CS 224S / LINGUIST 285
Spoken Language Processing

Andrew Maas
Stanford University
Spring 2022

Lecture 15: Text-to-Speech. Text Normalization. Prosody

Original slides by Dan Jurafsky, Alan Black, & Richard Sproat

Outline

- Text analysis
  - Text normalization
  - Letter-to-sound (grapheme-to-phoneme)
- TTS modeling history and overview
- Prosody and intonation

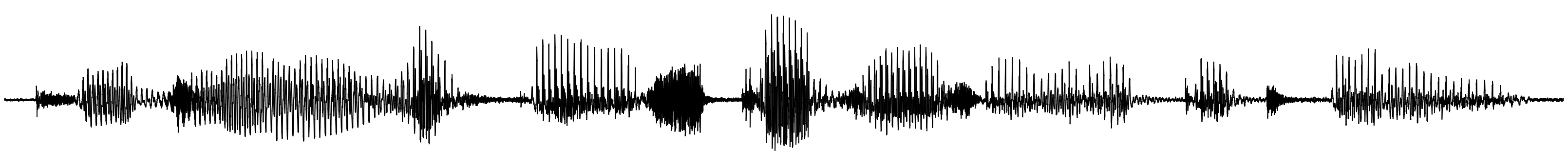
The two stages of TTS

PG&E will file schedules on April 20.

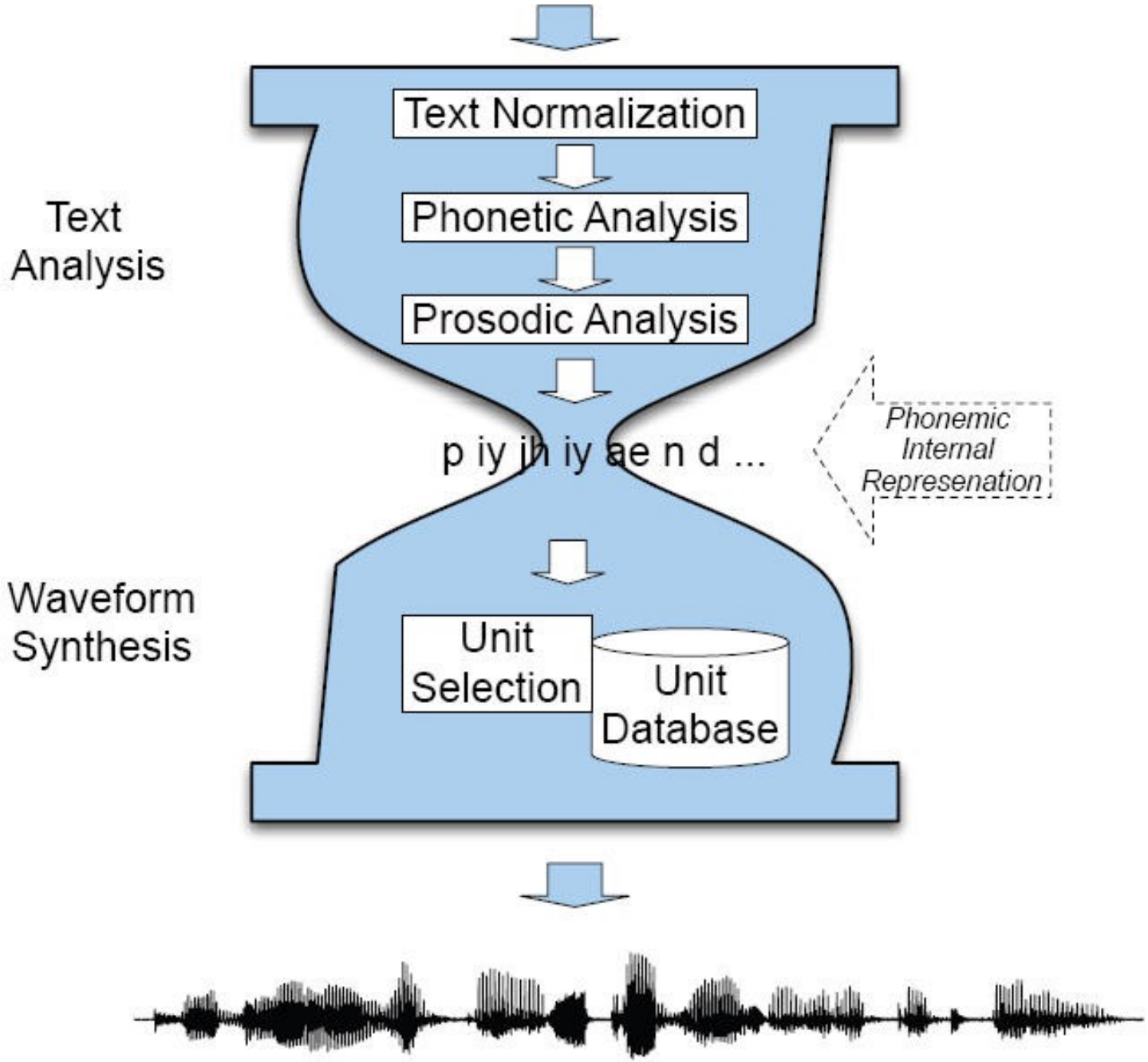
1. Text Analysis: Text into intermediate representation:

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

2. Waveform Synthesis: From the intermediate representation into waveform



PG&E will file schedules on April 20.



Text Normalization

Analysis of raw text into pronounceable words:

He said the increase in credit limits helped B.C. Hydro achieve record net income of about $1 billion during the year ending March 31. This figure does not include any write-downs that may occur if Powerex determines that any of its customer accounts are not collectible. Cousins, however, was insistent that all debts will be collected: "We continue to pursue monies owing and we expect to be paid for electricity we have sold."

- Sentence Tokenization
- Text Normalization
  - Identify tokens in text
  - Chunk tokens into reasonably sized sections
  - Map tokens to words
  - Tag the words

Text Processing

- He stole $100 million from the bank
- It's 13 St. Andrews St.
- The home page is http://www.stanford.edu
- Yes, see you the following tues, that's 11/12/01
- IV: four, fourth, I.V.
- IRA: I.R.A. or Ira
- 1750: seventeen fifty (date, address) or one thousand seven… (dollars)

Text Normalization Steps

1. Identify tokens in text
2. Chunk tokens
3. Identify types of tokens
4. Convert tokens to words

Step 1+2: identify tokens and chunk

- Whitespace can be viewed as separators
- Punctuation can be separated from the raw tokens
- For example, **Festival** converts text into
  - ordered list of tokens
  - each with features:
    - its own preceding whitespace
    - its own succeeding punctuation

Important issue in tokenization: end-of-utterance detection

- Relatively simple if utterance ends in ?!
- But what about ambiguity of "."?
- Ambiguous between end-of-utterance, end-of-abbreviation, and both
  - My place on Main St. is around the corner.
  - I live at 123 Main St.
  - (Not "I live at 151 Main St..")

Steps 3+4: Identify Types of Tokens, and Convert Tokens to Words

- Pronunciation of numbers often depends on type. Three ways to pronounce 1776:

  Date: seventeen seventy six

  Phone number: one seven seven six

  Quantifier: one thousand seven hundred (and) seventy six

  - Also:
    - 25 **Day**: twenty-fifth

Rules/features for end-of-utterance detection

- A dot with one or two letters is an abbrev
- A dot with 3 cap letters is an abbrev.
- An abbrev followed by 2 spaces and a capital letter is an end-of-utterance
- Non-abbrevs followed by capitalized word are breaks
- Modern approaches use deep learning sequence transduction

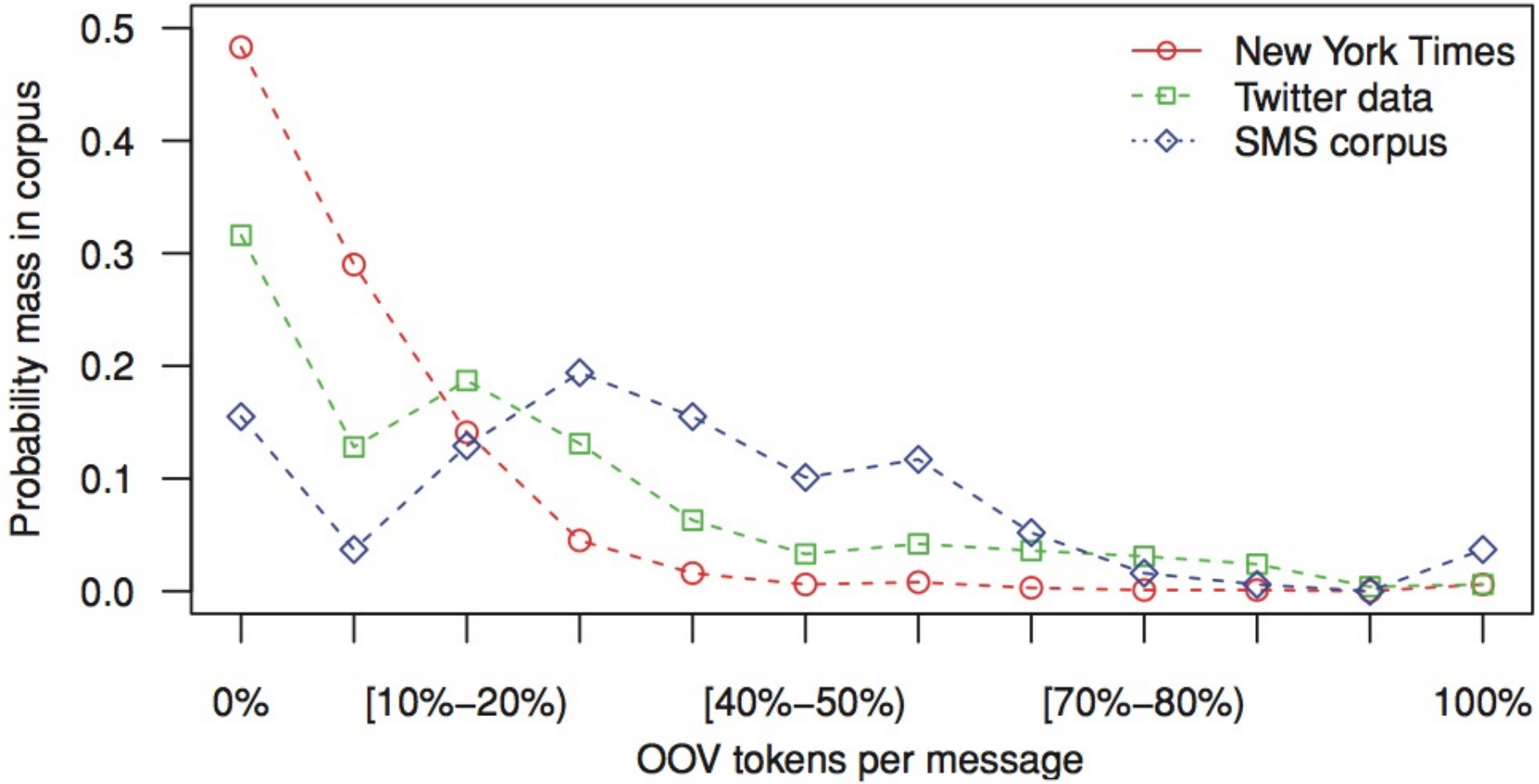
How common are non-standard words (NSWs)?

