

# LSA 352: Speech Recognition and Synthesis

Dan Jurafsky

## Lecture 3: Intro to Festival; Letter-to-Sound Rules Prosody

IP Notice: lots of info, text, and diagrams on these slides comes (thanks!) from Alan Black's excellent lecture notes and from Richard Sproat's great new slides.

LSA 352 Summer 2007

## Outline

1. Festival
  - Where it lives, its components
  - Its scripting language: Scheme
2. From words to strings of phones
  - Dictionaries
  - Letter-to-Sound Rules
    - ("Grapheme-to-Phoneme Conversion")
3. Prosody
  1. Linguistic Background
    - Prosody, F0, Pitch Accents, Boundaries, Tunes
  2. Producing Intonation in TTS
    - Predicting Accents
    - Predicting Boundaries
    - Predicting Duration
    - Generating F0
  3. Advanced: The TOBI Prosodic Transcription Theory

LSA 352 Summer 2007

## 1. Festival

- Open source speech synthesis system
- Designed for development and runtime use
  - Use in many commercial and academic systems
  - Distributed with RedHat 9.x
  - Hundreds of thousands of users
- Multilingual
  - No built-in language
  - Designed to allow addition of new languages
- Additional tools for rapid voice development
  - Statistical learning tools
  - Scripts for building models

Text from Richard Sproat

LSA 352 Summer 2007

## Festival as software

- <http://festvox.org/festival/>
- General system for multi-lingual TTS
- C/C++ code with Scheme scripting language
- General replaceable modules:
  - Lexicons, LTS, duration, intonation, phrasing, POS tagging, tokenizing, diphone/unit selection, signal processing
- General tools
  - Intonation analysis (f0, Tilt), signal processing, CART building, N-gram, SCFG, WFST

Text from Richard Sproat

LSA 352 Summer 2007

## Festival as software

- <http://festvox.org/festival/>
- No fixed theories
- New languages without new C++ code
- Multiplatform (Unix/Windows)
- Full sources in distribution
- Free software

Text from Richard Sproat

LSA 352 Summer 2007

## CMU FestVox project

- Festival is an engine, how do you make voices?
- Festvox: building synthetic voices:
  - Tools, scripts, documentation
  - Discussion and examples for building voices
  - Example voice databases
  - Step by step walkthroughs of processes
- Support for English and other languages
- Support for different waveform synthesis methods
  - Diphone
  - Unit selection
  - Limited domain

Text from Richard Sproat

LSA 352 Summer 2007

## Synthesis tools

- I want my computer to talk
  - Festival Speech Synthesis
- I want my computer to talk in my voice
  - FestVox Project
- I want it to be fast and efficient
  - Flite

Text from Richard Spina LSA 352 Summer 2007 7

## Using Festival

- How to get Festival to talk
- Scheme (Festival's scripting language)
- Basic Festival commands

Text from Richard Spina LSA 352 Summer 2007 8

## Getting it to talk

- Say a file
  - `festival --tts file.txt`
- From Emacs
  - `say region, say buffer`
- Command line interpreter
  - `festival> (SayText "hello")`

Text from Richard Spina LSA 352 Summer 2007 9

## Scheme: the scripting lg

- Advantages of a scripting lg
  - Convenient, easy to add functionality
- Why Scheme?
  - Holdover from the LISP days of AI.
  - Many people like it.
  - It's very simple
  - We're stuck with it.

LSA 352 Summer 2007  
Text adapted from Richard Spina 10

## Quick Intro to Scheme

- Scheme is a dialect of LISP
- expressions are
  - atoms or
  - lists
    - `a bcd "hello world" 12.3`
    - `(a b c)`
    - `(a (1 2) seven)`
- Interpreter evaluates expressions
  - Atoms evaluate as variables
  - Lists evaluate as functional calls
    - `bxx`
    - `3.14`
    - `(+ 2 3)`

LSA 352 Summer 2007  
Text from Richard Spina 11

## Quick Intro to Scheme

- Setting variables
  - `(set! a 3.14)`
- Defining functions
  - `(define (timestwo n) (* 2 n))`
  
