CS 294-5: Statistical Natural Language Processing

Speech Synthesis
Lecture 22: 12/4/05

Slides directly from Dan Jurafsky, indirectly many others

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

Types of Modern Synthesis

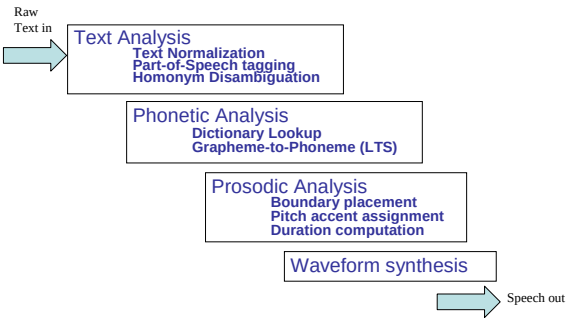
- Articulatory Synthesis:
  - Model movements of articulators and acoustics of vocal tract
- Formant Synthesis:
  - Start with acoustics, create rules/filters to create each formant
- Concatenative Synthesis:
  - Use databases of stored speech to assemble new utterances.

Text from Richard Sproat slides

TTS Demos (Mostly Unit-Selection)

- Comparisons:
  - http://www.tmaa.com/tts/companies.htm
- ATT:
  - http://www.naturalvoices.att.com/demos/
- Rhetorical (= Scansoft)
  - http://www.rhetorical.com/cgi-bin/demo.cgi
- Festival
  - http://www-2.cs.cmu.edu/~awb/festival_demos/index.html
- IBM
  - http://www-306.ibm.com/software/pervasive/tech/demos/tts.shtml

TTS Architecture



Text Normalization

- Analysis of raw text into pronounceable words
- Sample problems:
  - `He stole $100 million from the bank`
  - `It's 13 St. Andrews St.`
  - `The home page is http://www.cnn.com`
  - `yes, see you the following tues, that's 11/12/01`
- Steps
  - Identify tokens in text
  - Chunk tokens into reasonably sized sections
  - Map tokens to words
  - Identify types for words

Words to Phones

- Two methods:
  - Dictionary-based
  - Rule-based (Letter-to-sound=LTS)
- Early systems, all LTS
- MITalk was radical in having huge 10K word dictionary
- Now systems use a combination
  - Big dictionary
  - Special code for handling names
  - Machine learned LTS system for other unknown words
- CMU dictionary: 127K words
  - http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Letter-to-Sound Rules

- Festival LTS rules:
  (LEFTCONTEXT [ITEMS] RIGHTCONTEXT = NEWITEMS)
- Examples:
  - (# [c h] C = k)
  - (# [c h] = ch)
- Rules apply in order
  - "christmas" pronounced with [k]
  - But word with ch followed by non-consonant pronounced [ch]
    - E.g., "choice"
- More modern approach: learn HMMs / CRFs

Prosody

- Prosody:
  - Getting from words+phones to boundaries, accent, F0, duration
- Prosodic phrasing
  - Need to break utterances into phrases
  - Punctuation is useful, not sufficient
- Accents:
  - Predictions of accents: which syllables should be accented
  - Realization of F0 contour: given accents/tones, generate F0 contour
- Duration:
  - Predicting duration of each phone

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)

Prominence: Pitch Accents

A: What types of foods are a good source of vitamins?

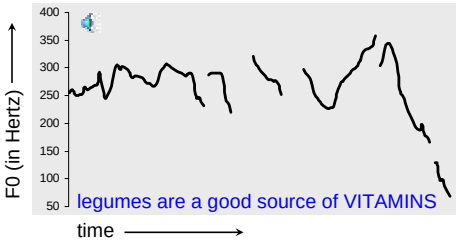
B1: Legumes are a good source of VITAMINS.

B2: LEGUMES are a good source of vitamins.

- Prominent syllables are:
  - Louder
  - Longer
  - Have higher F0 and/or sharper changes in F0 (higher F0 velocity)

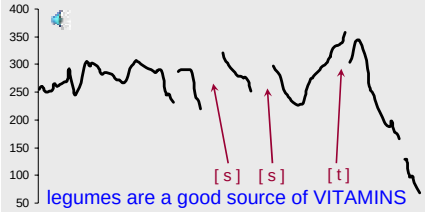
Slide from Jennifer Venditti

Graphic representation of F0



Slide from Jennifer Venditti

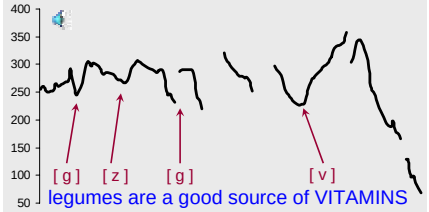
The ‘ripples’



F0 is not defined for consonants without vocal fold vibration.

