Computational Linguistics 1

CMSC/LING 723, LBSC 744

Kristy Hollingshead Seitz
Institute for Advanced Computer Studies
University of Maryland

Lecture 24: 1 December 2011

Agenda

- HW5 graded
- HW6 due next Tuesday
- Schedule changes
- IGERT
- Winter Storm
- Questions, comments, concerns?
- Text-to-Speech (TTS)

Speech Synthesis/Text-to-Speech (TTS)

- IP notice
  - The following slides are from Dan Jurafsky, Richard Sproat, and other researchers as noted on the slides
  - As presented in the Speech Synthesis lectures at the LSA Summer Institute

TTS: Outline

- From words to strings of phones
  - Dictionaries
  - Letter-to-Sound Rules
    - ("Grapheme-to-Phoneme Conversion")
- Prosody
  - Linguistic Background
  - Producing Intonation in TTS
  - Stress/accent
- TTS Systems
  - Diphone synthesis
  - Unit selection synthesis

From words to phones

- Two methods:
  - Dictionary-based
  - Rule-based (Letter-to-sound=LTS, grapheme-to-phoneme = G2P)
- Early systems, all LTS
- MITalk was radical in having 'huge' 10K word dictionary
- Modern systems use a combination

Pronunciation Dictionaries: CMU

- CMU dictionary: 127k words
  - http://www.speech.cs.cmu.edu/cgi-bin/cmudict
- Some problems:
  - Has errors
  - Only American pronunciations
  - No syllable boundaries
  - Doesn't tell us which pronunciation to use for which homophones
    - (no POS tags)
  - Doesn't distinguish case
    - The word US has 2 pronunciations
      - [AH1 S] and [Y UW1 EH1 S]

Dictionaries aren't sufficient

- Unknown words (OOVs)
  - Increase with the (sqrt of) number of words in unseen text
  - Black et al (1998) OALD on 1st section of Penn Treebank:
  - Out of 39923 word tokens,
    - 1775 tokens were OOV: 4.6% (943 unique types):

names	unknown	Typos/other
1360	351	64
76.6%	19.8%	3.6%

- So commercial systems have 4-part system:
  - Big **dictionary**
  - **Names** handled by special routines
  - **Acronyms** handled by special routines (previous lecture)
  - Machine learned **g2p** algorithm for other unknown words

Names

- Big problem area is names
- Names are common
  - 20% of tokens in typical newswire text will be names
  - 1987 Donnelly list (72 million households) contains about 1.5 million names
  - Personal names: McArthur, D'Angelo, Jiminez, Rajan, Raghavan, Sondhi, Xu, Hsu, Zhang, Chang, Nguyen
  - Company/Brand names: Infinit, Kmart, Cytyc, Medamicus, Inforte, Aaon, Idexx Labs, Bebe

Names

- Methods:
  - Can do morphology (Walters -> Walter, Lucasville)
  - Can write stress-shifting rules (Jordan -> Jordanian)
  - Rhyme analogy: Plotsky by analogy with Trostsky (replace tr with pl)
  - Liberman and Church:
    - for 250K most common names, got 212K (85%) from these modified-dictionary methods, used LTS for rest.
  - Can do automatic country detection (from letter trigrams) and then do country-specific rules
  - Can train **g2p** system specifically on names
    - Or specifically on types of names (brand names, Russian names, etc)

Acronyms

- We saw in the text normalization lecture
- Use machine learning to detect acronyms
  - EXPN
  - ASWORD
  - LETTERS
- Use acronym dictionary, hand-written rules to augment

Letter-to-Sound Rules

- Earliest algorithms: handwritten Chomsky+Halle-style rules:

$$c \rightarrow [k] \; / \; \_\_ \; \{a,o\}V \qquad \text{; context-dependent}$$
$$c \rightarrow [s] \qquad \text{; context-independent}$$

- Rules apply in order
  - "christmas" pronounced with [k]
  - But word with ch followed by non-consonant pronounced [ch]
    - e.g., "choice"
- English famously evil
  - in terms of pronunciation and stress rules

Modern method: Learning LTS rules automatically

- Induce LTS from a dictionary of the language
- Black et al. 1998
- Applied to English, German, French
- Two steps:
  - **alignment**
  - (CART-based) **rule-induction**