  - `(timestwo a)`
  - `6.28`

LSA 352 Summer 2007  
Text from Richard Spina 12

## Lists in Scheme

```
festival> (set! alist '(apples pears bananas))
(apples pears bananas)
festival> (car alist)
apples
festival> (cdr alist)
(pears bananas)
festival> (set! blist (cons 'oranges alist))
(oranges apples pears bananas)
festival> append alist blist
#<SUBR(6) append>
(apples pears bananas)
(oranges apples pears bananas)
festival> (append alist blist)
(apples pears bananas oranges apples pears bananas)
festival> (length alist)
3
festival> (length (append alist blist))
7
```

## Scheme: speech

```
Make an utterance of type text
festival> (set! uttl (Utterance Text "hello"))
#<Utterance 0xf6855718>

Synthesize an utterance
festival> (utt.synth uttl)
#<Utterance 0xf6855718>

Play waveform
festival> (utt.play uttl)
#<Utterance 0xf6855718>

Do all together
festival> (SayText "This is an example")
#<Utterance 0xf6961618>
```

## Scheme: speech

```
In a file
(define (SpeechPlus a b)
  (SayText
   (format nil
            "%d plus %d equals %d"
            a b (+ a b))))

Loading files
festival> (load "file.scm")
t

Do all together
festival> (SpeechPlus 2 4)
#<Utterance 0xf6961618>
```

## Scheme: speech

```
(define (sp_time hour minute)
  (cond
   (( < hour 12)
    (SayText
     (format nil
              "It is %d %d in the morning"
              hour minute )))
   (( < hour 18)
    (SayText
     (format nil
              "It is %d %d in the afternoon"
              (- hour 12) minute )))
   (t
    (SayText
     (format nil
              "It is %d %d in the evening"
              (- hour 12) minute )))))
```

## Getting help

- Online manual
  - <http://festvox.org/docs/manual-1.4.3>
- Alt-h (or esc-h) on current symbol short help
- Alt-s (or esc-s) to speak help
- Alt-m goto man page
- Use TAB key for completion

## Festival Structure

- Utterance structure in Festival
- [http://www.festvox.org/docs/manual-1.4.2/festival\\_14.html](http://www.festvox.org/docs/manual-1.4.2/festival_14.html)
  - Features in festival
  - [http://www.festvox.org/docs/manual-1.4.2/festival\\_32.html](http://www.festvox.org/docs/manual-1.4.2/festival_32.html)

## II. From Words to Phones

- Dictionaries
- Letter-to-Sound Rules

LSA 352 Summer 2007

19

## Converting 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

LSA 352 Summer 2007

20

## 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]

LSA 352 Summer 2007

21

## Pronunciation Dictionaries: UNISYN

- UNISYN dictionary: 110K words (Fitt 2002)
    - <http://www.cstr.ed.ac.uk/projects/unisyn/>
- going: { g \* ou } .> i n g >  
 antecedents: { \* a n . t i . s i i . d n ! t } > s >  
 dictionary: { d \* i k . s h @ . n e . r i i }
- Benefits:
    - Has syllabification, stress, some morphological boundaries
    - Pronunciations can be read off in
      - General American
      - RP British
      - Australia
      - Etc
  - (Other dictionaries like CELEX not used because too small, British-only)

LSA 352 Summer 2007

22

## Lexical Entries in Festival

- You can explicitly give pronunciations for words
  - Each lg/dialect has its own separate lexicon file
  - You can lookup words with
    - (lex.lookup WORD)
  - You can add entries to the current lexicon
    - (lex.add.entry NEWENTRY)
  - Entry: (WORD POS (SYLO SYLL...))
  - Syllable: ((PHONE PHONE1 ...) STRESS )
  - Example:
 

```
'("cepstra" n ((k eh p) 1) ((s t r aa) 0))
```

LSA 352 Summer 2007

23

## Dictionaries aren't sufficient

- Unknown words (= OOV = "out of vocabulary")
  - 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