Slide from Jennifer Venditti

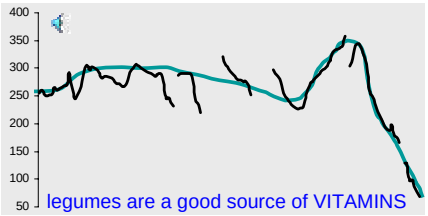
The ‘ripples’



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

Slide from Jennifer Venditti

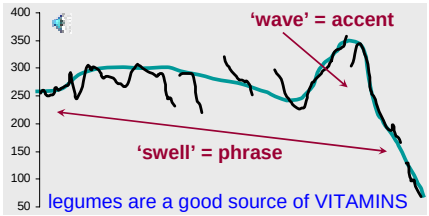
Abstraction of the F0 contour



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

Slide from Jennifer Venditti

The ‘waves’ and the ‘swells’



Slide from Jennifer Venditti

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	x	x	syllables
vi	ta	mins		Ca	li	for	nia

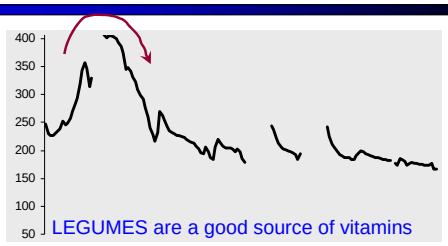
Slide from Jennifer Venditti

Which Word is Accented?

- **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.

Slide from Jennifer Venditti

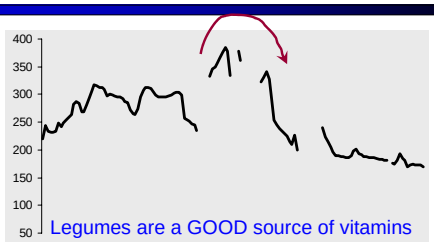
Same ‘tune’, different alignment



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

Slide from Jennifer Venditti

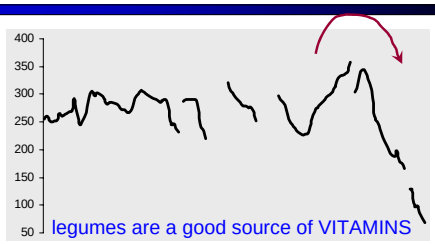
Same ‘tune’, different alignment



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

Slide from Jennifer Venditti

Same ‘tune’, different alignment

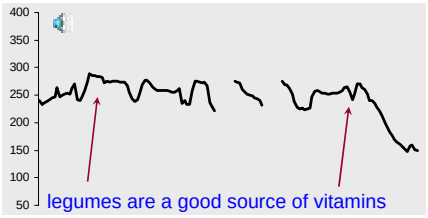


The main **rise-fall** accent (= “I assert this”) shifts locations.

Slide from Jennifer Venditti

Broad focus

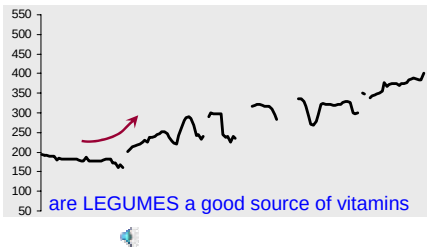
“Tell me something about the world.”



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

Slide from Jennifer Venditti

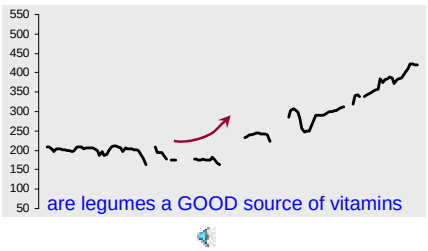
Yes-No question tune



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

Slide from Jennifer Venditti

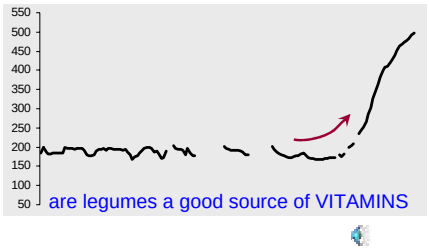
Yes-No question tune



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

Slide from Jennifer Venditti

Yes-No question tune

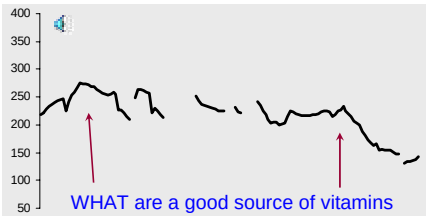


Rise from the main accent to the end of the sentence.

Slide from Jennifer Venditti

WH-questions

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

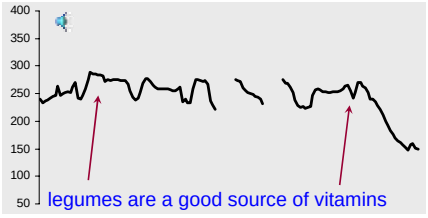


WH-questions typically have falling contours, like statements.

