

CS 224S / LINGUIST 281 Speech Recognition and Synthesis

Dan Jurafsky

Lecture 6: Waveform Synthesis in Concatenative TTS

IP Notice: many of these slides come directly from Richard Sproat's slides, and other (and some of Richard's) come from Alan Black's excellent TTS lecture notes. A couple also from Paul Taylor

1/29/06

CS 224S Winter 2006

1

Goal of Today's Lecture

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

1/29/06

CS 224S Winter 2006

2

Outline: Waveform Synthesis in Concatenative TTS

- **Diphone Synthesis**
- **Break: Final Projects**
- **Unit Selection Synthesis**
 - Target cost
 - Unit cost
- **Joining**
 - Dumb
 - PSOLA

1/29/06

CS 224S Winter 2006

3

Diphone TTS architecture

- **Training:**
 - Choose units (kinds of diphones)
 - Record diphones
 - Label diphones (decide where break is)
- **Synthesizing an utterance,**
 - grab relevant diphones from database,
 - use signal processing to change the prosody (F0, energy, duration) of selected sequence of diphones

1/29/06

CS 224S Winter 2006

4

Diphones

- mid-phone is more stable than edge
- Need $O(\text{phone}^2)$ number of units
 - Some combinations don't exist (hopefully)
 - May include stress, consonant clusters
 - Lots of phonetic knowledge in design
- Database relatively small (by today's standards)
 - Around 8 megabytes for English (16 KHz 16 bit)

1/29/06

CS 224S Winter 2006

Slide from Richard Sproat

Designing a diphone inventory: Nonsense words

- **Build set of carrier words:**
 - pau t aa b aa b aa pau
 - pau t aa m aa m aa pau
 - pau t aa m iy m aa pau
 - pau t aa m iy m aa pau
 - pau t aa m ih m aa pau
- **Advantages:**
 - Easy to get all diphones
 - Likely to be pronounced consistently
 - No lexical interference
- **Disadvantages:**
 - (possibly) bigger database
 - Speaker becomes bored

1/29/06

CS 224S Winter 2006

Slide from Richard Sproat

Designing a diphone inventory: Natural words

- **Greedily select sentences/words:**
 - Quebecois arguments
 - Brouhaha abstractions
 - Arkansas arranging
- **Advantages:**
 - Will be pronounced naturally
 - Easier for speaker to pronounce
 - Smaller database? (505 pairs vs. 1345 words)
- **Disadvantages:**
 - May not be pronounced correctly

1/29/06

CS 224S Winter 2006

7

Slide from Richard Sproat

Making recordings consistent:

- **Diiphone should come from mid-word**
 - Help ensure full articulation
- **Performed consistently**
 - Constant pitch (monotone), power, duration
- **Use (synthesized) prompts:**
 - Helps avoid pronunciation problems
 - Keeps speaker consistent
 - Used for alignment in labeling

1/29/06

CS 224S Winter 2006

8

Slide from Richard Sproat

Building diphone schemata

- **Find list of phones in language:**
 - Plus interesting allophones
 - Stress, tons, clusters, onset/coda, etc
 - Foreign (rare) phones.
- **Build carriers for:**
 - Consonant-vowel, vowel-consonant
 - Vowel-vowel, consonant-consonant
 - Silence-phone, phone-silence
 - Other special cases
- **Check the output:**
 - List *all* diphones and justify missing ones
 - Every *diphone* list has mistakes

1/29/06

CS 224S Winter 2006

9

Slide from Richard Sproat

Recording conditions

- **Ideal:**
 - Anechoic chamber
 - Studio quality recording
 - EGG signal
- **More likely:**
 - Quiet room
 - Cheap microphone/sound blaster
 - No EGG
 - Headmounted microphone
- **What we can do:**
 - Repeatable conditions
 - Careful setting on audio levels

1/29/06

CS 224S Winter 2006

10

Slide from Richard Sproat

Labeling Diphones

- **Much easier than phonetic labeling:**
 - The phone sequence is defined
 - They are clearly articulated
 - But sometimes speaker still pronounces wrong, so need to check.
- **Phone boundaries less important**
 - +- 10 ms is okay
- **Midphone boundaries important**
 - Where is the stable part
 - Can it be automatically found?