LSA 352 Summer 2007

24

## 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

LBA 352 Summer 2007

24

## 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)
  - Lieberman 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)

LBA 352 Summer 2007

26

## Acronyms

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

LBA 352 Summer 2007

27

## Letter-to-Sound Rules

- Earliest algorithms: handwritten Chomsky+Halle-style rules:
  - $c \rightarrow [k] / \_ \{a,o\} V$  ; context-dependent
  - $c \rightarrow [s] / \_$  ; context-independent
- Festival version of such LTS rules:
  - (LEFTCONTEXT [ ITEMS] RIGHTCONTEXT = NEWITEMS )
- Example:
  - ( # [ c h ] C = k )
  - ( # [ c h ] = ch )
- # denotes beginning of word
- C means all consonants
- Rules apply in order
  - "christmas" pronounced with [k]
  - But word with ch followed by non-consonant pronounced [ch]
    - E.g., "choice"

LBA 352 Summer 2007

28

## Stress rules in hand-written LTS

- English famously evil: one from Allen et al 1987
  - $V \rightarrow [+stress] / X \_ C^* \{V_{short} C C^?|V\} \{V_{short} C^*|V\}$
- Where X must contain all prefixes:
- Assign 1-stress to the vowel in a syllable preceding a weak syllable followed by a morpheme-final syllable containing a short vowel and 0 or more consonants (e.g. *difficult*)
- Assign 1-stress to the vowel in a syllable preceding a weak syllable followed by a morpheme-final vowel (e.g. *oregano*)
- etc

LBA 352 Summer 2007

29

## 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**

LBA 352 Summer 2007

30

## Alignment

- Letters: c h e c k e d L: c a k e
- Phones: ch \_ eh \_ k \_ t | | | |
- Black et al Method 1: P: K EY K ε
  - 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

LSA 352 Summer 2007

31

## Alignment

- Black et al method 2

c:k ch s sh t-s ε  
e:lh iy er ax ah eh ey uw ay ow y-uw oy aa ε

LSA 352 Summer 2007

32

## Hand specify which letters can be rendered as which phones

- C goes to k/ch/s/sh
- W goes to w/v/f, etc
- An actual list:
- Once mapping table is created, find all valid alignments, find p(letter|phone), score all alignments, take best

LSA 352 Summer 2007

33

## 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

LSA 352 Summer 2007

34

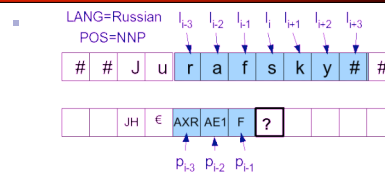
## Building CART trees

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

LSA 352 Summer 2007

35

## 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*

LSA 352 Summer 2007

36

## Prosody: Linguistic Background

LSA 352 Summer 2007

17

## 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.

LSA 352 Summer 2007

18

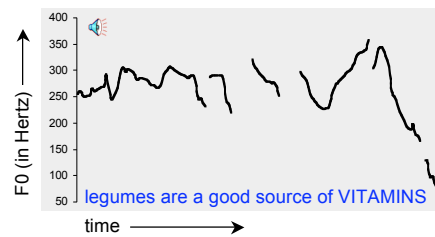
## Three aspects of prosody

- **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.

LSA 352 Summer 2007

19

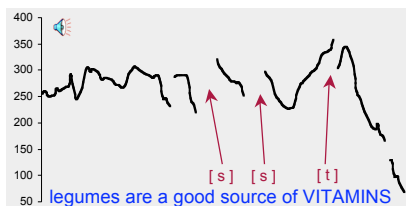
## Graphic representation of F0



LSA 352 Summer 2007

Slide From Jennifer Venditti

## The 'ripples'

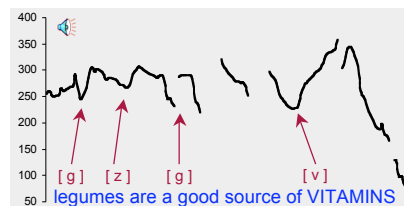


F0 is not defined for consonants without vocal fold vibration.