- Word not in lexicon, or with non-alphabetic characters (Sproat et al 2001, before SMS/Twitter)

Text Type	**% NSW**
novels	1.5%
press wire	4.9%
e-mail	10.7%
recipes	13.7%
classified	17.9%

How common are non-standard words (NSWs)?

Word not in gnu aspell dictionary (Han, Cook, Baldwin 2013) not counting @mentions, #hashtags, urls,



Twitter: 15% of tweets have 50% or more OOV

Homograph disambiguation

It's no use (/y uw s/) to ask to use (/y uw z/) the telephone.

Do you live (/l ih v/) near a zoo with live (/l ay v/) animals?

I prefer bass (/b ae s/) fishing to playing the bass (/b ey s/) guitar.

Final voicing		
	N (/s/)	**V** (/z/)
use	y uw s	y uw z
close	k l ow s	k l ow z
house	h aw s	h aw z

Stress shift		
	N (init. stress)	**V** (fin. stress)
record	r eh1 k axr0 d	r ix0 k ao1 r d
insult	ih1 n s ax0 l t	ix0 n s ah1 l t
object	aa1 b j eh0 k t	ax0 b j eh1 k t

-ate final vowel		
	N/A (final /ax/)	**V** (final /ey/)
estimate	eh s t ih m ax t	eh s t ih m ey t
separate	s eh p ax r ax t	s eh p ax r ey t
moderate	m aa d ax r ax t	m aa d ax r ey t

Letter to Sound Rules

- AKA Grapheme to Phoneme (G2P)
- Generally machine learning, induced from a dictionary
- Pick your favorite machine learning tool and go for it
- Earlier work: (Black et al. 1998)
  - Two steps: alignment and (CART-based) rule-induction
- Modern seq2seq approaches might handle tokenization and G2P jointly as a single text normalization module

Outline

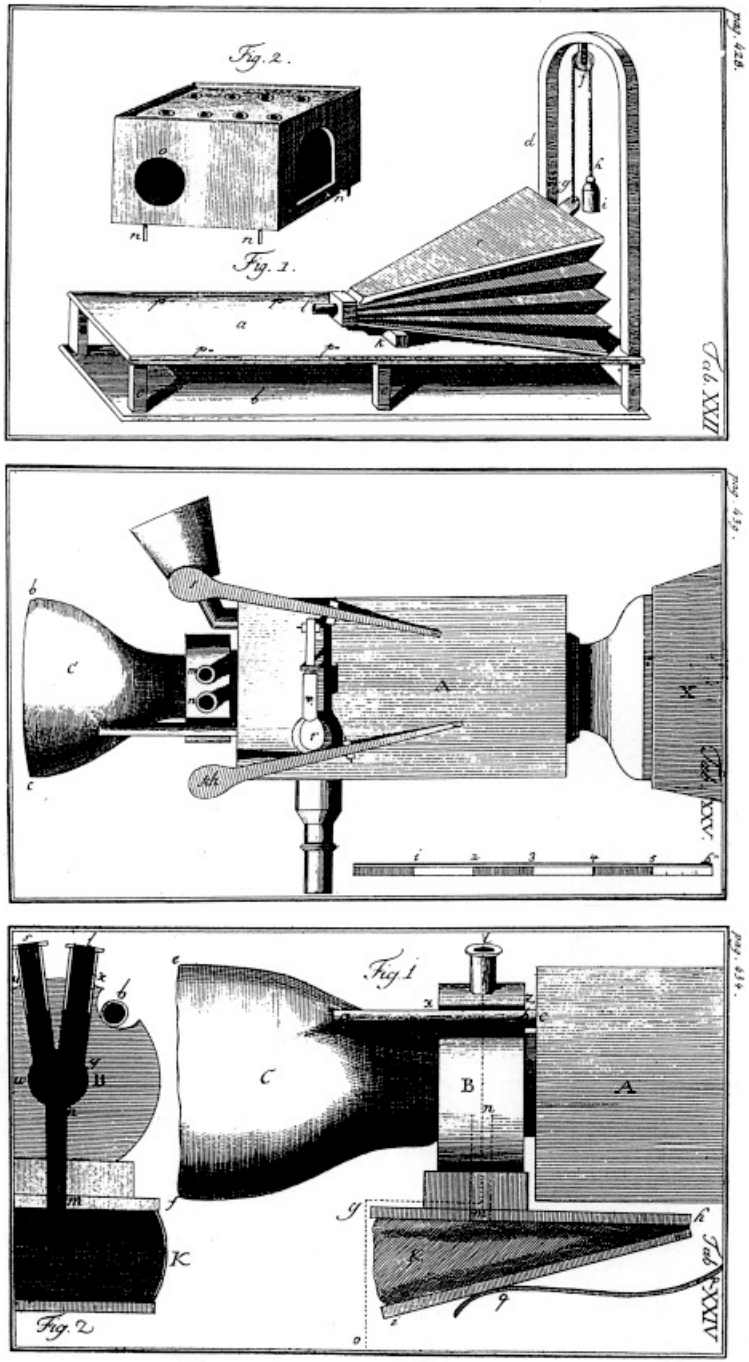
- Text analysis
  - Text normalization
  - Letter-to-sound (grapheme-to-phoneme)
- TTS modeling history and overview
- Prosody and intonation

History of TTS

- Pictures and some text from Hartmut Traunmüller's web site:
  http://www.ling.su.se/staff/hartmut/kemplne.htm
- Von Kempeln 1780 b. Bratislava 1734 d. Vienna 1804
- Leather resonator manipulated by the operator to try and copy vocal tract configuration during sonorants (vowels, glides, nasals)
- Bellows provided air stream, counterweight provided inhalation
- Vibrating reed produced periodic pressure wave

Von Kempelen:

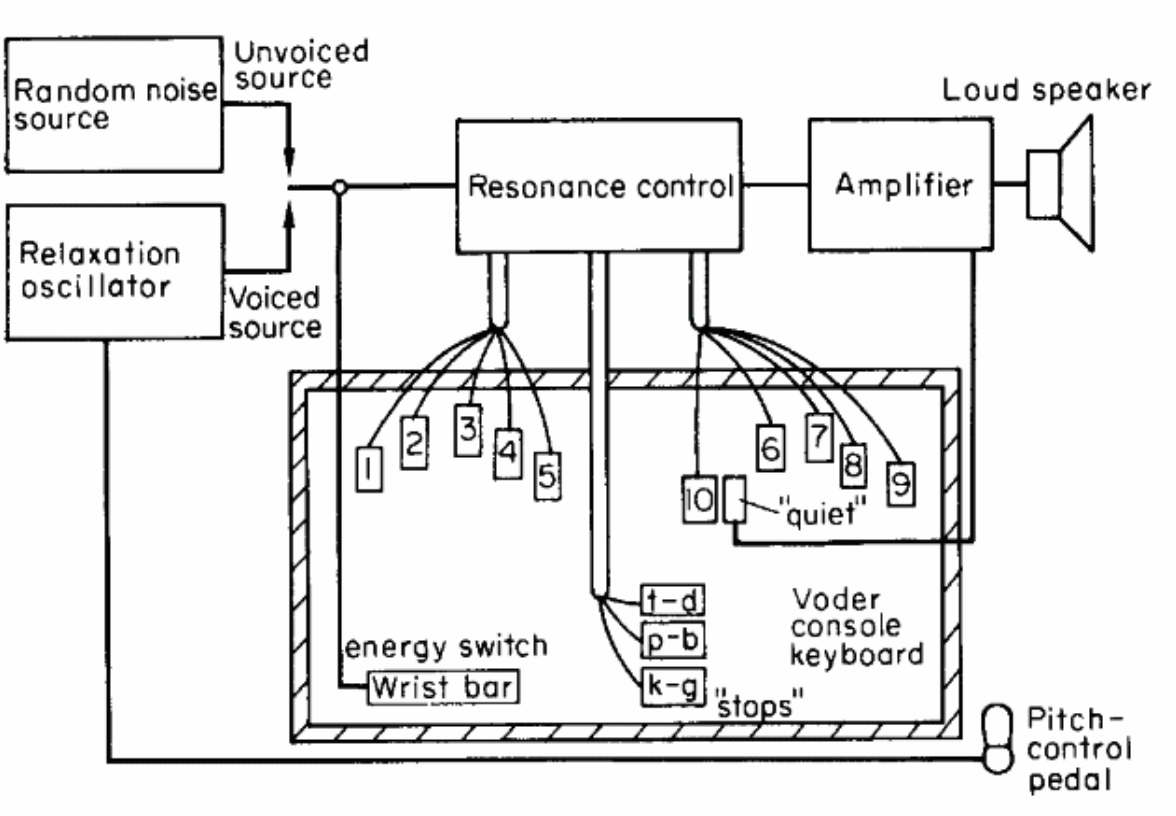
- Small whistles controlled consonants
- Rubber mouth and nose; nose had to be covered with two fingers for non-nasals
- Unvoiced sounds: mouth covered, auxiliary bellows driven by string provides puff of air



From Traunmüller's web site

Homer Dudley 1939 VODER

- Synthesizing speech by electrical means
- 1939 World's Fair



Homer Dudley's VODER

- Manually controlled through complex keyboard
- Operator training was a problem

An aside on demos

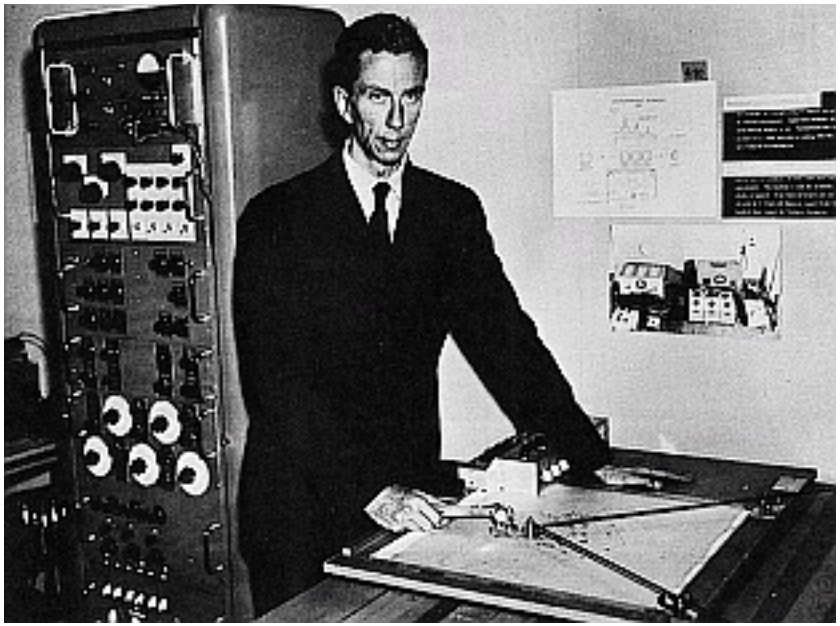
That last slide exhibited:

Rule 1 of playing a speech synthesis demo:

Always have a human say what the words are right before you have the system say them

Gunnar Fant's OVE synthesizer

- Of the Royal Institute of Technology, Stockholm
- Formant Synthesizer for vowels
- F1 and F2 could be controlled

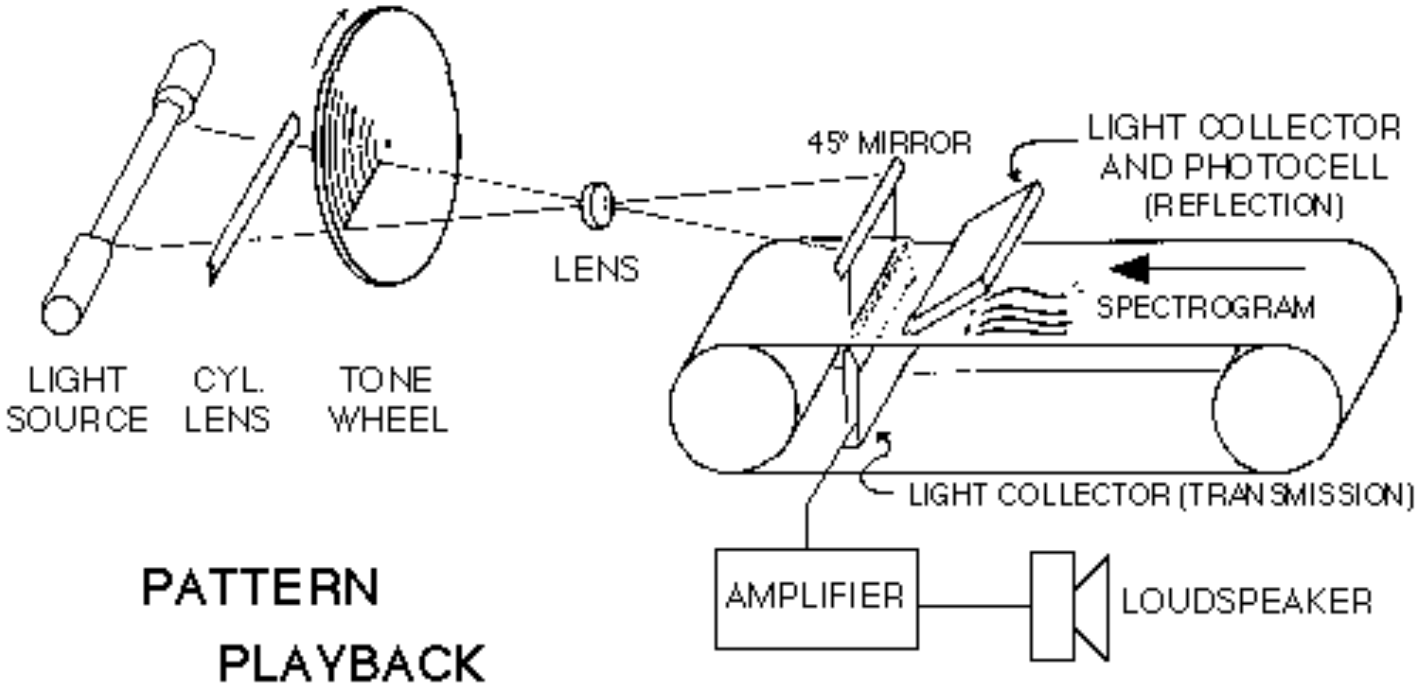


From Traunmüller's web site

Cooper's Pattern Playback

- Haskins Labs for investigating speech perception
- Works like an inverse of a spectrograph
- Light from a lamp goes through a rotating disk then through spectrogram into photovoltaic cells
- Thus amount of light that gets transmitted at each frequency band corresponds to amount of acoustic energy at that band

Cooper's Pattern Playback



Pre-modern TTS systems

- 1960's first full TTS: Umeda et al (1968)
- 1970's
  - Joe Olive 1977 concatenation of linear-prediction diphones
  - Texas Instruments Speak and Spell,
    - June 1978
    - Paul Breedlove

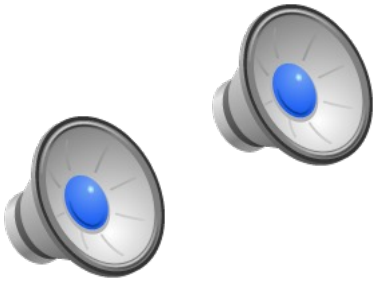
Types of Synthesis

- 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.
    - Diphone
    - Unit Selection
- Parametric Synthesis
  - Learn parameters on databases of speech

Text modified from Richard Sproat slides

1980s: Formant Synthesis

- Were the most common commercial systems when computers were slow and had little memory.
- 1979 MIT MITalk (Allen, Hunnicut, Klatt)
- 1983 DECtalk system based on Klatttalk
  - “Perfect Paul” (The voice of Stephen Hawking)
  - “Beautiful Betty”



1990s: 2nd Generation Synthesis 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
- ~1400 recordings
- Paste them together and modify prosody.

Modern Unit Selection Systems

- Most current commercial systems.
- Larger units of variable length
- Record one speaker speaking 10 hours or more,
  - Have multiple copies of each unit
- Use search to find best sequence of units

Parametric Synthesis

- Train a statistical model on large amounts of data. Learn text → sound mapping.
- Very active area of research
- Previously associated with HMM Synthesis
- Deep learning approaches (Wavenet, Tacotron) are the latest generation of parametric
  - Quickly becoming adopted in industry due to improved naturalness and compact models

Outline

- Text analysis
  - Text normalization
  - Letter-to-sound (grapheme-to-phoneme)
- TTS modeling history and overview
- **Prosody and intonation**

Prosody and Intonation in TTS

Prominence/Accent: Decide which words are accented, which syllable has accent, what sort of accent

Boundaries: Decide where intonational boundaries are

Duration: Specify length of each segment

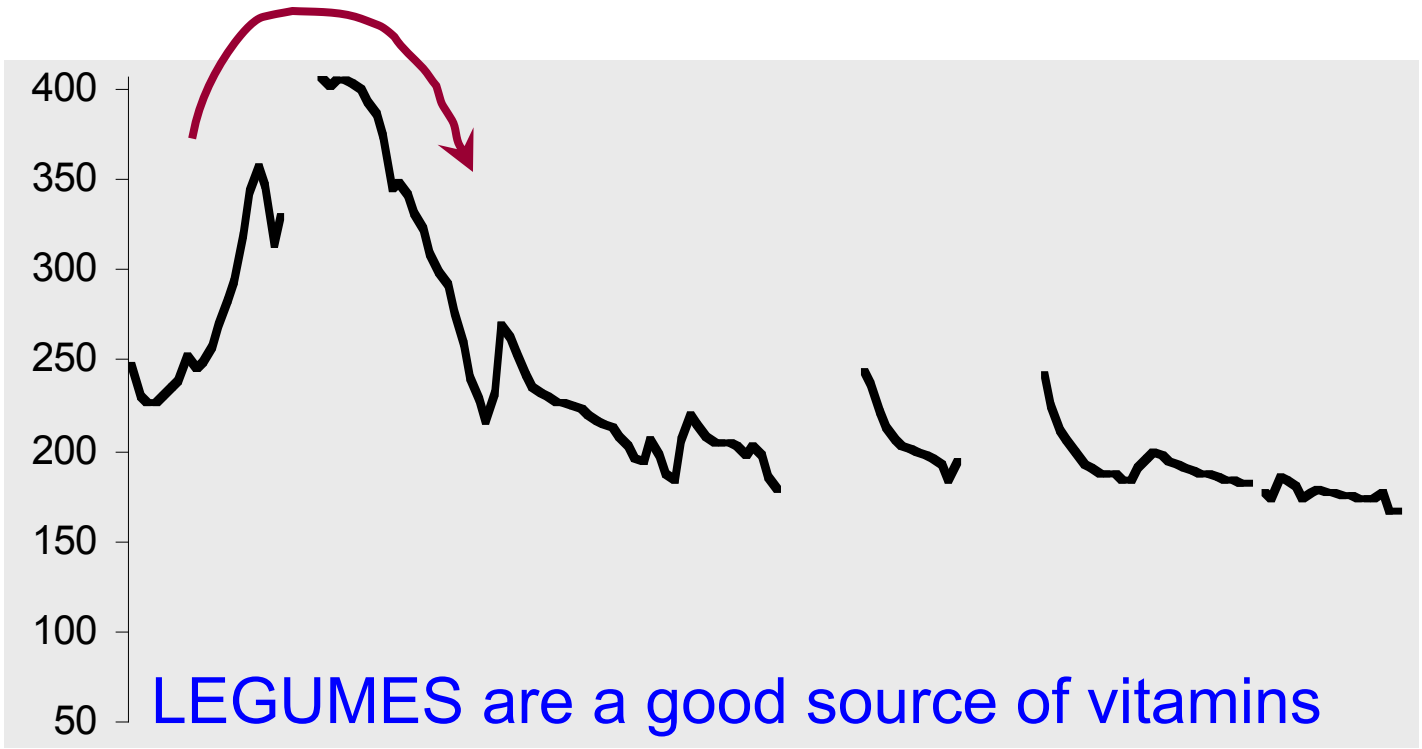
F0: Generate F0 contour from these

Stress vs. accent

- **Stress**: structural property of a word
  - Fixed in the lexicon: marks a potential (arbitrary) location for an accent to occur, if there is one
- **Accent**: property of a word in context
  - Context-dependent. Marks 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	