1/29/06

CS 224S Winter 2006

11

Slide from Richard Sproat

Diphone auto-alignment

- **Given**
 - synthesized prompts
 - Human speech of same prompts
- **Do a dynamic time warping alignment of the two**
 - Using euclidean distance
- **Works very well 95%+**
 - Errors are typically large (easy to fix)
 - Maybe even automatically detected
- **Malfrere and Dutoit (1997)**

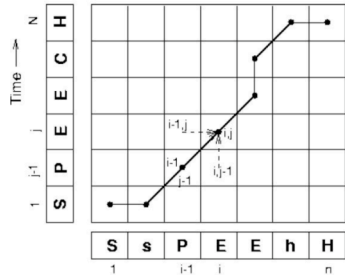
1/29/06

CS 224S Winter 2006

12

Slide from Richard Sproat

Dynamic Time Warping



1/29/06

CS 224S Winter 2006

13

Slide from Richard Sproat

Finding diphone boundaries

- **Stable part in phones**
 - For stops: one third in
 - For phone-silence: one quarter in
 - For other diphones: 50% in
- **In time alignment case:**
 - Given explicit known diphone boundaries in prompt in the label file
 - Use dynamic time warping to find same stable point in new speech
- **Optimal coupling**
 - Conkie and Isard 1996
 - Find optimal join points by measure cepstral distance at potential join points, pick best

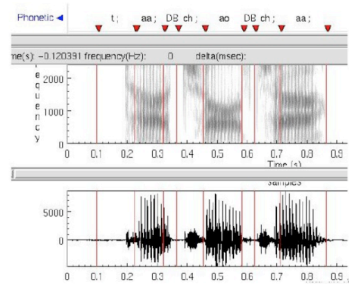
1/29/06

CS 224S Winter 2006

14

Slide from Richard Sproat

Diphone boundaries in stops

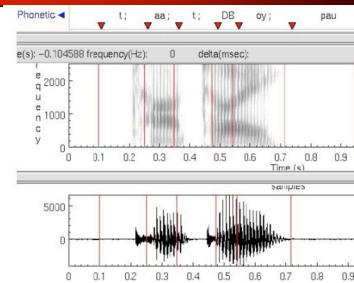


1/29/06

Slide from Richard Sproat

15

Diphone boundaries in end phones



1/29/06

CS 224S Winter 2006

16

Slide from Richard Sproat

Summary: Diphone Synthesis

- **Well-understood, mature technology**
- **Augmentations**
 - Stress
 - Onset/coda
 - Demi-syllables
- **Problems:**
 - Signal processing still necessary for modifying durations
 - Source data is still not natural
 - Units are just not large enough; can't handle word-specific effects, etc

1/29/06

CS 224S Winter 2006

17

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)

1/29/06

CS 224S Winter 2006

18

Why Unit Selection Synthesis

- Natural data solves problems with diphones
 - Diphone databases are carefully designed but:
 - Speaker makes errors
 - Speaker doesn't speak intended dialect
 - Require database design to be right
 - If it's automatic
 - Labeled with what the speaker actually said
 - Coarticulation, schwas, flaps are natural
- "There's no data like mo' data"
 - Lots of copies of each unit mean you can choose just the right one for the context
 - Larger units mean you can capture wider effects

1/29/06

CS 224S Winter 2006

19

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
 - FO, stress, phrase position
 - "Join cost": Best join with neighboring units
 - Matching formants + other spectral characteristics
 - Matching energy
 - Matching FO

1/29/06

CS 224S Winter 2006

20

Targets and Target Costs

- A measure of how well a particular unit in the database matches the internal representation produced by the prior stages
- Features, costs, and weights
- Examples:
 - /ih-t/ from stressed syllable, phrase internal, high FO, content word
 - /n-t/ from unstressed syllable, phrase final, low FO, content word
 - /dh-ax/ from unstressed syllable, phrase initial, high FO, from function word "the"

1/29/06

CS 224S Winter 2006

21

Slide from Paul Taylor

Target Costs

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

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

1/29/06

CS 224S Winter 2006

22

Slide from Paul Taylor

How to set target cost weights (1)

- What you REALLY want as a target cost is the perceivable acoustic difference between two units
- But we can't use this, since the target is NOT ACOUSTIC yet, we haven't synthesized it!
- We have to use features that we get from the TTS upper levels (phones, prosody)
- But we DO have lots of acoustic units in the database.
- We could use the acoustic distance between these to help set the WEIGHTS on the acoustic features.

1/29/06

CS 224S Winter 2006

23

How to set target cost weights (2)

- Clever Hunt and Black (1996) idea:
- Hold out some utterances from the database
- Now synthesize one of these utterances
 - Compute all the phonetic, prosodic, duration features
 - Now for a given unit in the output
 - For each possible unit that we COULD have used in its place
 - We can compute its acoustic distance from the TRUE ACTUAL HUMAN utterance.
 - This acoustic distance can tell us how to weight the phonetic/prosodic/duration features

1/29/06

CS 224S Winter 2006

24

How to set target cost weights (3)