LSA 352 Summer 2007

Slide From Jennifer Venditti

## The 'ripples'

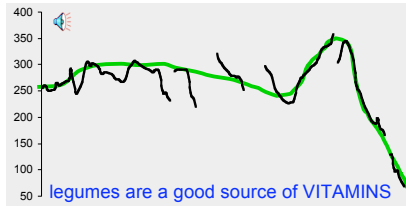


... and F0 can be perturbed by consonants with an extreme constriction in the vocal tract.

LSA 352 Summer 2007

Slide From Jennifer Venditti

## Abstraction of the F0 contour



Our perception of the intonation contour abstracts away from these perturbations.

LSA 352 Summer 2007 Slide From Jennifer Venditti

## Prominence: Placement of Pitch Accents

LSA 352 Summer 2007 44

## Stress vs. accent

- **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	syllables
vi	ta	mins	Ca	li for nia

LSA 352 Summer 2007 Slide From Jennifer Venditti

## Stress vs. accent (2)

- 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**

LSA 352 Summer 2007 46

## 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.

LSA 352 Summer 2007 Slide From Jennifer Venditti

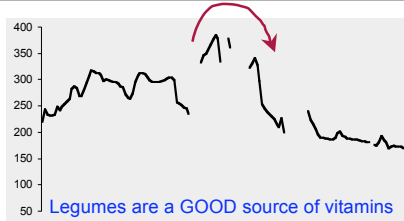
## Same 'tune', different alignment



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

LSA 352 Summer 2007 Slide From Jennifer Venditti

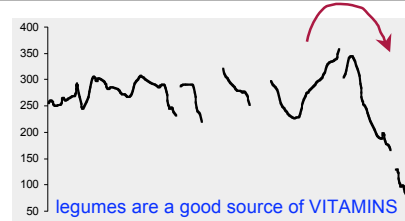
## Same 'tune', different alignment



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

LBA 352 Summer 2007 Slide From Jennifer Venditti

## Same 'tune', different alignment



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

LBA 352 Summer 2007 Slide From Jennifer Venditti

## 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**

LBA 352 Summer 2007

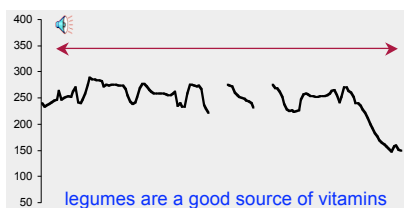
81

## Intonational phrasing/boundaries

LBA 352 Summer 2007

82

## A single intonation phrase



Broad focus statement consisting of one intonation phrase (that is, one intonation tune spans the whole unit).

LBA 352 Summer 2007 Slide From Jennifer Venditti

## Multiple phrases



Utterances can be 'chunked' up into smaller phrases in order to signal the importance of information in each unit.

LBA 352 Summer 2007 Slide From Jennifer Venditti

I wanted to go to London, but could only get tickets for France

LSA 352 Summer 2007  
Slide From Jennifer Vendetti

### Phrasing sometimes helps disambiguate

- Temporary ambiguity:**  
When Madonna sings the song ...

LSA 352 Summer 2007  
Slide From Jennifer Vendetti

### Phrasing sometimes helps disambiguate

- Temporary ambiguity:**  
When Madonna sings the song is a hit.

LSA 352 Summer 2007  
Slide From Jennifer Vendetti

### 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)]

LSA 352 Summer 2007  
Slide From Jennifer Vendetti

### Phrasing sometimes helps disambiguate

I met Mary and Elena's mother at the mall yesterday

One intonation phrase with relatively flat overall pitch range.

LSA 352 Summer 2007  
Slide From Jennifer Vendetti

### Phrasing sometimes helps disambiguate

I met Mary and Elena's mother at the mall yesterday

Separate phrases, with expanded pitch movements.