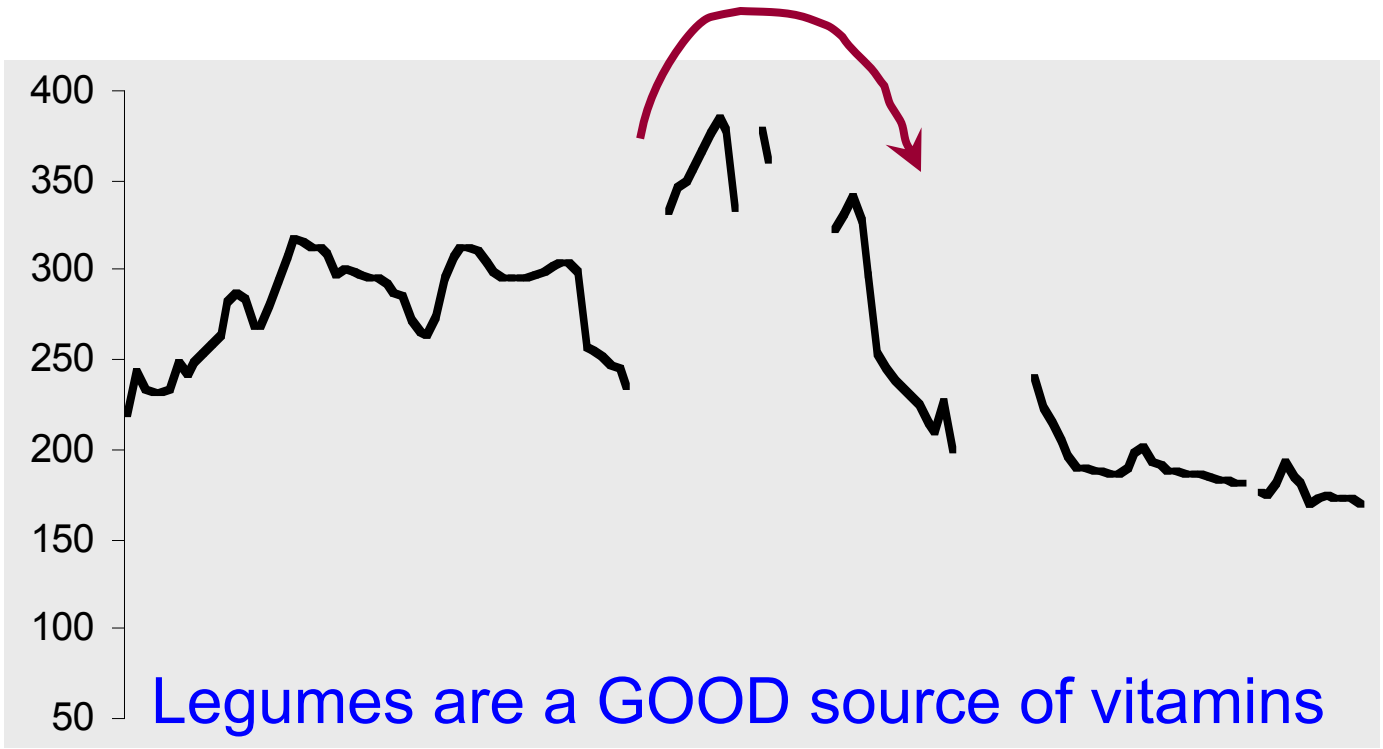
Same ‘tune’, different alignment



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

Slide from Jennifer Venditti

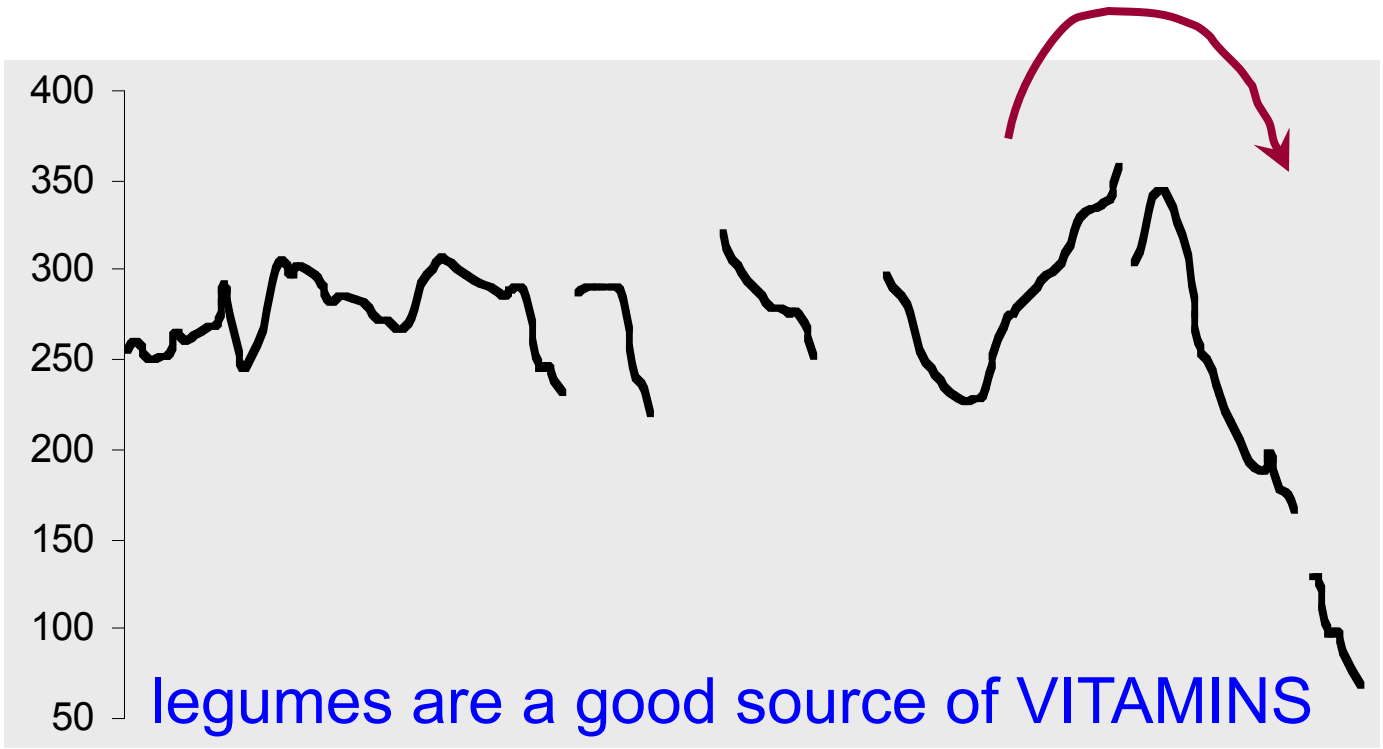
Same ‘tune’, different alignment



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

Levels of prominence

- Most phrases have more than one accent
- **Nuclear Accent:** Last accent in a phrase, perceived as more prominent
  - Plays semantic role like indicating a word is contrastive or focus.
  - Modeled via ***s in IM, or capitalized letters
  - 'I know **SOMETHING** interesting is sure to happen,' she said
- Can also have reduced words that are **less** prominent than usual (especially function words)
- Sometimes use 4 classes of prominence:
  - **Emphatic accent, pitch accent, unaccented, reduced**

Predicting Boundaries: Full || versus intermediate |

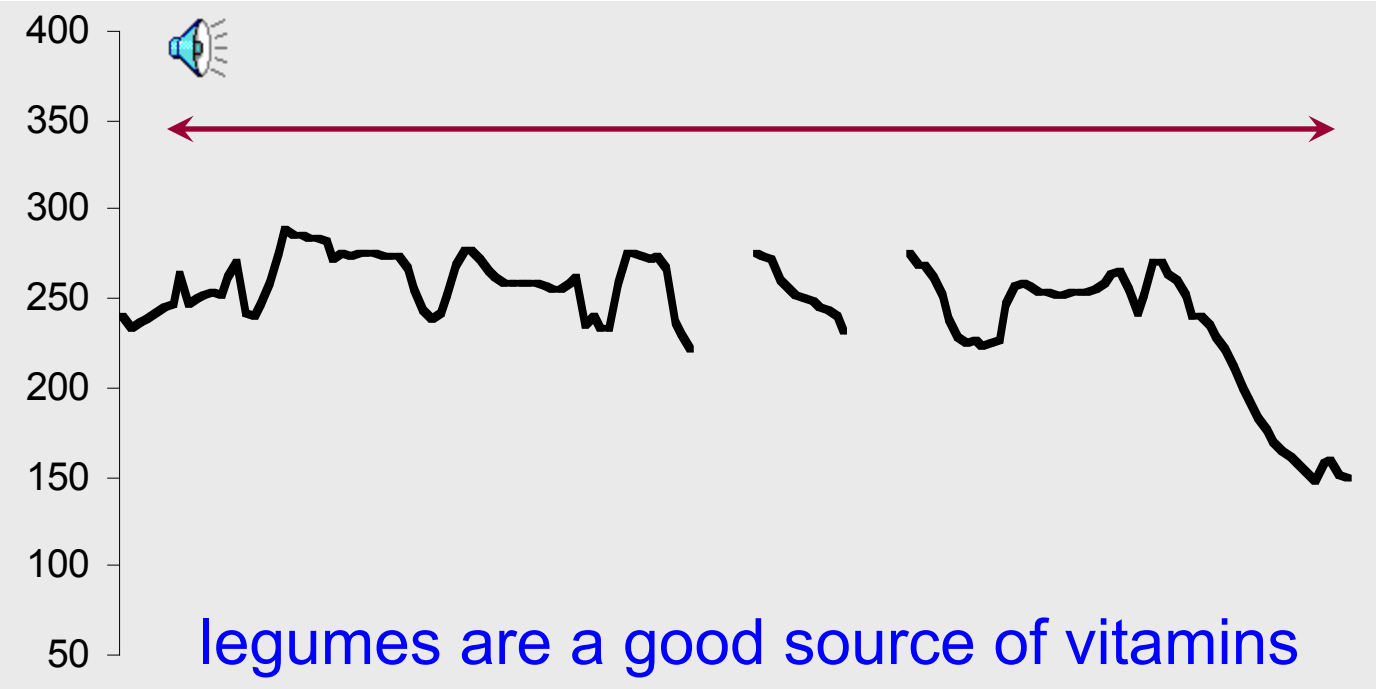
Ostendorf and Veilleux. 1994 “Hierarchical Stochastic model for Automatic Prediction of Prosodic Boundary Location”, Computational Linguistics 20:1

Computer phone calls, || which do everything | from selling magazine subscriptions || to reminding people about meetings || have become the telephone equivalent | of junk mail. ||

Doctor Norman Rosenblatt, || dean of the college | of criminal justice at Northeastern University, || agrees.||

For WBUR, || I’m Margo Melnicove.