Slide from Jennifer Venditti

Broad focus

"Tell me something about the world."



Slide from Jennifer Venditti

Rising statements

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

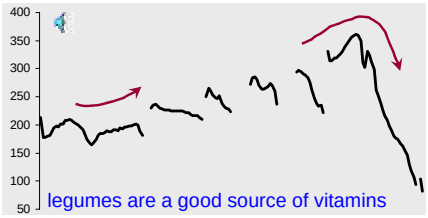
[... does this statement qualify?]

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

Slide from Jennifer Venditti

'Surprise-redundancy' tune

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

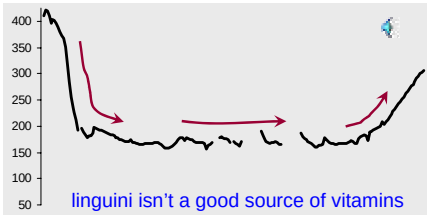


Low beginning followed by a gradual rise to a high at the end.

Slide from Jennifer Venditti

'Contradiction' tune

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

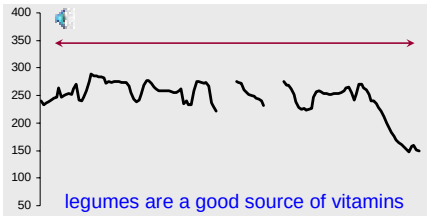


[... how could you think that?]

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

Slide from Jennifer Venditti

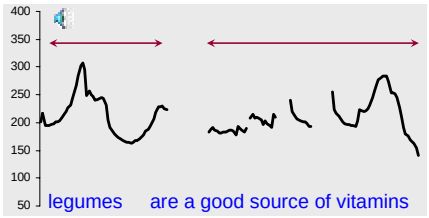
A single intonation phrase



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

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.

Slide from Jennifer Venditti

Phrasing can disambiguate

- **Global ambiguity:**

The old men and women stayed home.
The old men **%** and women **%** stayed home.

Sally saw **%** the man with the binoculars.
Sally saw the man **%** with the binoculars.

John doesn't drink because he's unhappy.
John doesn't drink **%** because he's unhappy.

Slide from Jennifer Venditti

Phrasing can disambiguate

- **Temporary ambiguity:**

When Madonna sings the song ...

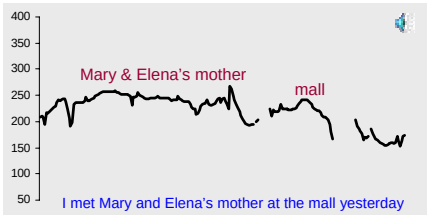
When Madonna sings **%** the song is a hit.

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

[from Speer & Kjelgaard (1992)]

Slide from Jennifer Venditti

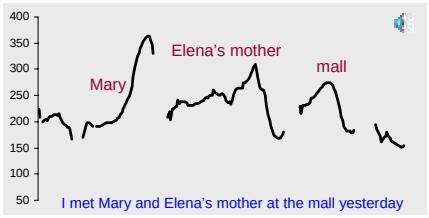
Phrasing can disambiguate



One intonation phrase with relatively flat overall pitch range.

Slide from Jennifer Venditti

Phrasing can disambiguate



Separate phrases, with expanded pitch movements.

Slide from Jennifer Venditti

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 queston)
- Break indices
  - 0: clitics, 1, word boundaries, 2 short pause
  - 3 intermediate intonation phrase
  - 4 full intonation phrase/final boundary.

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- H*L-L%
  3

Slide from Lavoie and Podesva

Intonation in TTS

1) 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

Factors in accent prediction

- Contrast
  - Legumes are poor source of VITAMINS
  - No, legumes are a GOOD source of vitamins
  - I think JOHN and MARY should go
  - No, I think JOHN AND MARY should go

But it's more than just contrast

- List intonation:
  - I went and saw ANNA, LENNY, MARY, and NORA.
- Part of speech:
  - Content words are usually accented
  - Function words are rarely accented
- 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.
- Information Status:
  - New versus old information.
  - Old information is not deaccented
  - There are LAWYERS, and there are GOOD lawyers
  - EACH NATION DEFINES its OWN national INTERST.