- Hunt and Black (1996)
- Database and target units labeled with:
 - phone context, prosodic context, etc.
- Need an acoustic similarity between units too
- Acoustic similarity based on perceptual features
 - MFCC (spectral features)
 - FO (normalized)
 - Duration penalty

$$AC^i(t_i, u_i) = \sum_{n=1}^p w_n^i \text{abs}(P_i(u_n) - P_i(u_m))$$

1/29/06

CS 224S Winter 2006

Richard Sproat slide 25

How to set target cost weights (3)

- Collect phones in classes of acceptable size
 - E.g., stops, nasals, vowel classes, etc
- Find AC between all of same phone type
- Find C^t between all of same phone type
- Estimate w_{1-j} using linear regression

1/29/06

CS 224S Winter 2006

26

How to set target cost weights (4)

- Target distance is

$$C^i(t_i, u_i) = \sum_{k=1}^p w_k^i C_k^i(t_i, u_i)$$
- For examples in the database, we can measure

$$AC^i(t_i, u_i) = \sum_{n=1}^p w_n^i \text{abs}(P_i(u_n) - P_i(u_m))$$
- Therefore, estimate weights w from all examples of

$$AC^i(t_i, u_i) \approx \sum_{n=1}^p w_n^i C_k^i(t_i, u_i)$$
- Use linear regression

1/29/06

CS 224S Winter 2006

Richard Sproat slide 27

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
 - FO
 - Energy
- Join cost:

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

1/29/06

CS 224S Winter 2006

28

Slide from Paul Taylor

Join costs

- Hunt and Black 1996
- If $u_{i-1} = \text{prev}(u_i)$ C^j=0
- Used
 - MFCC (mel cepstral features)
 - Local FO
 - Local absolute power
 - Hand tuned weights

1/29/06

CS 224S Winter 2006

29

Join costs

- The join cost can be used for more than just part of search
- Can use the join cost for *optimal coupling* (Conkie 1996), i.e., finding the best place to join the two units.
 - Vary edges within a small amount to find best place for join
 - This allows different joins with different units
 - Thus labeling of database (or diphones) need not be so accurate

1/29/06

CS 224S Winter 2006

30

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:

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

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

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

1/29/06

CS 224S Winter 2006

31

Slide from Paul Taylor

Improvements

- Taylor and Black 1999: Phonological Structure Matching
- Label whole database as trees:
 - Words/phrases, syllables, phones
- For target utterance:
 - Label it as tree
 - Top-down, find subtrees that cover target
 - Recurse if no subtree found
- Produces list of target subtrees:
 - Explicitly longer units than other techniques
- Selects on:
 - Phonetic/metrical structure
 - Only indirectly on prosody
 - No acoustic cost

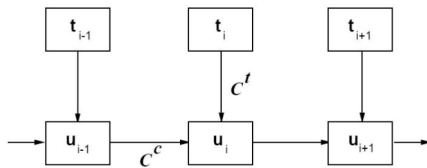
1/29/06

CS 224S Winter 2006

32

Slide from Richard Sproat

Unit Selection Search



1/29/06

CS 224S Winter 2006

33

Slide from Richard Sproat

Database creation (1)

- **Good speaker**
 - Professional speakers are always better:
 - Consistent style and articulation
 - Although these databases are carefully labeled
 - Ideally (according to AT&T experiments):
 - Record 20 professional speakers (small amounts of data)
 - Build simple synthesis examples
 - Get many (200?) people to listen and score them
 - Take best voices
 - Correlates for human preferences:
 - High power in unvoiced speech
 - High power in higher frequencies
 - Larger pitch range

1/29/06

CS 224S Winter 2006

34

Text from Paul Taylor and Richard Sproat

Database creation (2)

- **Good recording conditions**
- **Good script**
 - Application dependent helps
 - Good word coverage
 - News data synthesizes as news data
 - News data is bad for dialog.
 - Good phonetic coverage, especially wrt context
 - Low ambiguity
 - Easy to read
- **Annotate at phone level, with stress, word information, phrase breaks**

1/29/06

CS 224S Winter 2006

35

Text from Paul Taylor and Richard Sproat

Creating database

- **Unliked diphones, prosodic variation is a good thing**
- **Accurate annotation is crucial**
- **Pitch annotation needs to be very very accurate**
- **Phone alignments can be done automatically, as described for diphones**

1/29/06

CS 224S Winter 2006

36

Practical System Issues

- Size of typical system (Rhetorical rVoice):
 - ~300M
- Speed:
 - For each diphone, average of 1000 units to choose from, so:
 - 1000 target costs
 - 1000x1000 join costs
 - Each join cost, say 30x30 float point calculations
 - 10-15 diphones per second
 - 10 billion floating point calculations per second
- But commercial systems must run ~50x faster than real time
- Heavy pruning essential: 1000 units -> 25 units

1/29/06

CS 224S Winter 2006

37

Slide from Paul Taylor

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

1/29/06

CS 224S Winter 2006

38

Slide from Richard Sproat

Joining Units (+FO + duration)

- Both diphone and unit selection synthesis need to join the units
- For diphone synthesis, need to modify FO and duration
- For unit selection, in principle also need to modify FO and duration of selection units
- But in practice, if unit-selection database is big enough (commercial systems) often avoid prosodic modifications altogether, as selected targets may already be close to desired prosody.