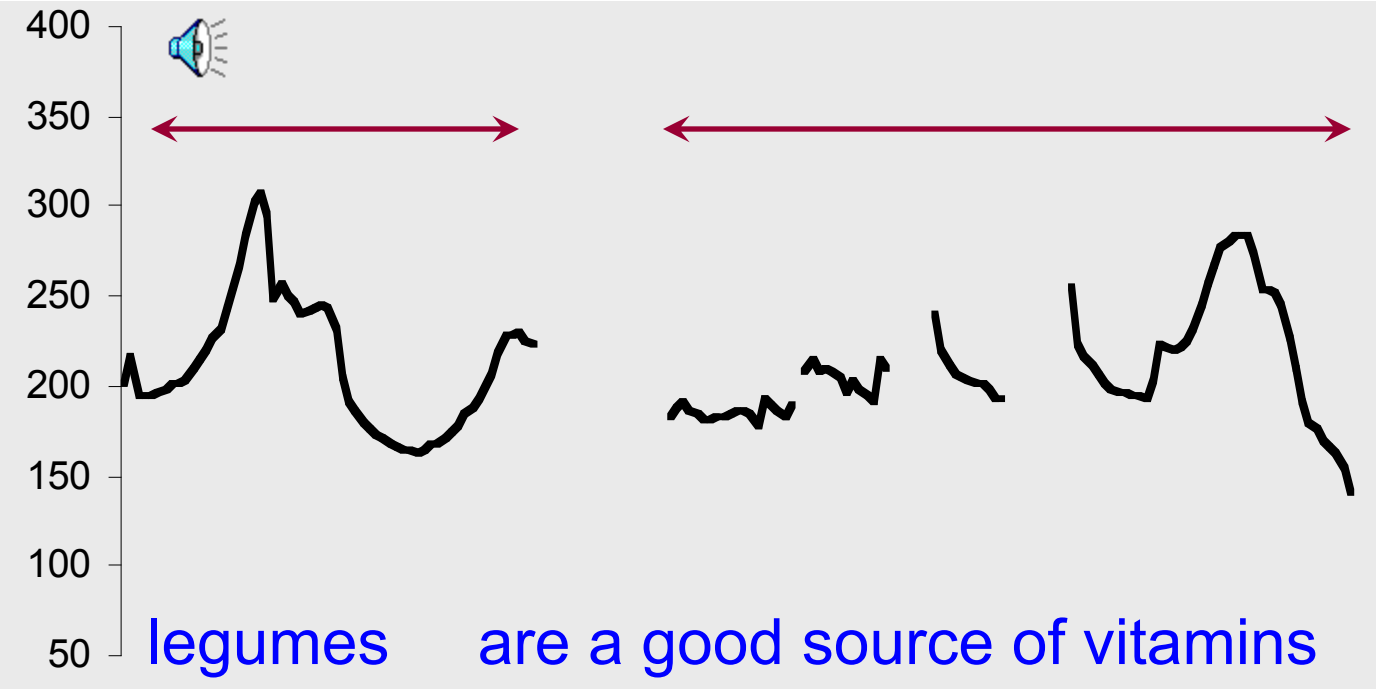
A single intonation phrase



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

Slide from Jennifer Venditti

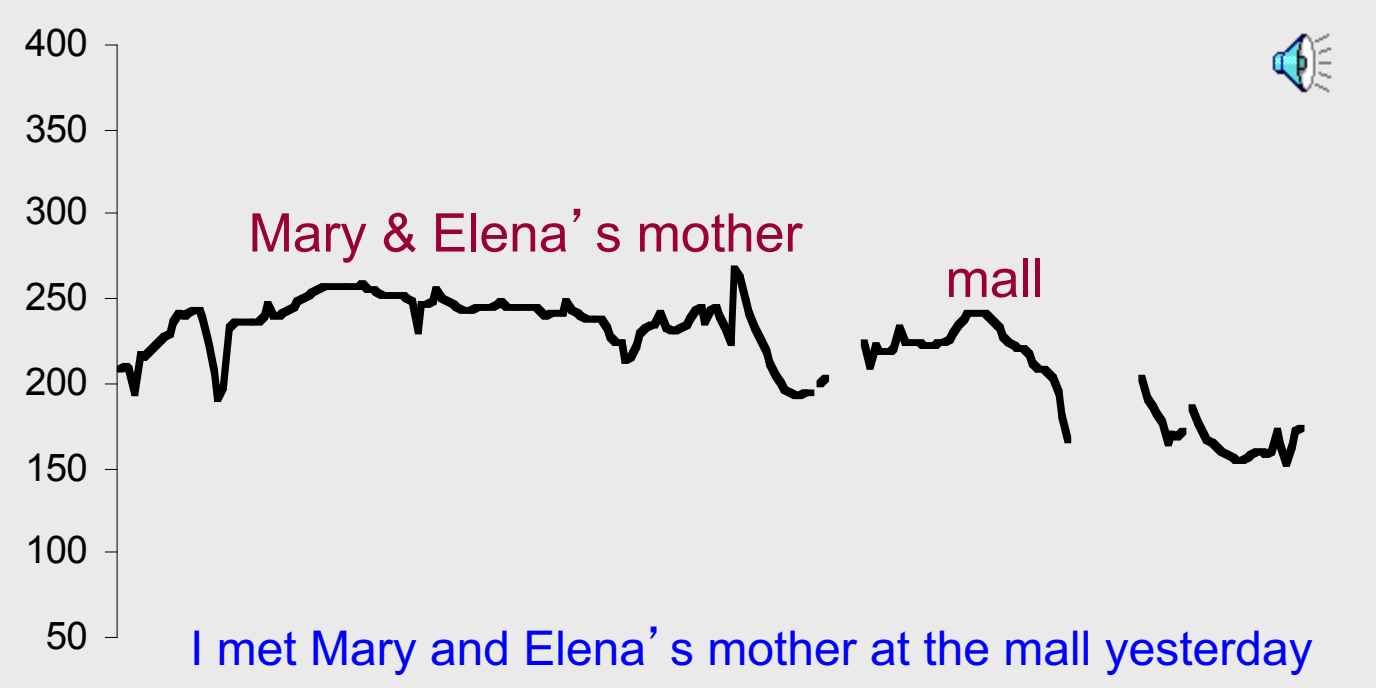
Multiple phrases



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

Slide from Jennifer Venditti

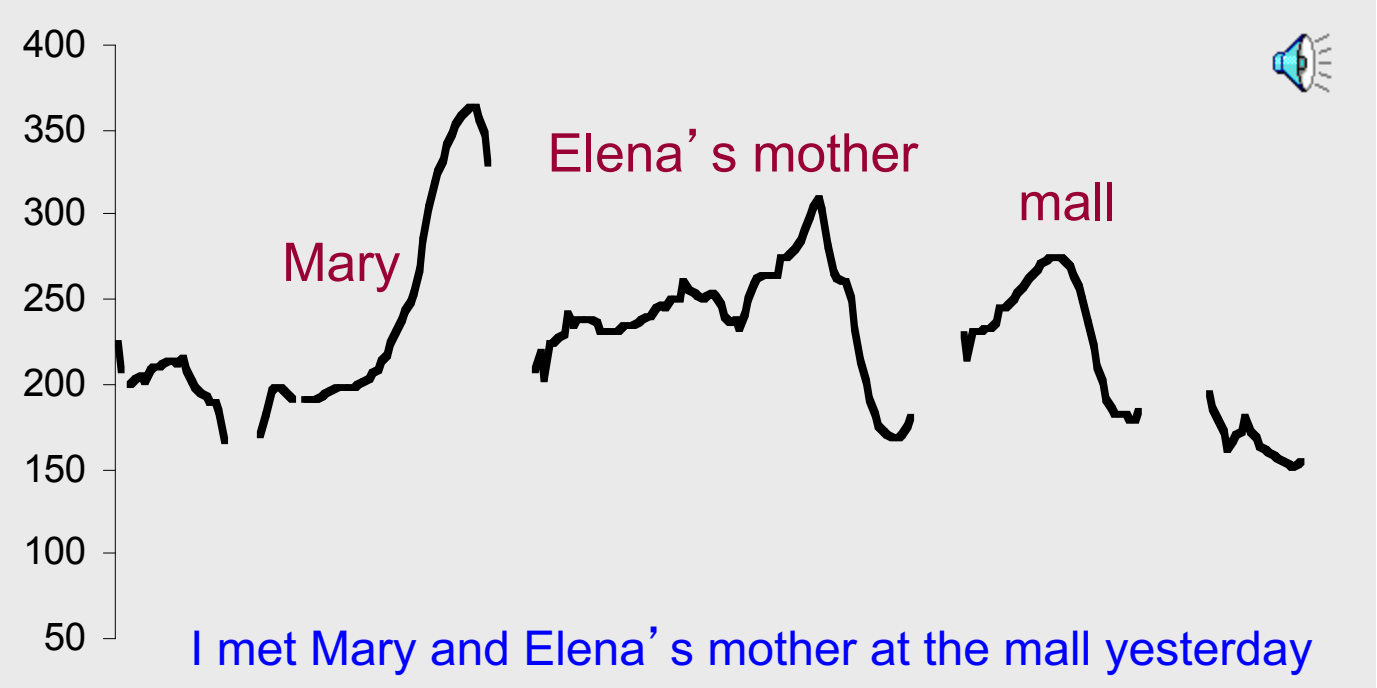
Phrasing sometimes helps disambiguate



One intonation phrase with relatively flat overall pitch range.

Slide from Jennifer Venditti

Phrasing sometimes helps disambiguate



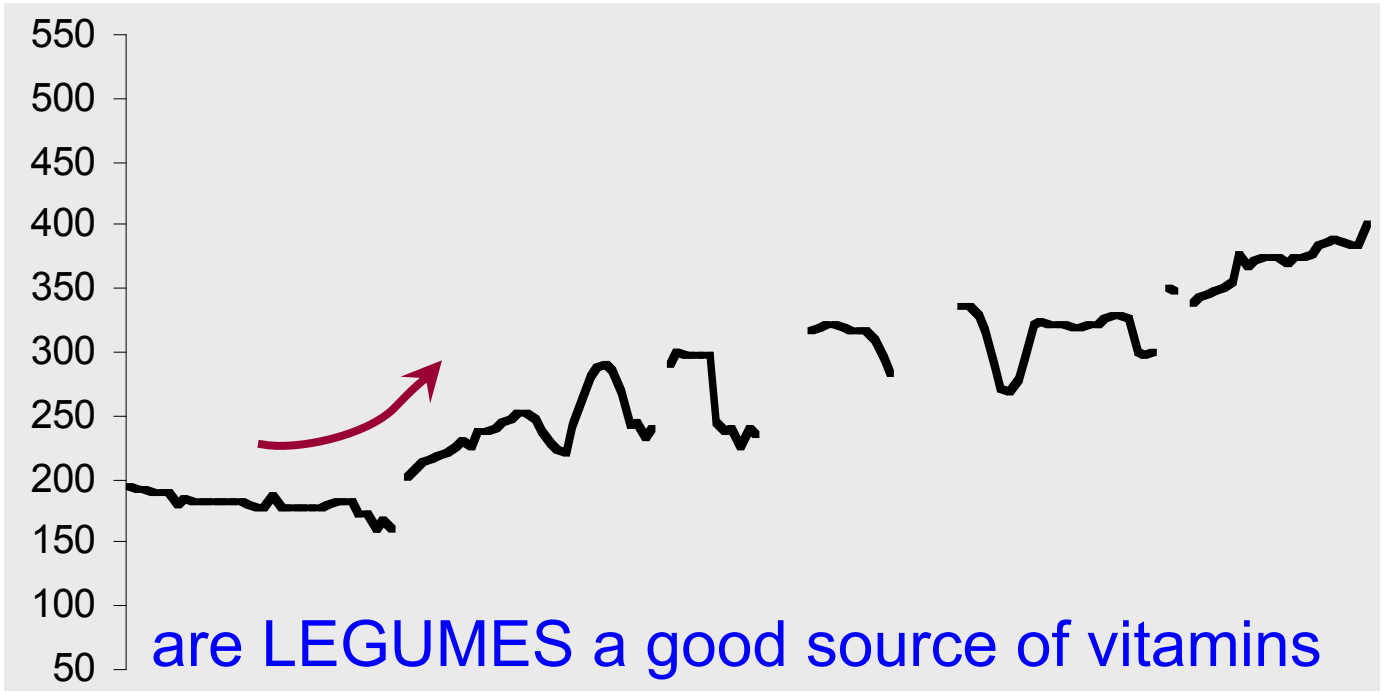
Separate phrases, with expanded pitch movements.