Alignment

- Letters: c h e c k e d
- Phones: ch _ eh _ k _ t
- Black et al Method 1:
  - First scatter epsilons in all possible ways to cause letters and phones to align
  - Then collect stats for P(phone|letter) and select best to generate new stats

$$p(p_i|l_j) = \frac{\text{count}(p_i, l_j)}{\text{count}(l_j)}$$

  - This iterated a number of times until settles (5-6)
  - This is EM (expectation maximization) alg

L: c a k e
| | | |
P: K EY K ε

Hand-specified letters-to-phones

- Hand specify which letters can be rendered as which phones
  - C goes to k/ch/s/sh
  - W goes to w/v/f, etc
- Once mapping table is created, find all valid alignments, find p(letter|phone), score all alignments, take best

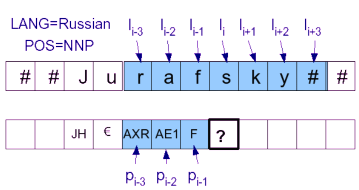
Alignment

- Some alignments will turn out to be really bad.
- These are just the cases where pronunciation doesn't match letters:
  - Dept d ih p aa r t m ah n t
  - CMU s iy eh m y uw
  - Lieutenant l eh f t eh n ax n t (British)
- Also foreign words
- These can just be removed from alignment training

Building CART trees

- Build a CART tree for each letter in alphabet (26 plus accented) using context of +-3 letters
- # # # c h e c -> ch
- c h e c k e d -> _

Add more features



- Even more: for French liaison, we need to know what the next word is, and whether it starts with a vowel
- French six
  - [s iy s] in *j'en veux six*
  - [s iy z] in *six enfants*
  - [s iy] in *six filles*

TTS: Outline

- From words to strings of phones
  - Dictionaries
  - Letter-to-Sound Rules
    - ("Grapheme-to-Phoneme Conversion")
- Prosody
  - Linguistic Background
  - Producing Intonation in TTS
  - Stress/accent
- TTS Systems
  - Diphone synthesis
  - Unit selection synthesis

Defining Intonation

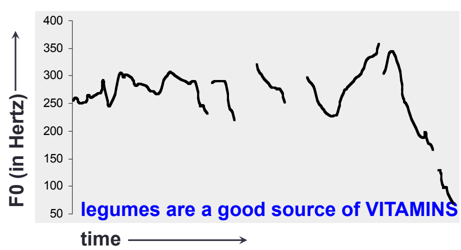
- Ladd (1996) "Intonational Phonology"
- "The use of suprasegmental phonetic features...
  Suprasegmental = above and beyond the segment/phone
  - F0
  - Intensity (energy)
  - Duration
- ...to convey sentence-level pragmatic meanings"
  - i.e., meanings that apply to phrases or utterances as a whole, not lexical stress, not lexical tone.

Three aspects of prosody

- Prominence: some syllables/words are more prominent than others
- Structure/boundaries: sentences have prosodic structure
  - Some words group naturally together
  - Others have a noticeable break or disjuncture between them
- Tune: the intonational melody of an utterance.

From Ladd (1996)

Graphic representation of F0



(F0 is not defined for consonants without vocal fold vibration.)

Prominence: Stress vs. Accent

- ***Prominence*** is the placement of pitch accents
- ***Stress* is a structural property of a word — it marks a potential (arbitrary) location for an accent to occur, if there is one.**
- ***Accent* is a property of a word in context — it is a way to mark intonational prominence in order to 'highlight' important words in the discourse.**

(x)					(x)			(accented syll)
x					x			stressed syll
x			x		x			full vowels
x	x	x	x	x	x	x		syllables
vi	ta	mins	Ca	li	for	nia		

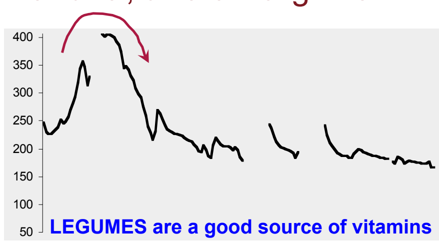
Stress vs. Accent

- The speaker decides to make the word vitamin more prominent by accenting it.
- Lexical stress tell us that this prominence will appear on the first syllable, hence VItamin.
- So we will have to look at both the lexicon and the context to predict the details of prominence
- I'm a little **surPRISED** to hear it **CHARacterized** as **upBEAT**

Which word receives an accent?