1/29/06

CS 224S Winter 2006

39

Alan Black

Joining Units

- **Dumb:**
 - just join
 - Better: at zero crossings
- **TD-PSOLA**
 - Time-domain pitch-synchronous overlap-and-add
 - Join at pitch periods (with windowing)

1/29/06

CS 224S Winter 2006

40

Alan Black

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

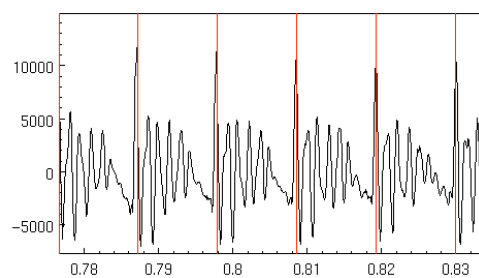
1/29/06

CS 224S Winter 2006

41

Text from Alan Black

Speech as Short Term signals



1/29/06

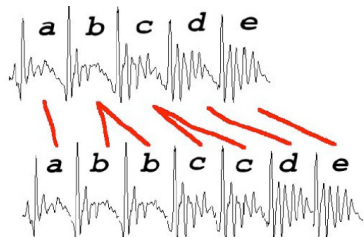
CS 224S Winter 2006

42

Alan Black

Duration modification

- Duplicate/remove short term signals



1/29/06

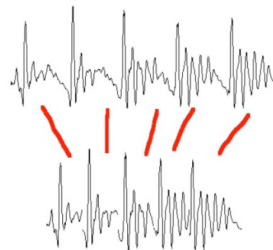
CS 224S Winter 2006

43

Slide from Richard Sproat

Pitch Modification

- Move short-term signals closer together/further apart

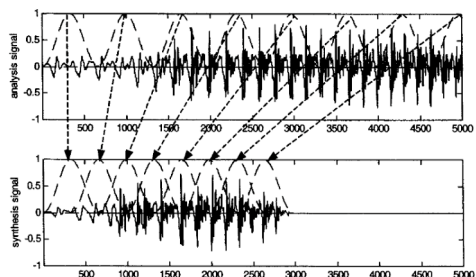


1/29/06

44

Slide from Richard Sproat

Overlap-and-add (OLA)



1/29/06

CS 224S Winter 2006

45

Huang, Acero and Hon

Overlap and Add (OLA)

- Hanning windows of length $2N$ used to multiply the analysis signal
- Resulting windowed signals are added
- Analysis windows, spaced $2N$
- Synthesis windows, spaced N
- Time compression is uniform with factor of 2
- Pitch periodicity somewhat lost around 4th window

1/29/06

CS 224S Winter 2006

46

Huang, Acero, and Hon

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

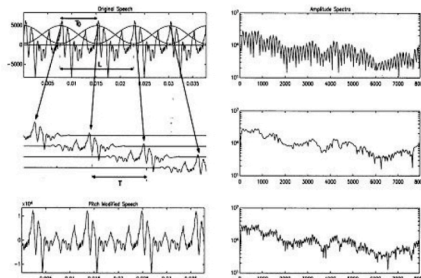
1/29/06

CS 224S Winter 2006

47

Slide from Richard Sproat

TD-PSOLA™



1/29/06

CS 224S Winter 2006

48

Thierry Dutoit

Evaluation of TTS

- **Intelligibility Tests**
 - **Diagnostic Rhyme Test (DRT)**
 - Humans do listening identification choice between two words differing by a single phonetic feature
 - Voicing, nasality, sustenation, sibilation
 - 96 rhyming pairs
 - Veal/feel, meat/beat, vee/bee, zee/thee, etc
 - Subject hears "veal", chooses either "veal" or "feel"
 - Subject also hears "feel", chooses either "veal" or "feel"
 - % of right answers is intelligibility score.
 - **Overall Quality Tests**
 - Have listeners rate space on a scale from 1 (bad) to 5 (excellent)
 - **Preference Tests (prefer A, prefer B)**

1/29/06

CS 224S Winter 2006

49

Huang, Acero, Hon

Summary

- **Diphone Synthesis**
- **Unit Selection Synthesis**
 - Target cost
 - Unit cost
- **Joining**
 - Dumb
 - PSOLA

1/29/06

CS 224S Winter 2006

50