Slide from Jennifer Venditti

More prosodic tunes

Slide from Jennifer Venditti

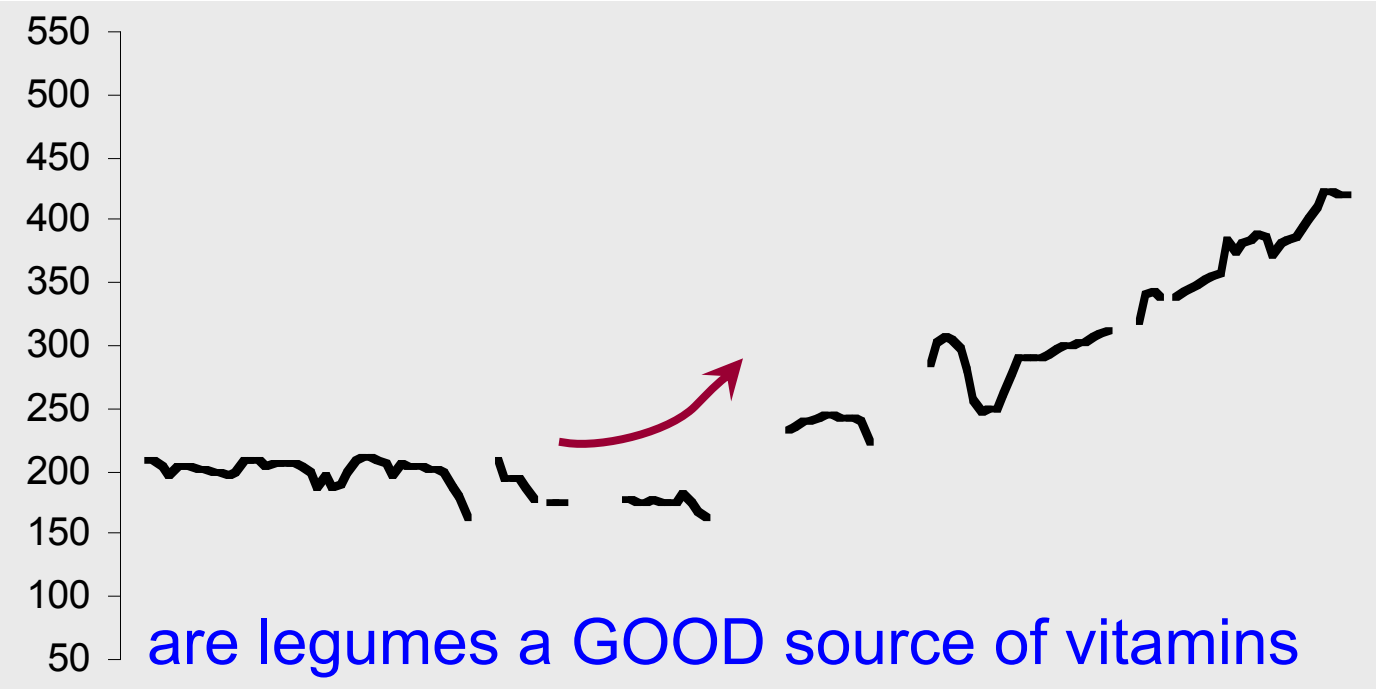
Yes-No question tune



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

Slide from Jennifer Venditti

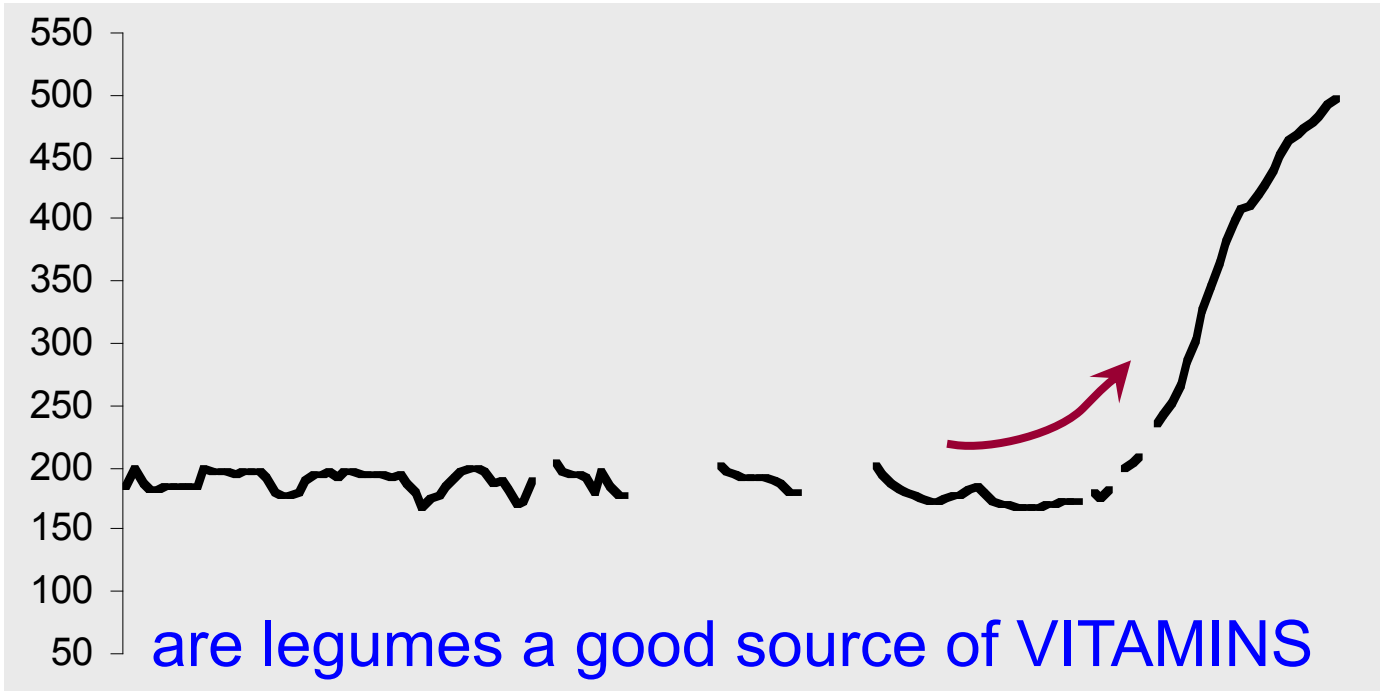
Yes-No question tune



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

Slide from Jennifer Venditti

Yes-No question tune

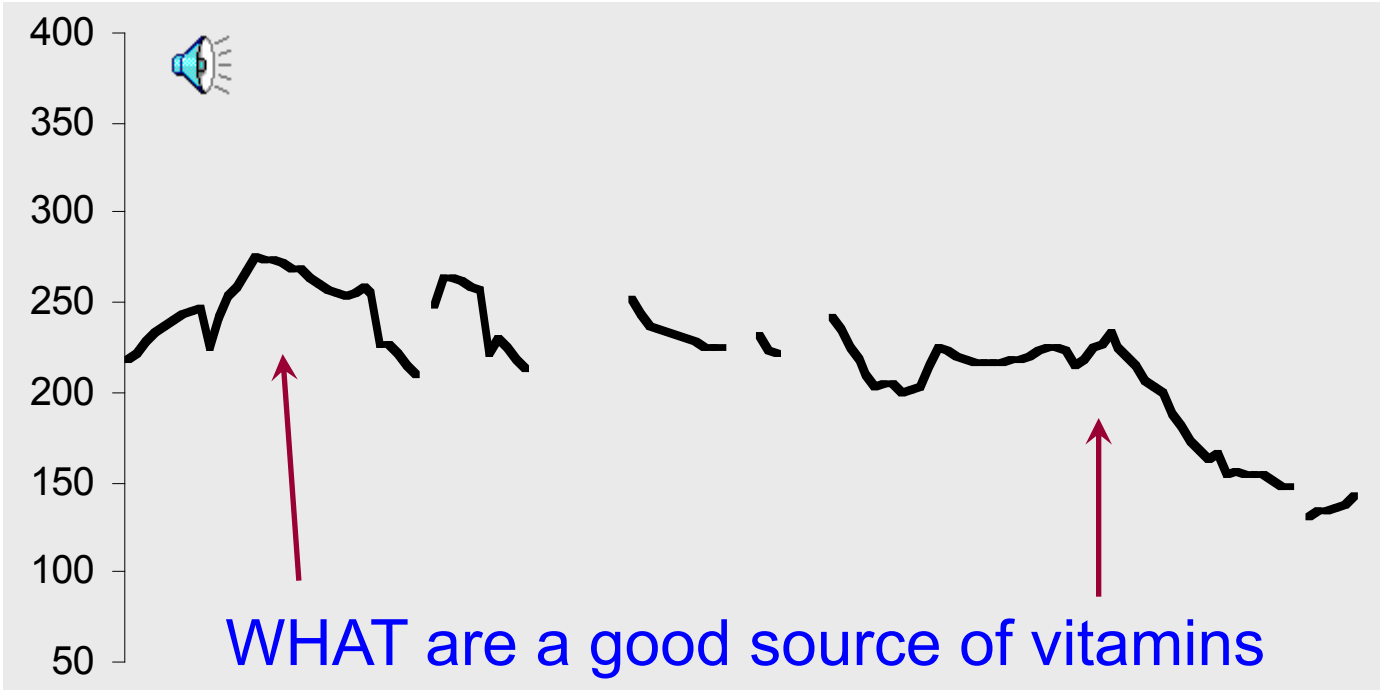


Rise from the main accent to the end of the sentence.