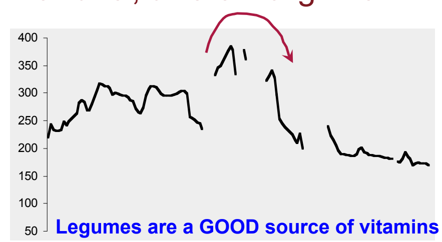
- **It depends on the context.**
- **For example, the 'new' information in the answer to a question is often accented, while the 'old' information usually is not.**
  - Q1: What types of foods are a good source of vitamins?
  - A1: LEGUMES are a good source of vitamins.
  - Q2: Are legumes a source of vitamins?
  - A2: Legumes are a GOOD source of vitamins.
  - Q3: I've heard that legumes are healthy, but what are they a good source of ?
  - A3: Legumes are a good source of VITAMINS.

Same 'tune', different alignment



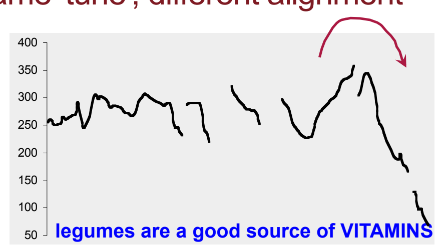
The main rise-fall accent (= "I assert this") shifts locations.

Same 'tune', different alignment



The main rise-fall accent (= "I assert this") shifts locations.

Same 'tune', different alignment



The main rise-fall accent (= "I assert this") shifts locations.

Levels of Prominence

- Most phrases have more than one accent
- The last accent in a phrase is perceived as more prominent
  - Called the Nuclear Accent
- **Emphatic** accents like nuclear accent often used for semantic purposes, such as indicating that a word is contrastive, or the semantic focus.
  - The kind of thing you represent via ***s in IM, or capitalized letters
  - "I know **SOMETHING** interesting is sure to happen," she said to herself.
- Can also have words that are **less** prominent than usual
  - Reduced words, especially function words.
- Often use 4 classes of prominence:
  1. **emphatic accent,**
  2. **pitch accent,**
  3. **unaccented,**
  4. **reduced**

Three Aspects of Prosody

- Prominence: some syllables/words are more prominent than others
- Structure/boundaries: sentences have prosodic structure
  - Some words group naturally together
  - Others have a noticeable break or disjuncture between them
- Tune: the intonational melody of an utterance.

From Ladd (1996)

Intonational Phrasing/Boundaries

- A single intonation phrase
  - Broad focus statement consisting of one intonation phrase (that is, one intonation tune spans the whole unit).
  - "Legumes are a good source of vitamins."
- Multiple phrases
  - Utterances can be 'chunked' up into smaller phrases in order to signal the importance of information in each unit.
  - "Legumes are a good source of vitamins"
- Disambiguation

Phrasing sometimes helps disambiguate

- **Temporary ambiguity:**

When Madonna sings the song ...

Phrasing sometimes helps disambiguate

- **Temporary ambiguity:**

When Madonna sings the song is a hit.

Phrasing sometimes helps disambiguate

- **Temporary ambiguity:**

When Madonna sings % the song is a hit.

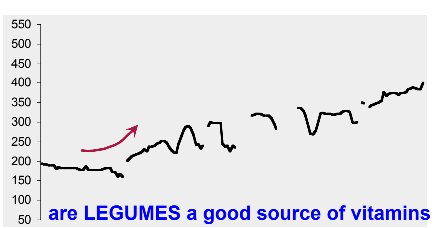
When Madonna sings the song % it's a hit.

[from Speer & Kjelgaard (1992)]

Intonational Tunes

- Yes-No question tune
- WH-questions
- Rising statements
- 'Surprise-redundancy' tune
- 'Contradiction' tune

Yes-No question tune

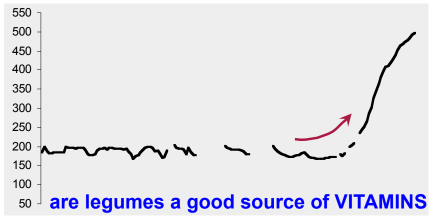


Rise from the main accent to the end of the sentence.

Yes-No question tune

Rise from the main accent to the end of the sentence.

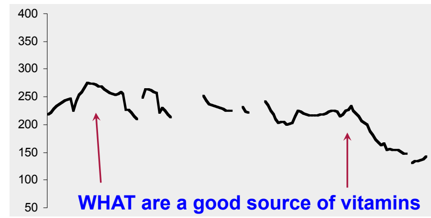
Yes-No question tune



Rise from the main accent to the end of the sentence.