LSA 352 Summer 2007  
Slide From Jennifer Vendetti

## Intonational tunes

---

LSA 352 Summer 2007 41

## Yes-No question tune

---

are LEGUMES a good source of vitamins

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

LSA 352 Summer 2007 Slide From Jennifer Vend...

## Yes-No question tune

---

are legumes a GOOD source of vitamins

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

LSA 352 Summer 2007 Slide From Jennifer Vend...

## Yes-No question tune

---

are legumes a good source of VITAMINS

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

LSA 352 Summer 2007 Slide From Jennifer Vend...

## WH-questions

---

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

WHAT are a good source of vitamins

WH-questions typically have falling contours, like statements.

LSA 352 Summer 2007 Slide From Jennifer Vend...

## Broad focus

---

"Tell me something about the world."

legumes are a good source of vitamins

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

LSA 352 Summer 2007 Slide From Jennifer Vend...

## Rising statements

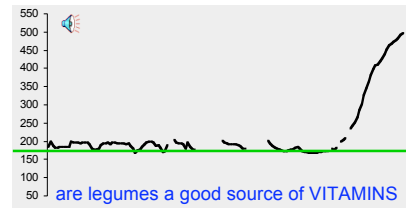
"Tell me something I didn't already know."



High-rising statements can signal that the speaker is seeking approval.

LSA 352 Summer 2007 Slide 11 Jennifer Venditti

## Yes-No question

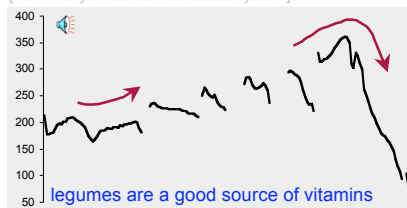


Rise from the main accent to the end of the sentence.

LSA 352 Summer 2007 Slide 12 Jennifer Venditti

## 'Surprise-redundancy' tune

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

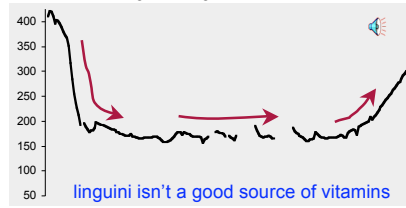


Low beginning followed by a gradual rise to a high at the end.

LSA 352 Summer 2007 Slide 13 Jennifer Venditti

## 'Contradiction' tune

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



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

LSA 352 Summer 2007 Slide 14 Jennifer Venditti

## 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

LSA 352 Summer 2007 11

## Predicting pitch accent

LSA 352 Summer 2007 12

## Factors in accent prediction

- 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

LSA 352 Summer 2007

74

## Simplest possible algorithm for pitch accent assignment

```
(set! simple_accent_cart_tree
  '
  (
    (R:SylStructure.parent.gpos is content)
    ( (stress is 1)
      ((Accented))
      ((NONE))
    )
  )
)
```

LSA 352 Summer 2007

74

## But not just function/content:

- A Broadcast News example from Hirschberg (1993)
- SUN MICROSYSTEMS INC, the UPSTART COMPANY that HELPED LAUNCH the DESKTOP COMPUTER industry TREND TOWARD HIGH powered WORKSTATIONS, was UNVEILING an ENTIRE OVERHAUL of its PRODUCT LINE TODAY. SOME of the new MACHINES, PRICED from FIVE THOUSAND NINE hundred NINETY five DOLLARS to seventy THREE thousand nine HUNDRED dollars, BOAST SOPHISTICATED new graphics and DIGITAL SOUND TECHNOLOGIES, HIGHER SPEEDS AND a CIRCUIT board that allows FULL motion VIDEO on a COMPUTER SCREEN.

LSA 352 Summer 2007

75

## Factors in accent prediction

- 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

LSA 352 Summer 2007

76

## List intonation

- I went and saw ANNA, LENNY, MARY, and NORA.

LSA 352 Summer 2007

77

## Word order

- Preposed items are accented more frequently
- TODAY we will BEGIN to LOOK at FROG anatomy.
- We will BEGIN to LOOK at FROG anatomy today.

