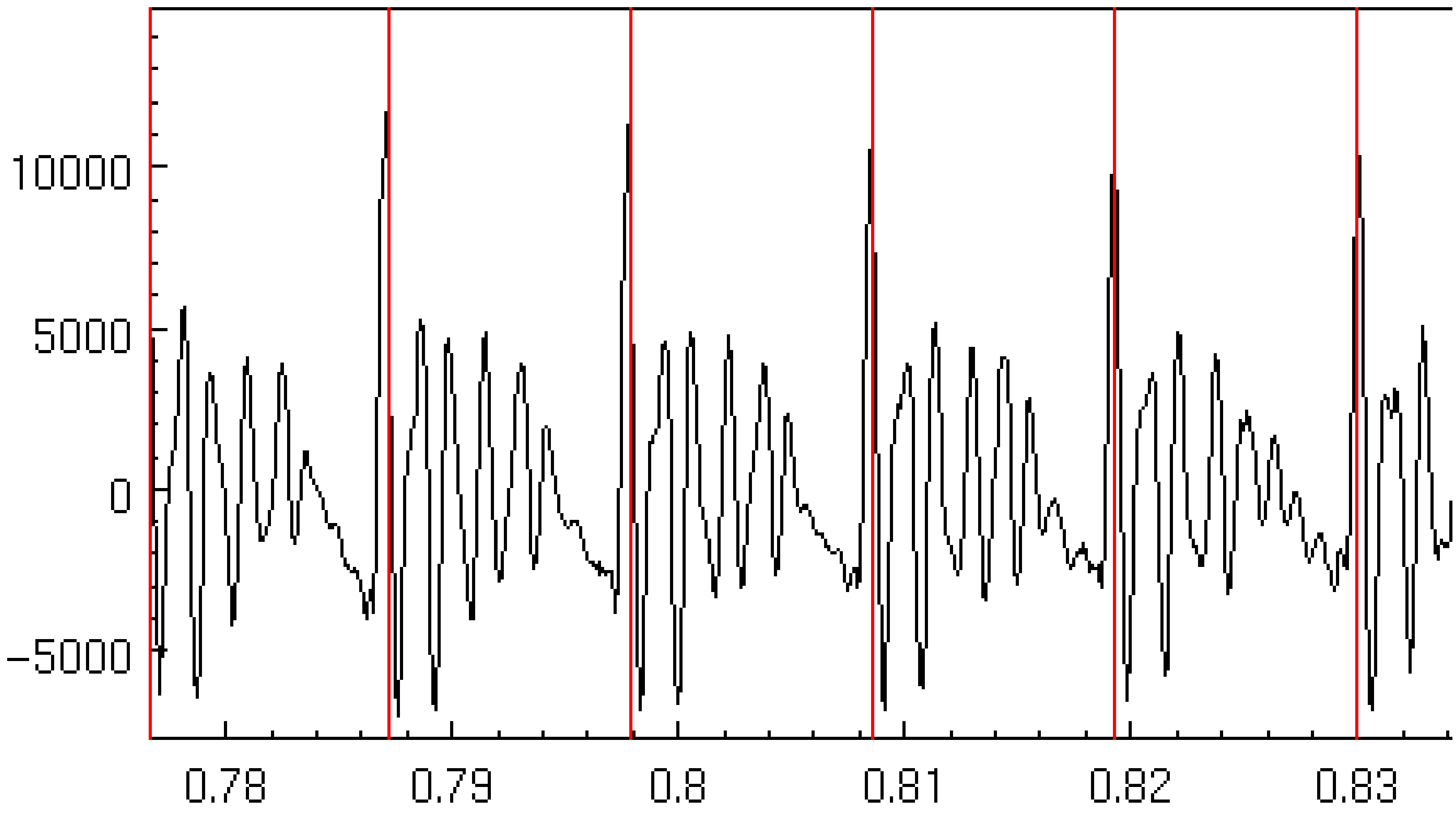
- Speaker
  - Called the **voice talent**
  - **How to choose?**
- Diphone database
  - Called a **voice**
  - Modern TTS systems have multiple voices