WH-questions

[I know that many natural foods are healthy, but ...]



WH-questions typically have **falling** contours, like statements.

Broad focus

"Tell me something about the world."

In the absence of narrow focus, English tends to mark the **first** and **last** 'content' words with perceptually prominent accents.

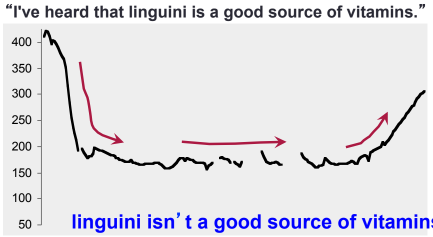
'Surprise-redundancy' tune

[How many times do I have to tell you ...]

Low beginning followed by a gradual rise to a **high** at the end.

'Contradiction' tune

"I've heard that linguini is a good source of vitamins."



[... how could you think that?]

Sharp fall at the beginning, **flat and low**, then **rising** at the end.

Using Intonation in TTS

1) Prominence/Accent: Decide which words are accented, which syllable has accent, what sort of accent
2) Boundaries: Decide where intonational boundaries are
3) Duration: Specify length of each segment
4) F0: Generate F0 contour from these

Predicting Pitch Accent: Factors

- Part of speech
  - Content words are usually accented
  - Function words are rarely accented
    - Of, for, in on, that, the, a, an, no, to, and but or will may would can her is their its our there is am are was were, etc.
- But it's not just function/content
  - Contrast
    - Legumes are poor source of VITAMINS
      No, legumes are a GOOD source of vitamins
    - I think JOHN or MARY should go
      No, I think JOHN AND MARY should go
  - List intonation
  - Information status
  - Syntactic structure

Predicting Pitch Accent: Other Features

- POS
- POS of previous word
- POS of next word
- Stress of current, previous, next syllable
- Unigram probability of word
- Bigram probability of word
- Position of word in sentence

Predicting Pitch Accent: State-of-the-art

- Hand-label large training sets
- Use CART, SVM, CRF, etc to predict accent
- Lots of rich features from context
- Classic lit:
  - Hirschberg, Julia. 1993. Pitch Accent in context: predicting intonational prominence from text. Artificial Intelligence 63, 305-340

Predicting Boundaries: Features

- Intonation phrase boundaries
  - Intermediate phrase boundaries
  - Full intonation phrase boundaries
- Based just on punctuation and clauses?

Police also say | Levy's blood alcohol level | was twice the legal limit ||

Predicting Boundaries: More Features

- Length features:
  - Phrases tend to be of roughly equal length
  - Total number of words and syllables in utterance
  - Distance of juncture from beginning and end of sentence (in words or syllables)
- Neighboring POS, punctuation
- Syntactic structure (parse trees)
  - Largest syntactic category dominating preceding word but not succeeding word
  - How many syntactic units begin/end between words
- Other:
  - English: boundaries are more likely between content words and function words
  - Type of function word to right
  - Capitalized names
  - # of content words since previous function word

TTS Intonation Prediction

- Predict duration
- Predict F0

TTS: Outline

- From words to strings of phones
  - Dictionaries
  - Letter-to-Sound Rules
    - ("Grapheme-to-Phoneme Conversion")
- Prosody
  - Linguistic Background
  - Producing Intonation in TTS
  - Stress/accent
- TTS Systems
  - Diphone synthesis
  - Unit selection synthesis