Complex NP 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:
  - Madison AVENUE, park STREET, MEDICAL building
  - APPLE cake, cherry PIE
- Some Rules:
  - Furniture+Room -> RIGHT (e.g., kitchen TABLE)
  - Proper-name + Street -> LEFT (e.g. PARK street)

State-of-the-Art

- Supervised systems
  - Hand-labeled accented data
  - Feature driven
- More 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

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

Duration

- Klatt duration rules: modify duration based on:
  - Position in clause
  - Syllable position in word
  - Syllable type
  - Lexical stress
  - Left+right context phone
  - Prepausal lengthening
- Supervised systems now used

F0 generation by regression

- Supervised learning again
- 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

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

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

Diphone TTS architecture

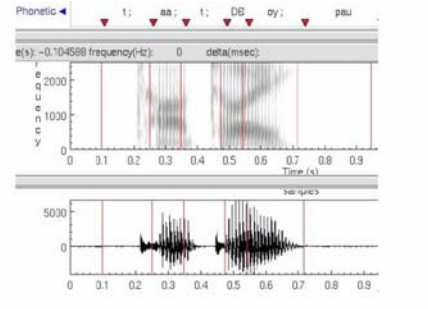
- Collecting diphones:
  - Record diphones in correct contexts
    - l sounds different in onset than coda
    - t is flapped sometimes, etc.
  - Need quiet recording room, etc.
  - Need to label them very very exactly
- Training:
  - Choose units (kinds of diphones)
  - Record diphones
  - Label diphones (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

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

Slide from Richard Sproat

Diphone Boundaries, Ends



Slide from Richard Sproat

Diphones

- Mid-phone is more stable than edge
- Need O(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 MB for English (16 KHz 16 bit)

Slide from Richard Sproat

Diphone Synthesis

- 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

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)

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
- 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

- 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"

Slide from Paul Taylor

Target Costs

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

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

Slide from Paul Taylor

How to set target cost weights

- 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

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:

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

Slide from Paul Taylor

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

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:

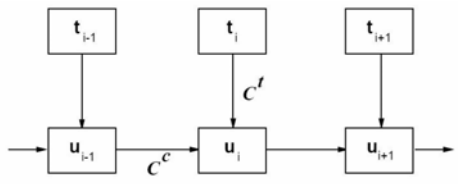
$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^{target}(t_i, u_i) + \sum_{i=2}^{n} C^{join}(u_{i-1}, u_i)$$

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

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

Slide from Paul Taylor

Unit Selection Search



Slide from Richard Sproat

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

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

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

Text from Paul Taylor and Richard Sproat

Creating database

- Unlike 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

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

Slide from Richard Sproat

Joining Units (+F0 + duration)

- Both diphone and unit selection synthesis need to join the units
- For diphone synthesis, need to modify F0 and duration
- For unit selection, in principle also need to modify F0 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.

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)

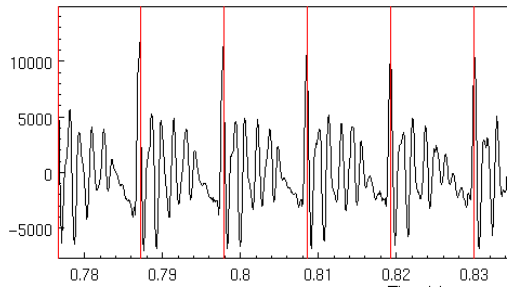
Alan Black

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

Text from Alan Black

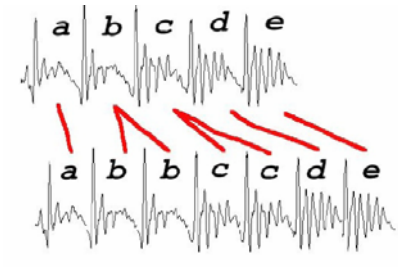
Speech as Short Term signals



Alan Black

Duration modification

- Duplicate/remove short term signals



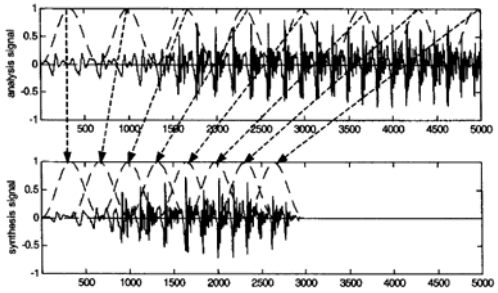
Slide from Richard Sproat

Pitch Modification

- Move short term signals closer together/further apart

Slide from Richard Sproat

Overlap-and-add (OLA)



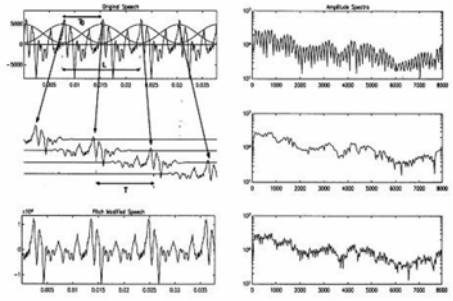
Huang, Acero and Hon

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

Slide from Richard Sproat

TD-PSOLA ™



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)

Huang, Acero, Hon