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

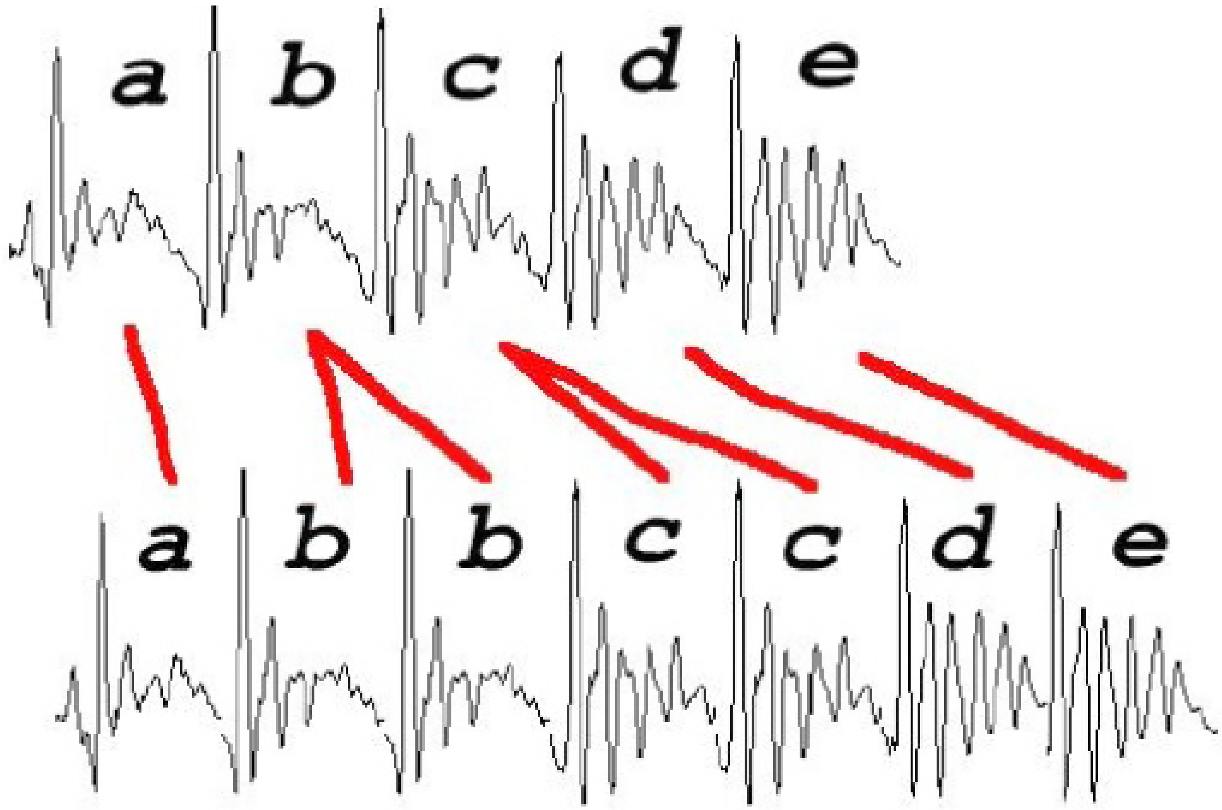
Text from Alan Black

Speech as Short Term signals



Duration modification

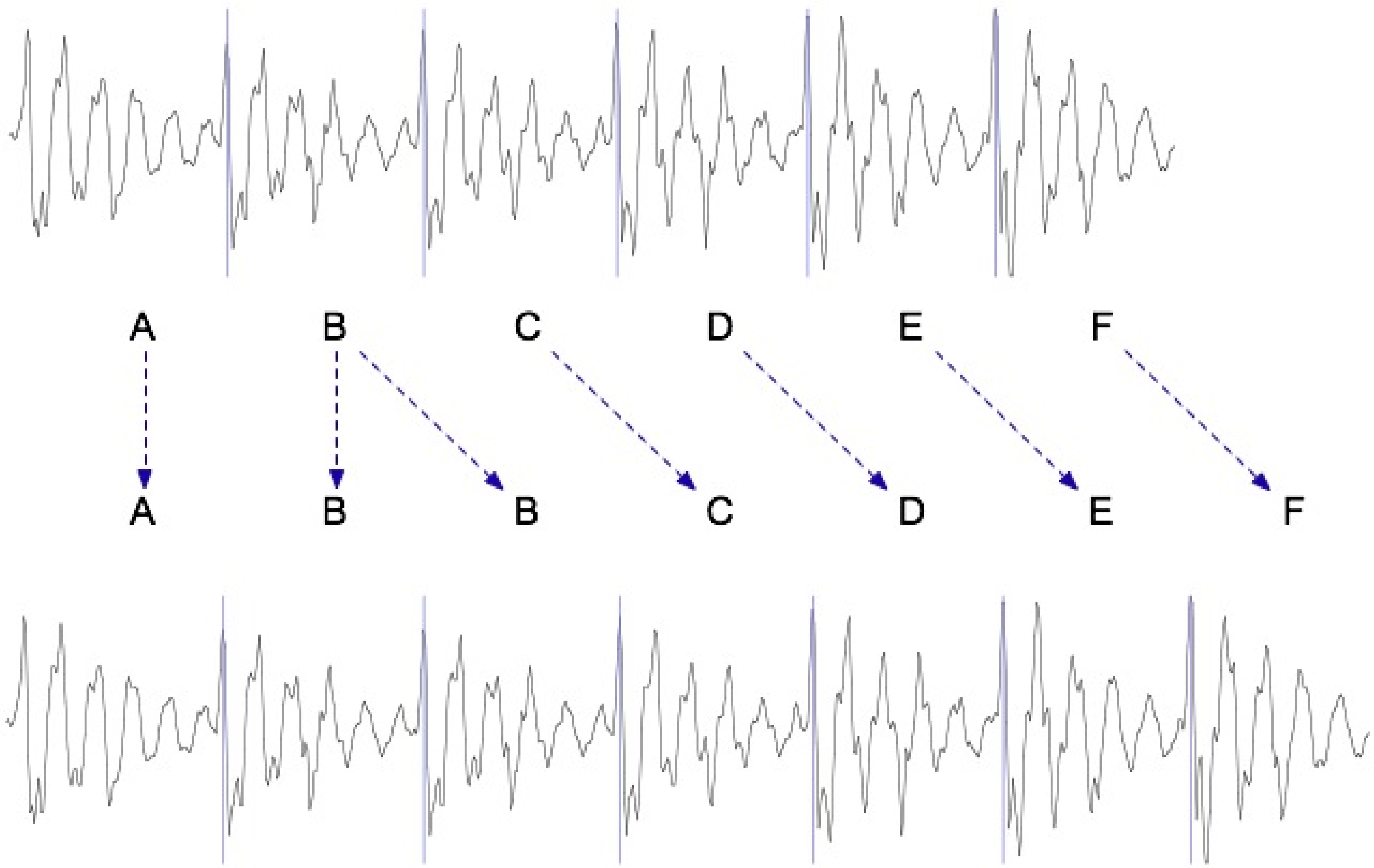
- Duplicate/remove short term signals



Slide from Richard Sproat

Duration modification

- Duplicate/remove short term signals



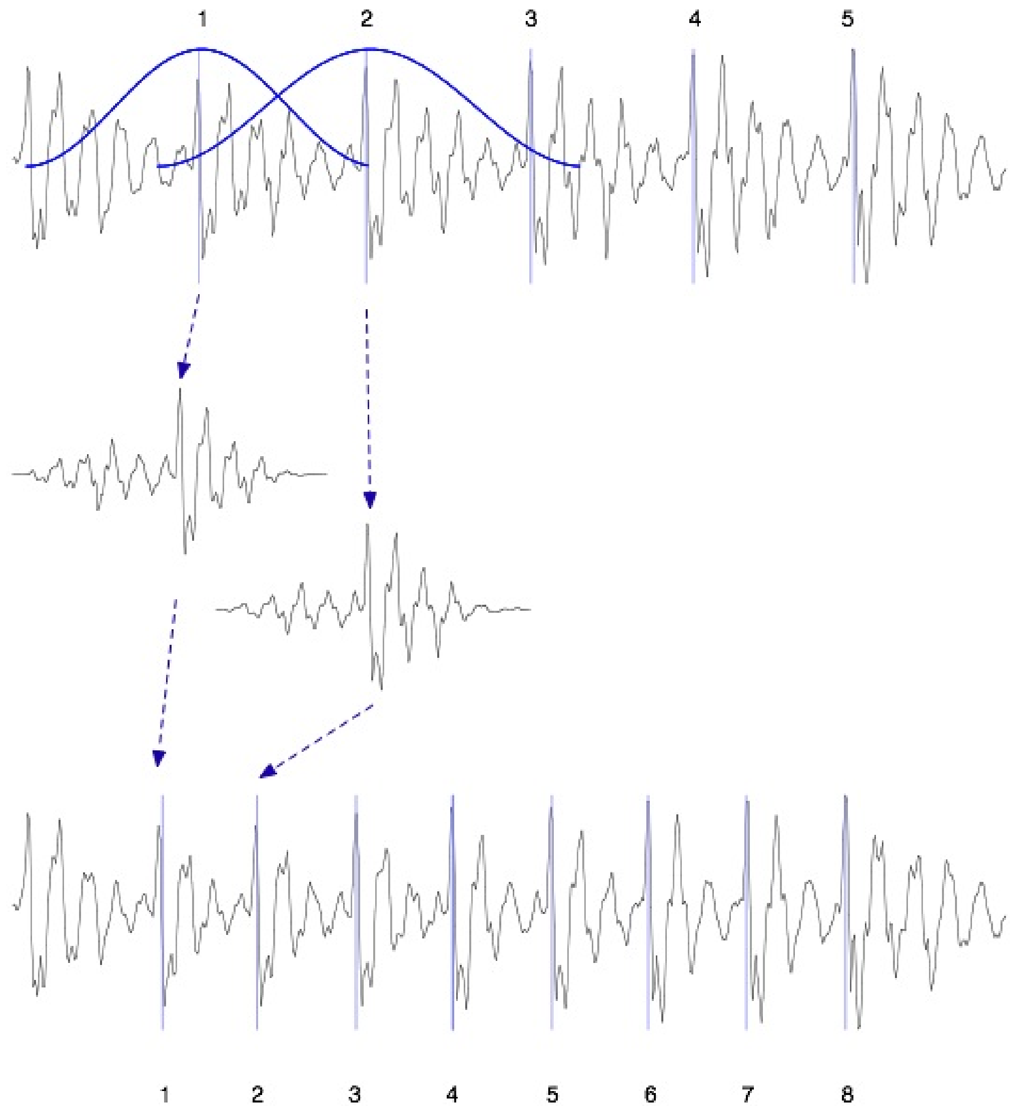
Pitch Modification

- Move short-term signals closer together/further apart: more cycles per sec means higher pitch and vice versa
- Add frames as needed to maintain desired duration

Slide from Richard Sproat

TD-PSOLA ™

- Time-Domain Pitch Synchronous Overlap and Add
- Patented by France Telecom (CNET)
- Epoch detection and windowing
- 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
  - Record many copies of each unit
    - E.g.10 hours of speech instead of 1500 diphones (a few minutes of speech)

Unit Selection Intuition

- Given a large ***labeled*** database, find the unit that best matches the desired synthesis specification
- What does “best” mean?
  - Target cost: Find closest match in terms of
    - Phonetic context
    - F0, stress, phrase position
  - Join cost: Find best join with neighboring units
    - Matching formants + other spectral characteristics
    - Matching energy
    - Matching F0

Targets and Target Costs

- Target cost T(ut,st): How well does the target specification st match the potential unit in the database ut?
- Goal: find the unit ***least unlike*** the target
- Examples of labeled diphone midpoints
  - /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*
- Costs of different features have different weights

Target Costs

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

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

Join (Concatenation) Cost

- Measure of smoothness of join between two database units (target irrelevant)
- Features, costs, and weights
- Comprised of p subcosts:
  - Spectral features
  - F0
  - Energy
- Join cost: $$C^j(u_{i-1},u_i) = \sum_{k=1}^{p} w_k^j C_k^j(u_{i-1},u_i)$$