Slide from Jennifer Venditti

WH-questions

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

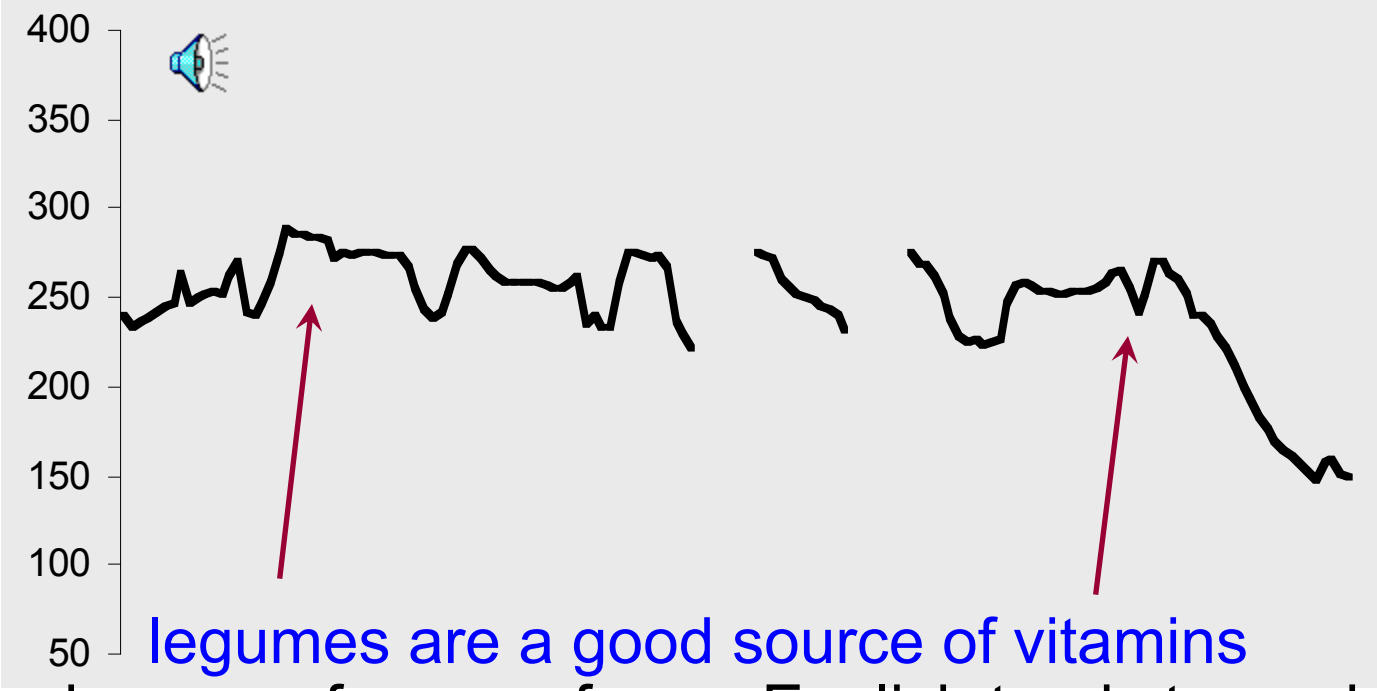


WH-questions typically have falling contours, like statements.

Slide from Jennifer Venditti

Broad focus

“Tell me something about the world.”

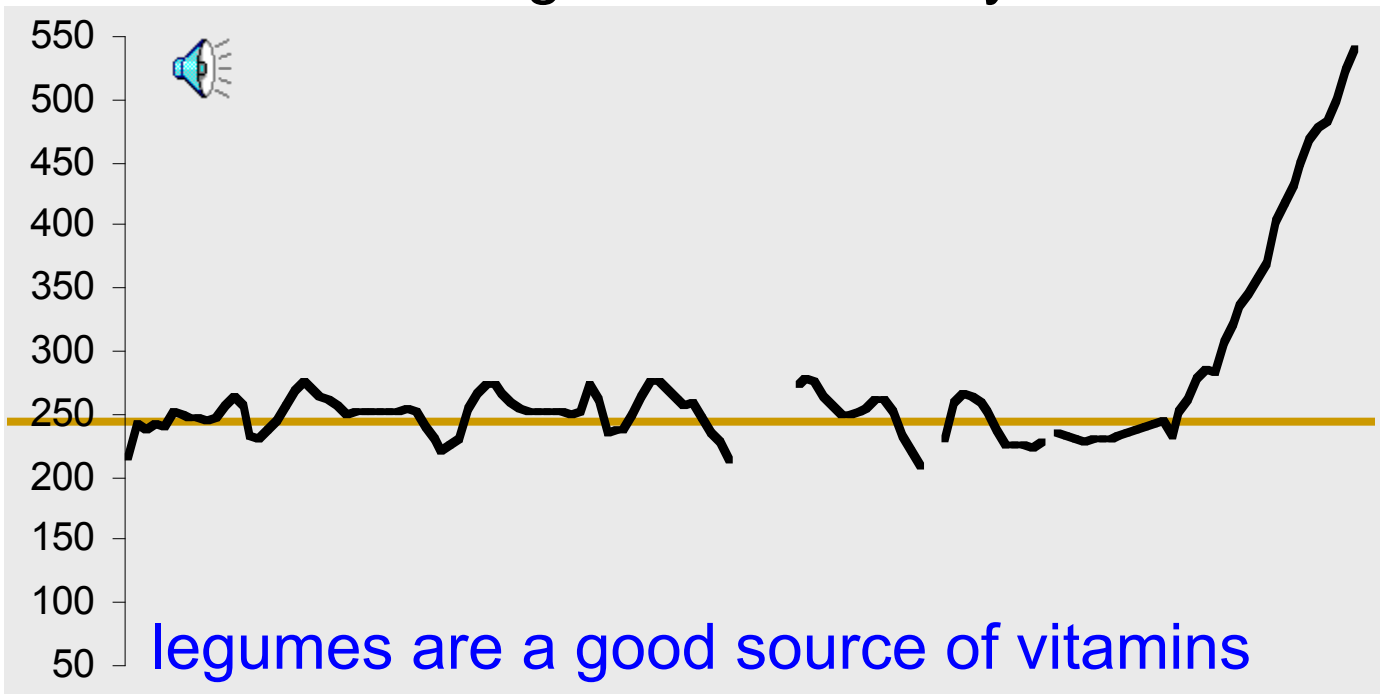


In the absence of narrow focus, English tends to mark the first and last ‘content’ words with perceptually prominent accents.

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

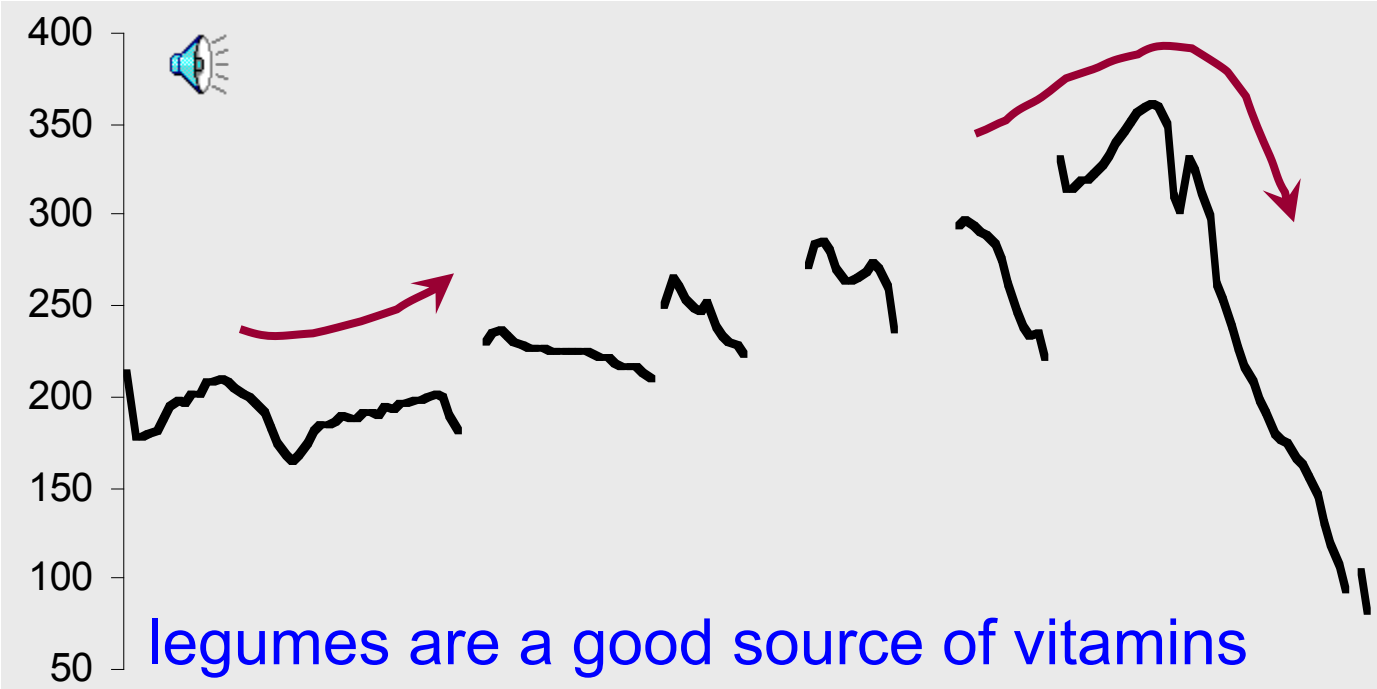
Yes-No question

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

Slide from Jennifer Venditti

‘Surprise-redundancy’ tune

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

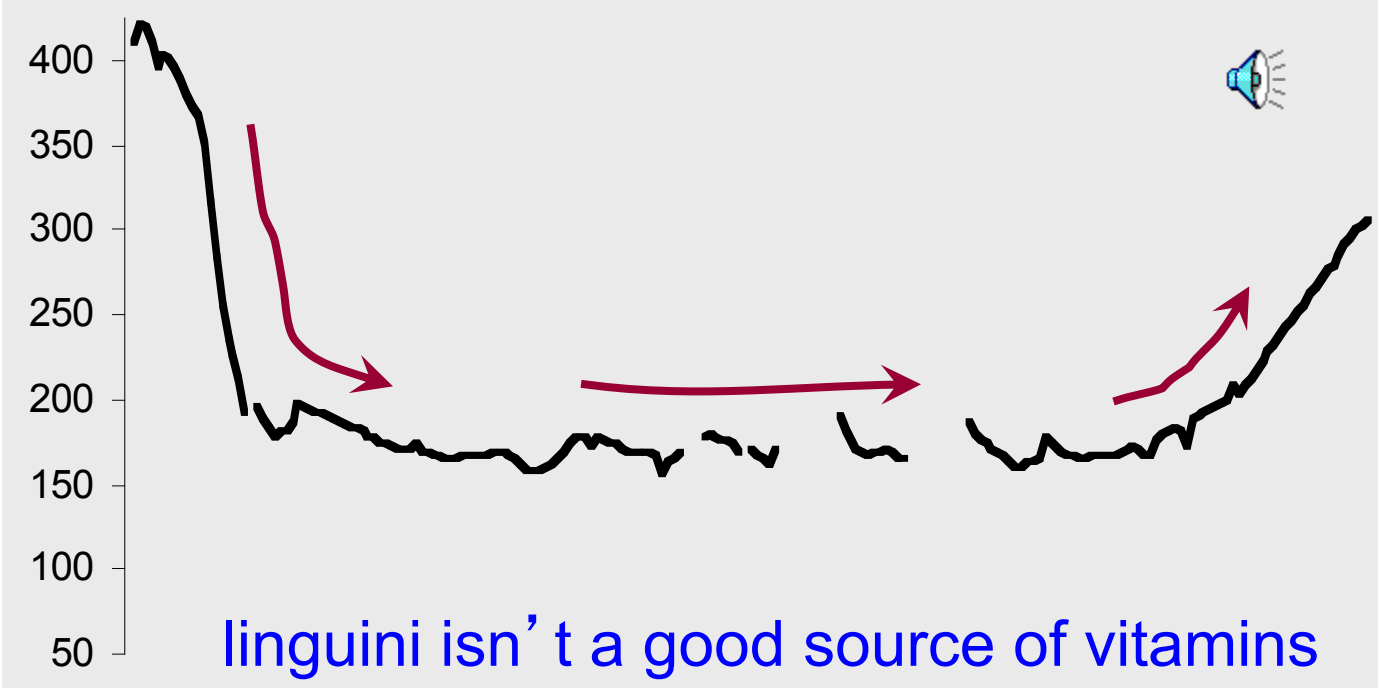


Low beginning followed by a gradual rise to a high at the end.