LSA 352 Summer 2007

78

## Information Status

- New versus old information.
- Old information is deaccented
- Something can be old because of explicit lexical repetition, or more subtly:
  - There are LAWYERS, and there are GOOD lawyers
  - EACH NATION DEFINES its OWN national INTERST.
  - I LIKE GOLDEN RETRIEVERS, but MOST dogs LEAVE me COLD.

LSA 352 Summer 2007

79

## Complex Noun Phrase Structure

- Sproat, R. 1994. English noun-phrase accent prediction for text-to-speech. *Computer Speech and Language* 8:79-94.
- Proper Names, stress on right-most word
  - New York CITY; Paris, FRANCE
- Adjective-Noun combinations, stress on noun
  - Large HOUSE, red PEN, new NOTEBOOK
- Noun-Noun compounds: stress left noun
  - HOTdog (food) versus HOT DOG (overheated animal)
  - WHITE house (place) versus WHITE HOUSE (made of stucco)
- examples:
  - MEDICAL Building, APPLE cake, cherry PIE...
  - What about: Madison avenue, Park street ???
- Some Rules:
  - Furniture+Room -> RIGHT (e.g., kitchen TABLE)
  - Proper-name + Street -> LEFT (e.g. PARK street)

LSA 352 Summer 2007

80

## 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

LSA 352 Summer 2007

81

## Advanced features

- Accent is often deflected away from a word due to focus on a neighboring word.
- Could use syntactic parallelism to detect this kind of contrastive focus:
  - .....driving [FIFTY miles] an hour in a [THIRTY mile] zone
  - [WELD] [APPLAUDS] mandatory recycling. [SILBER] [DISMISSES] recycling goals as meaningless.
  - ...but while Weld may be [LONG] on people skills, he may be [SHORT] on money

LSA 352 Summer 2007

82

## 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

LSA 352 Summer 2007

83

## Predicting boundaries

LSA 352 Summer 2007

84

## Predicting Boundaries

- Intonation phrase boundaries
  - Intermediate phrase boundaries
  - Full intonation phrase boundaries
- Police also say | Levy's blood alcohol level | was twice the legal limit ||

LBA 352 Summer 2007

85

## Simplest CART

```
(set! simple_phrase_cart_tree
'
((lisp_token_end_punc in ("?" "." ":"))
((BB))
((lisp_token_end_punc in ("'" "\"" ","
";"))
((B))
((n.name is 0) ;; end of utterance
((BB))
((NB))))))
```

LBA 352 Summer 2007

86

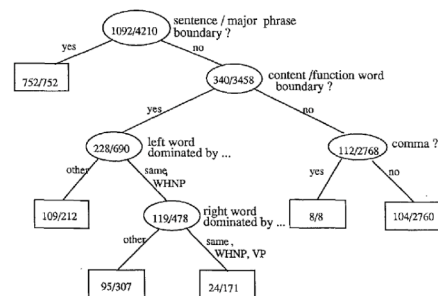
## More complex features

- 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

LBA 352 Summer 2007

87

## Ostendorf and Veilleux CART



LBA 352 Summer 2007

88

## TOPIC II.3

### Predicting duration

LBA 352 Summer 2007

89

## Duration

- Simplest: fixed size for all phones (100 ms)
- Next simplest: average duration for that phone (from training data). Samples from SWBD in ms:
 

aa	118	b	68
ax	59	d	68
ay	138	dh	44
eh	87	f	90
ih	77	g	66
- Next Next Simplest: add in phrase-final and initial lengthening plus stress