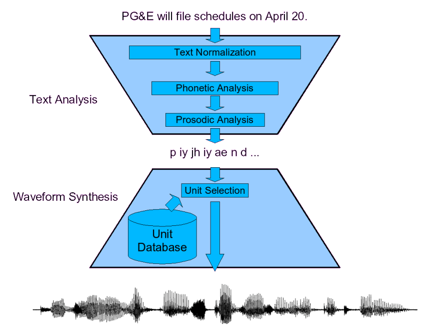
Goal of Speech Synthesis Systems

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

Waveform Synthesis in Concatenative TTS

- Diphone Synthesis
- Unit Selection Synthesis
  - Target cost
  - Unit cost

TTS Architecture



Diphone TTS Architecture

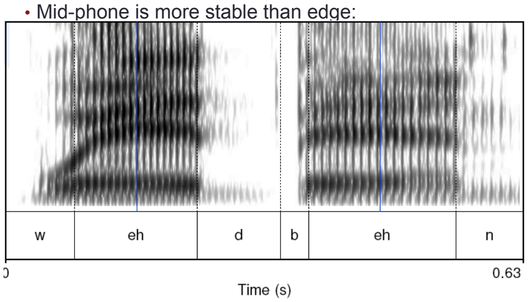
- Training:
  - Choose units (kinds of diphones)
  - Record 1 speaker saying 1 example of each diphone
  - Mark the boundaries of each diphones,
    - cut each diphone out and create a diphone database
- Synthesizing an utterance:
  - Grab relevant sequence of diphones from database
  - Concatenate the 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
- 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)

Diphones

- Mid-phone is more stable than edge:



Voice

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

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

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

Labeling Diphones

- Run a speech recognizer in **forced** alignment mode
  - Forced alignment:
    - A trained ASR system
    - A wavefile
    - A word transcription of the wavefile
    - Returns an alignment of the phones in the words to the wavefile.
- *Much* easier than phonetic labeling:
  - The words are defined
  - The phone sequence is generally 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?

Concatenating diphones: junctures

- If waveforms are very different, will perceive a click at the junctures
  - So need to window them
- Also if both diphones are voiced
  - Need to join them **pitch-synchronously**
- That means we need to know where each pitch period begins, so we can paste at the same place in each pitch period.
  - **Pitch marking** or **epoch detection**: mark where each **pitch pulse** or **epoch** occurs
    - Finding the Instant of Glottal Closure (IGC)
  - (note difference from **pitch tracking**)

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

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.

Problems with Diphone Synthesis

- Signal processing methods leave artifacts, making the speech sound unnatural
- Diphone synthesis only captures local effects
  - But there are many more global effects (syllable structure, stress pattern, word-level effects)

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)
  - Little or no signal processing applied to each unit
    - Unlike diphones

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 more 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

Unit Selection Intuition

- Given a big database
- For each segment (diphone) that we want to synthesize
  - Find the unit in the database that is the *best* to synthesize this 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

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

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 F0, content word
  - /n-t/ from unstressed syllable, phrase final, low F0, content word
  - /dh-ax/ from unstressed syllable, phrase initial, high F0, from function word "the"

Target Costs

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

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

How to set target cost weights

- 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.

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:

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

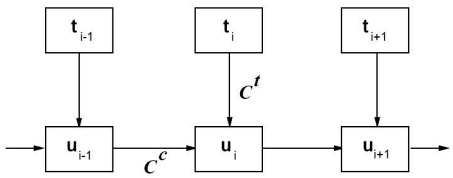
Join costs

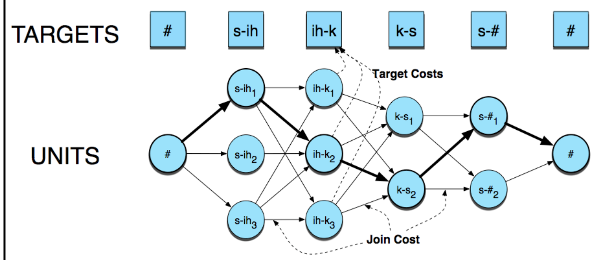
- Hunt and Black 1996
- If u_{i-1}==prev(u_i) C^c=0
- Used
  - MFCC (mel cepstral features)
  - Local F0
  - Local absolute power
  - Hand tuned weights

Join costs

- The join cost can be used for more than just part of search
- Can use the join cost for *optimal coupling* (Isard and Taylor 1991, 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

Unit Selection Search





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

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

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

(Relatively) Recent Advances

- Problems with Unit Selection Synthesis
  - Can't modify signal (mixing modified and unmodified sounds bad)
  - But database often doesn't have exactly what you want
- Solution: HMM Synthesis
  - Won the last TTS bakeoff
  - Sounds unnatural to researchers
  - But naïve subjects preferred it
  - Has the potential to improve on both diphone and unit selection

HMM Synthesis

Unit selection (Roger)
HMM (Roger)

Unit selection (Nina)
HMM (Nina)

Agenda

- HW5 graded
- HW6 due next Tuesday
- Schedule changes
- Questions, comments, concerns?
- Text-to-Speech (TTS)
  - Text-to-Movies: xtranormal.com