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

Appendix: Synthesis tools

Synthesis tools (older)

- 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 Sproat

Festival

- Open source speech synthesis system
- Designed for development and runtime use
  - Use in many commercial and academic systems
  - 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

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

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

Text from Richard Sproat

Dictionaries

- CMU dictionary: 127K words
  - http://www.speech.cs.cmu.edu/cgi-bin/cmudict
- Unisyn dictionary
  - Significantly more accurate, includes multiple dialects
  - http://www.cstr.ed.ac.uk/projects/unisyn/

```
going:        { g * ou }.> i ng >
antecedents:    { * a n . t^ i . s ~ ii . d n! t }> s >
dictionary:     { d * i k . sh @ . n ~ e . r ii }
```

Dictionaries aren't always sufficient: Unknown words

- Go up with square root of # of words in text
- Mostly person, company, product names
  - From a Black et al analysis
    - 5% of tokens in a sample of WSJ not in the OALD dictionary

      77%: Names

      20% Other Unknown Words

      4%: Typos
- So commercial systems have 3-part system:
  - Big dictionary
  - Special code for handling names
  - Machine learned LTS system for other unknown words

Names

- Big problem area is names
- Names are common
  - 20% of tokens in typical newswire text
  - Spiegel (2003) estimate of US names:
    - 2 million surnames
    - 100,000 first 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

Methods for Names

- 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

Homograph disambiguation

- 19 most frequent homographs, from Liberman and Church 1992
- Counts are per million, from an AP news corpus of 44 million words
- Not a huge problem, but still important

use	319
increase	230
close	215
record	195
house	150
contract	143
lead	131
live	130
lives	105
protest	94
survey	91
project	90
separate	87
present	80
read	72
subject	68
rebel	48
finance	46
estimate	46

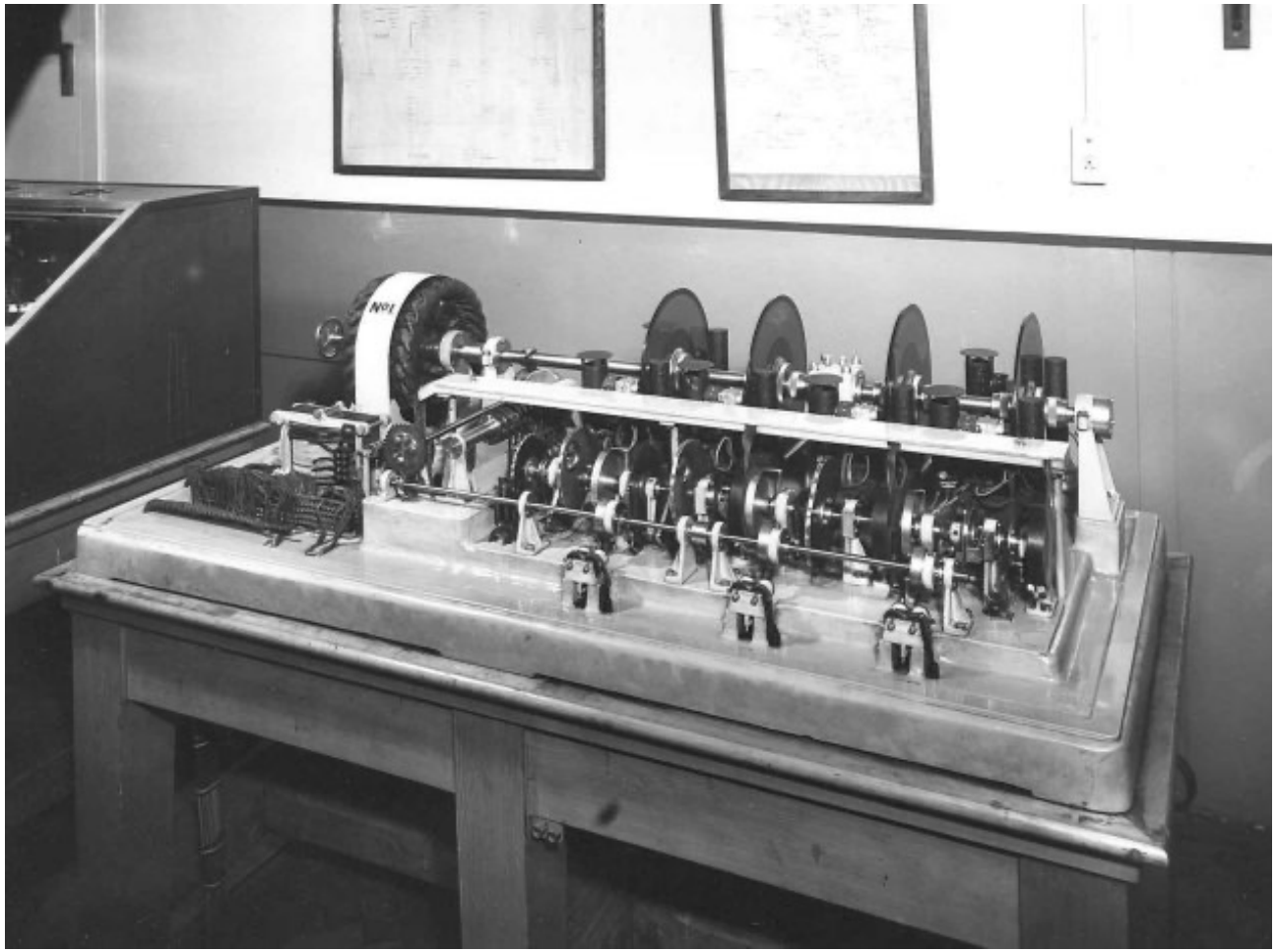
POS Tagging for homograph disambiguation

- Many homographs can be distinguished by POS

use	y uw s	y uw z
close	k l ow s	k l ow z
house	h aw s	h aw z
live	l ay v	l ih v
REcord	reCORD	
INsult	inSULT	
OBject	objECT	
OVERflow	overFLOW	
DIScount	disCOUNT	
CONtent	conTENT	

- POS tagging also useful for CONTENT/FUNCTION distinction, which is useful for phrasing

The 1936 UK Speaking Clock



From http://web.ukonline.co.uk/freshwater/clocks/spkgclock.htm

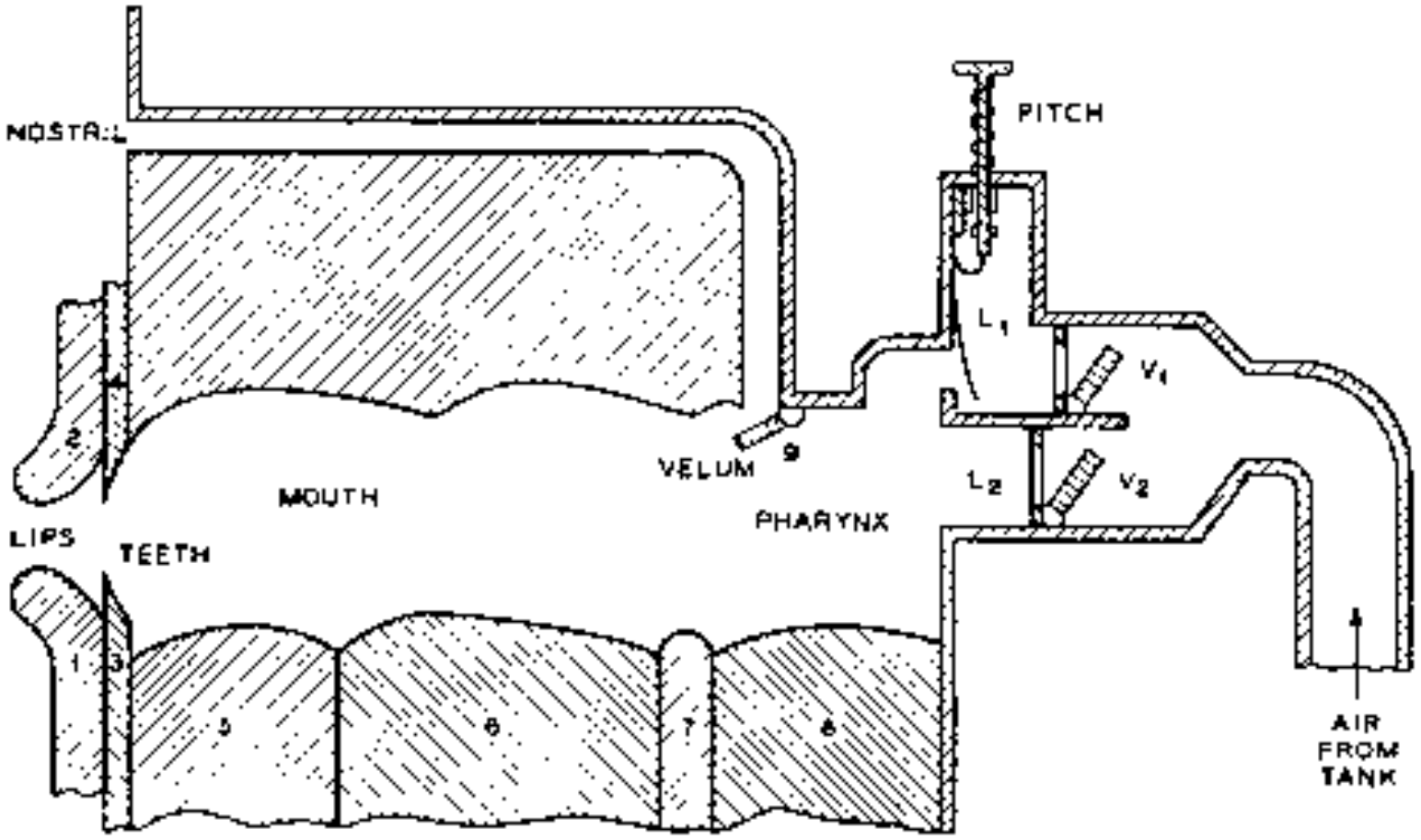
The UK Speaking Clock

- July 24, 1936
- Photographic storage on 4 glass disks
- 2 disks for minutes, 1 for hour, one for seconds.
- Other words in sentence distributed across 4 disks, so all 4 used at once.
- Voice of “Miss J. Cain”

A technician adjusts the amplifiers of the first speaking clock

From http://web.ukonline.co.uk/freshwater/clocks/spkgclock.htm

Closer to a natural vocal tract: Riesz 1937



Some modern examples

- Sample 1

- Sample 2

- Sample 3

- Sample 4