LBA 352 Summer 2007

90

## Klatt duration rules

Models how context-neutral duration of a phone lengthened/shortened by context

- While staying above a min duration  $d_{\min}$
- Prepausal lengthening:
  - The vowel or syllabic consonant in the syllable before a pause is lengthened by 1.4
- Non-phrase-final shortening
  - Segments which are not phrase-final are shortened by 0.6. Phrase-final postvocalic liquids and nasals are lengthened by 1.4
- Unstressed shortening
  - Unstressed segments are more compressible, so their minimum duration  $d_{\min}$  is halved, and are shortened by .7 for most phone types.
- Lengthening for accent
  - A vowel which bears accent is lengthened by 1.4
- Shortening in clusters
  - A consonant followed by a consonant is shortened by 0.5
- Pre-voiceless shortening
  - Vowels are shortened before a voiceless plosive by 0.7

LBA 352 Summer 2007

81

## Klatt duration rules

- Klatt formula for phone durations:

$$d = d_{\min} + \prod_{i=1}^N f_i \times (\bar{d} - d_{\min})$$

- Festival: 2 options
  - Klatt rules
  - Use labeled training set with Klatt features to train CART
    - Identity of the left and right context phone
    - Lexical stress and accent values of current phone
    - Position in syllable, word, phrase
    - Following pause

LBA 352 Summer 2007

82

## Duration: state of the art

- Lots of fancy models of duration prediction:
  - Using Z-scores and other clever normalizations
  - Sum-of-products model
  - New features like word predictability
    - Words with higher bigram probability are shorter

LBA 352 Summer 2007

83

## Duration in Festival

```
(set: spanish_dur_tree
'
((R:SylStructure.parent.R:Syllable.p.syl_break > 1)
;; clause initial
((R:SylStructure.parent.stress is 1)
((1.5))
((1.2)))
((R:SylStructure.parent.syl_break > 1) ;; clause
final
((R:SylStructure.parent.stress is 1)
((2.0))
((1.5))
((R:SylStructure.parent.stress is 1)
((1.2))
((1.0))))))
```

LBA 352 Summer 2007

84

## F0 Generation

- Generation in Festival
  - F0 Generation by rule
  - F0 Generation by linear regression
- Some constraints
  - F0 is constrained by accents and boundaries
  - F0 declines gradually over an utterance ("declination")

LBA 352 Summer 2007

85

LBA 352 Summer 2007

86

## F0 generation by rule

- F0 is generally defined relative to **pitch range**
  - A speaker's pitch range is the range between
    - Baseline frequency: lowest freq in a particular utterance
    - Topline frequency: highest freq in a particular utterance
- Jilka et al (1999)
  - Beginning of utterance: target point of 50%
  - Target point for H\* accent: 100%
  - Target point for L\* accent: 0%
  - Target point for L+H\* accent: 20% and 100%
  - Target point for H-H%: extra-high 120%
  - Target point for L-L% extra-low -20%
- Alignment: where accent lies in syllable
  - H\* accent: aligned 60% through syllable
  - IP-initial accent: somewhat earlier

LSA 352 Summer 2007

97

## F0 Generation by rule in Festival

- Generate a list of target F0 points for each syllable
- Here's a rule to generate a simple H\* "hat" accent (with fixed = speaker-specific F0 values):

```
(define (targ_func1 utt syl)
  "(targ_func1 UTF STREAMITEM)
  Returns a list of targets for the given syllable."
  (let ((start (item.feats syl 'syllable_start))
        (end (item.feats syl 'syllable_end)))
    (if (equal? (item.feats syl "R:Intonation.daughter1.name")
              "Accented")
        (list
         (list start 110)
         (list (/ (+ start end) 2.0) 140)
         (list end 100))))))
```

LSA 352 Summer 2007

98

## F0 generation by regression

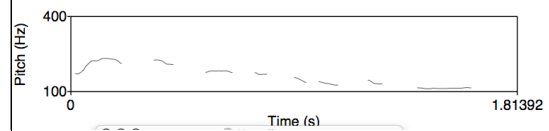
- Supervised machine learning again
- We predict: value of F0 at 3 places in each syllable
- Predictor features:
  - Accent of current word, next word, previous
  - Boundaries
  - Syllable type, phonetic information
  - Stress information
- Need training sets with pitch accents labeled