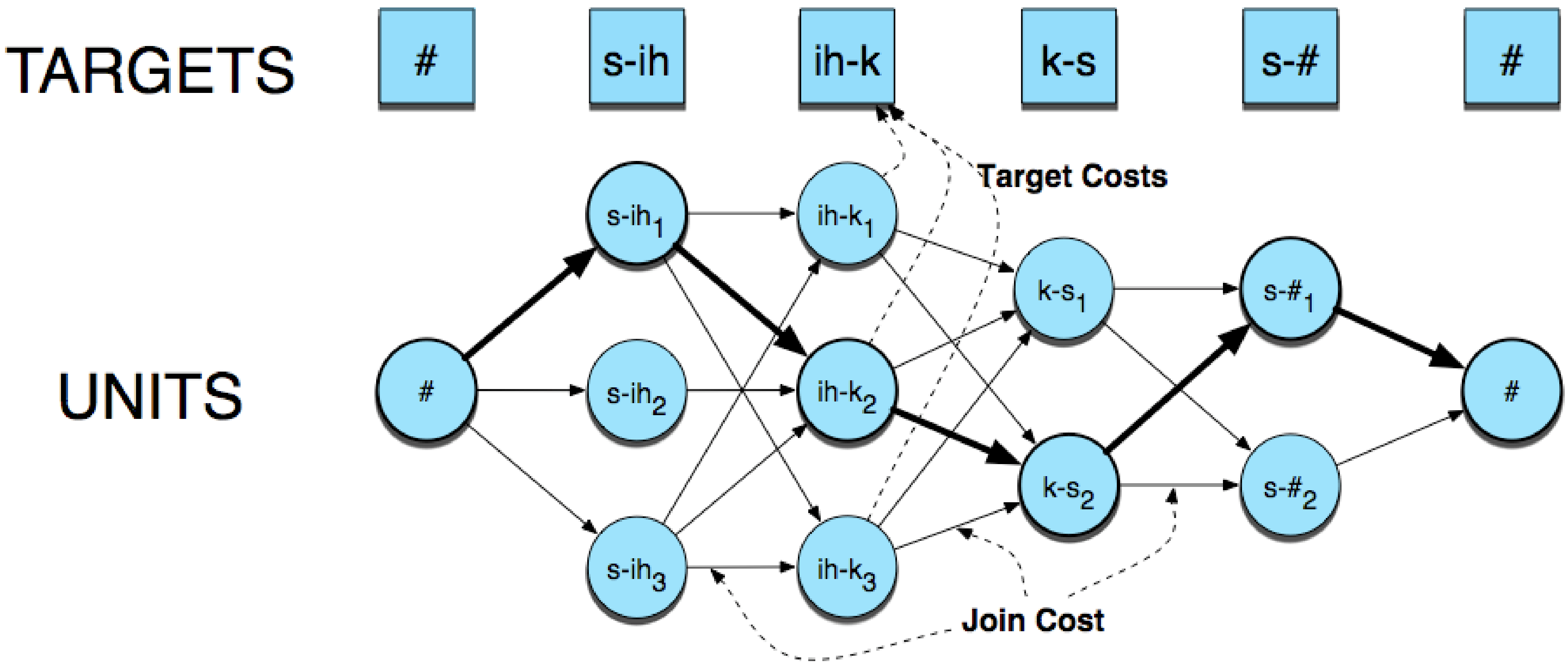
Total Costs

- 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


TARGETS
#
s-ih
ih-k
k-s
s-#
#
Target Costs
UNITS
#
s-ih1
s-ih2
s-ih3
ih-k1
ih-k2
ih-k3
k-s1
k-s2
s-#1
s-#2
#
Join Cost