LSA 352 Summer 2007

99

## Declination

- F0 tends to decline throughout a sentence



LSA 352 Summer 2007

100

## Advanced: Intonational Transcription Theories: ToBI and Tilt

LSA 352 Summer 2007

101

## ToBI: Tones and Break Indices

- Pitch accent tones
  - H\* "peak accent"
  - L\* "low accent"
  - L+H\* "rising peak accent" (contrastive)
  - L\*+H "scooped accent"
  - H+H\* downstepped high
- Boundary tones
  - L-L% (final low; Am Eng. Declarative contour)
  - L-H% (continuation rise)
  - H-H% (yes-no question)
- Break indices
  - 0: clitics, 1, word boundaries, 2 short pause
  - 3 intermediate intonation phrase
  - 4 full intonation phrase/final boundary.

LSA 352 Summer 2007

102

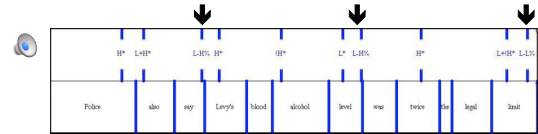
## Examples of the TOBI system

- I don't eat beef.
  - L\* L\* L\*L-L%
- Marianna made the marmalade.
  - H\* L-L%
  - L\* H-H%
- "I" means insert.
  - H\* H\* H\*L-L%
  - 1
  - H\*L- 3
  - H\*L-L%

LSA 352 Summer 2007 104

## Predicting Boundaries

- Intonation phrase boundaries
  - Intermediate phrase boundaries
  - Full intonation phrase boundaries
- Police also say | Levy's blood alcohol level | was twice the legal limit ||



LSA 352 Summer 2007 104

## ToBI

- <http://www.ling.ohio-state.edu/~tobi/>
- TOBI for American English
  - [http://www.ling.ohio-state.edu/~tobi/ame\\_tobi/](http://www.ling.ohio-state.edu/~tobi/ame_tobi/)
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: a standard for labelling English prosody. In *Proceedings of ICSLP92*, volume 2, pages 867-870
- Pitrelli, J. F., Beckman, M. E., and Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *ICSLP94*, volume 1, pages 123-126
- Pierrehumbert, J., and J. Hirschberg (1990) The meaning of intonation contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, and M. E. Pollack, eds., *Plans and Intentions in Communication and Discourse*, 271-311. MIT Press.
- Beckman and Elam. Guidelines for ToBI Labelling. Web.

LSA 352 Summer 2007 105

## TILT

- Like ToBI, a sequence of intonational events like accents and boundary tones
- But instead of ToBI-style phonemic categories
  - Each event modeled by continuous parameters representing F0 shape
  - Trained on a corpus labeled for pitch accents and boundary tones
  - Human label just specifies syllable; parameters learned automatically

LSA 352 Summer 2007 106

## TILT

- - Each accent in tilt is (optional) rise followed by (optional) fall
  - Tilt value: 1.0=rise, -1.0 = fall, 0=equal rise and fall

$$\text{tilt} = \frac{\text{tilt}_{\text{amp}} + \text{tilt}_{\text{dur}}}{2}$$

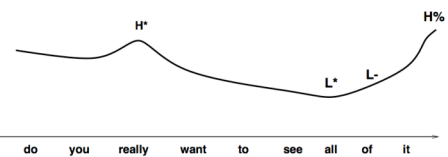
$$= \frac{|A_{\text{rise}}| - |A_{\text{fall}}|}{|A_{\text{rise}}| + |A_{\text{fall}}|} + \frac{D_{\text{rise}} - D_{\text{fall}}}{D_{\text{rise}} + D_{\text{fall}}}$$

LSA 352 Summer 2007 107

## Intermediate representation: using Festival

- Do you really want to see all of it?

do	you	really	want	to	see	all	of	it													
d	uw	y	r	ih	l	iy	w	aa	n	t	t	ax	s	iy	ao	l	ah	v	ih	t	
110	110	50	50	75	64	57	82	57	50	72	41	43	47	54	130	76	90	44	62	46	220



LSA 352 Summer 2007 108