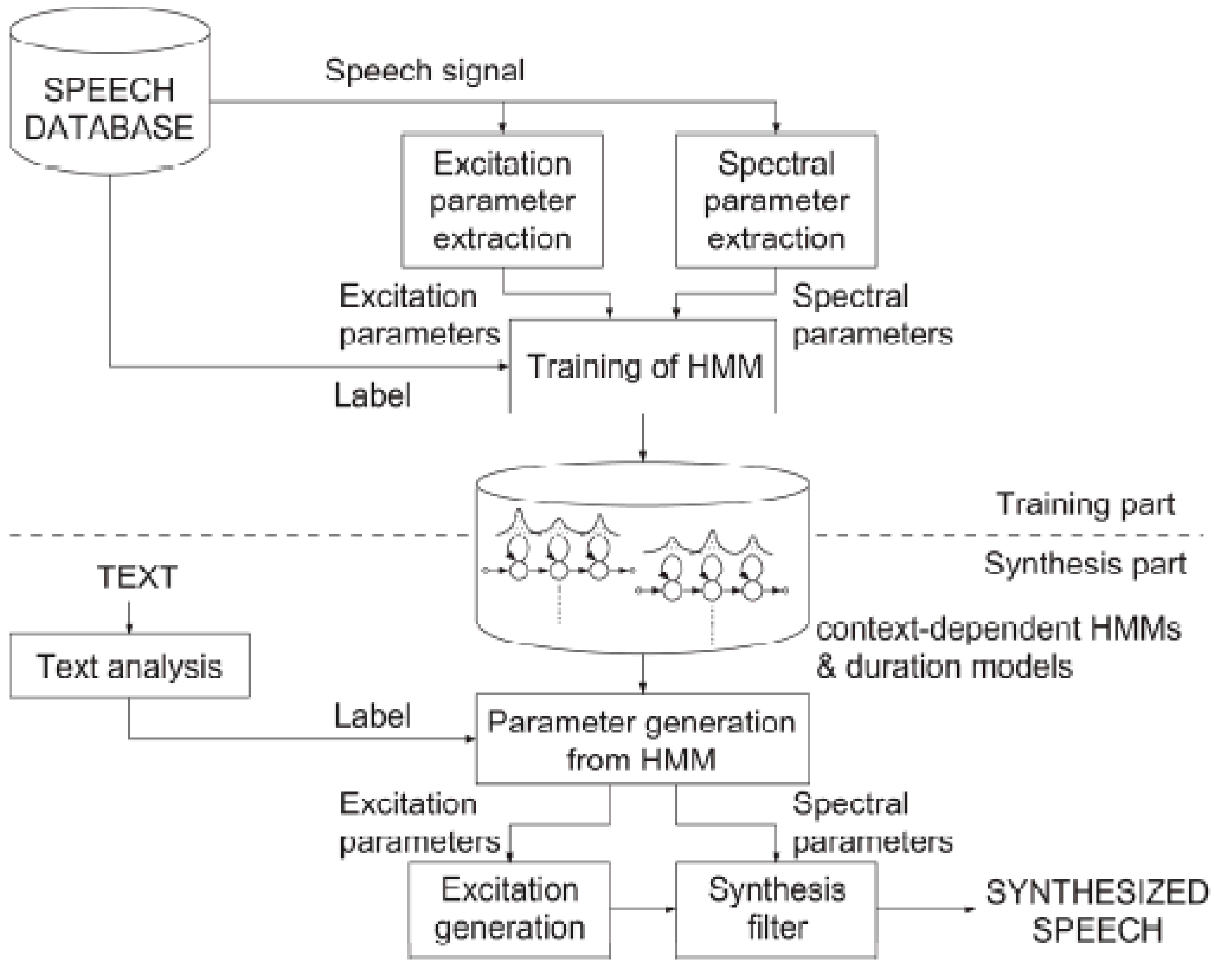
Unit Selection Summary

- Advantages
  - Quality is far superior to diphones: fewer joins, more choices
  - Natural prosody selection sounds better
- Disadvantages:
  - Quality can be very bad when no good match in database
    - HCI problem: mix of very good and very bad is quite annoying
  - Synthesis is computationally expensive
  - Can't synthesize everything you want. e.g.
    - Diphone technique can vary emphasis
    - Unit selection can give result that conveys wrong meaning

New Trend

- Problems with Unit Selection Synthesis
  - Can’t modify signal
    - Mixing modified and unmodified sounds unpleasant
    - But database often doesn’t have exactly what you want
- Solution: HMM (Hidden Markov Model) Synthesis
  - Won recent TTS bakeoff
    - Sounds less natural to researchers but naïve subjects preferred it
    - Has the potential to improve over both diphone and unit selection

Tokuda et al ’02



HMM Synthesis

- Unit selection (Roger)
- HMM (Roger)

- Unit selection (Nina)
- HMM (Nina)

TTS Evaluation

- Intelligibility Tests
- Mean Opinion Scores
- Preference Tests

Intelligibility Tests

- Diagnostic Rhyme Test (DRT)
  - Listening test
  - Listeners choose between two words differing by a single phonetic feature (voicing, nasality, sustenation, sibilation)
  - DRT: 96 rhyming pairs
    - Dense/tense, bond/pond, …
      - Subject hears dense, chooses either dense or tense
      - % of correct answers is intelligibility score
  - Problem: Only tests single word synthesis

- Modified DRT:
  - 300 words, 50 sets of 6 words (went, sent, bent, tent, dent, rent)
  - Embedded in carrier phrases:
    - Now we will say dense again
- Mean Opinion Score
  - Have listeners rate output on a scale from 1 (bad) to 5 (excellent)
- Preference tests:
  - Reading addresses out loud, reading news text, using two different systems or systems against human voice
  - Do a **preference test** (prefer A, prefer B)

Next Class

- Speech Recognition Overview
- HW 4 due: Can you come up with ways to evaluate TTS systems better?
- Happy Spring Break