CS 224S / Linguist 285
Spoken Language Processing

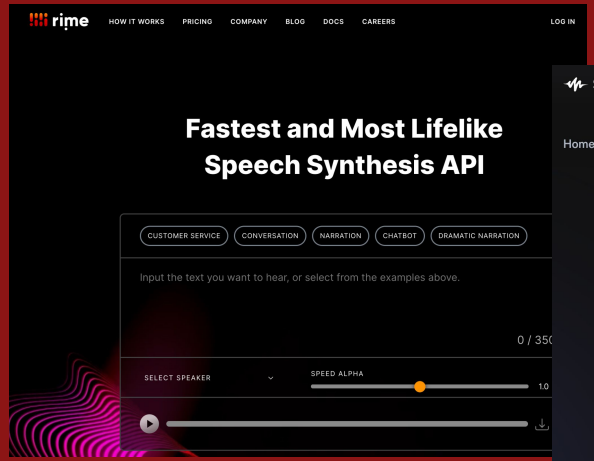
Andrew Maas | Stanford University | Spring 2024

Lecture 3: Text-to-Speech (TTS) Overview

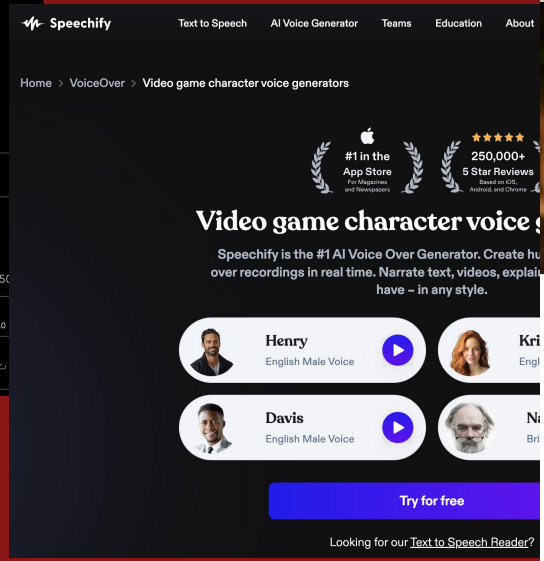
Outline

- Modern speech synthesis (= TTS). State of the art + challenges
- Prosody and Intonation
- Early speech synthesis systems
- TTS Overview: The basic modules
  - Text Analysis
    - Text Normalization
    - Letter-to-sound (Grapheme-to-phoneme)
  - (Next class) Waveform synthesis

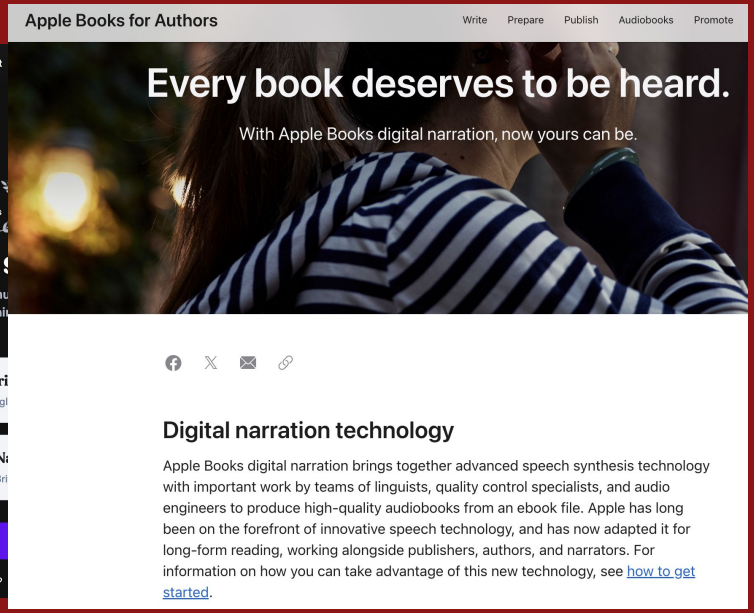
Modern TTS Systems are good!



Rime.ai



Speechify.com



https://authors.apple.com/support/4519-digital-narration-audiobooks

Many cloud provider APIs and public models

Where can modern systems improve?

- Emotional expression
- Inferring expression from text
- Singing, accents, and controllability
- Realtime factor and low-latency for spoken dialog systems
- Hardware constraints, cloud vs on-device synthesis
- The dream:
  Large range of voices and emotional expression, controllable voice types, works on any text, runs on smartphone without internet

Prosody and Intonation

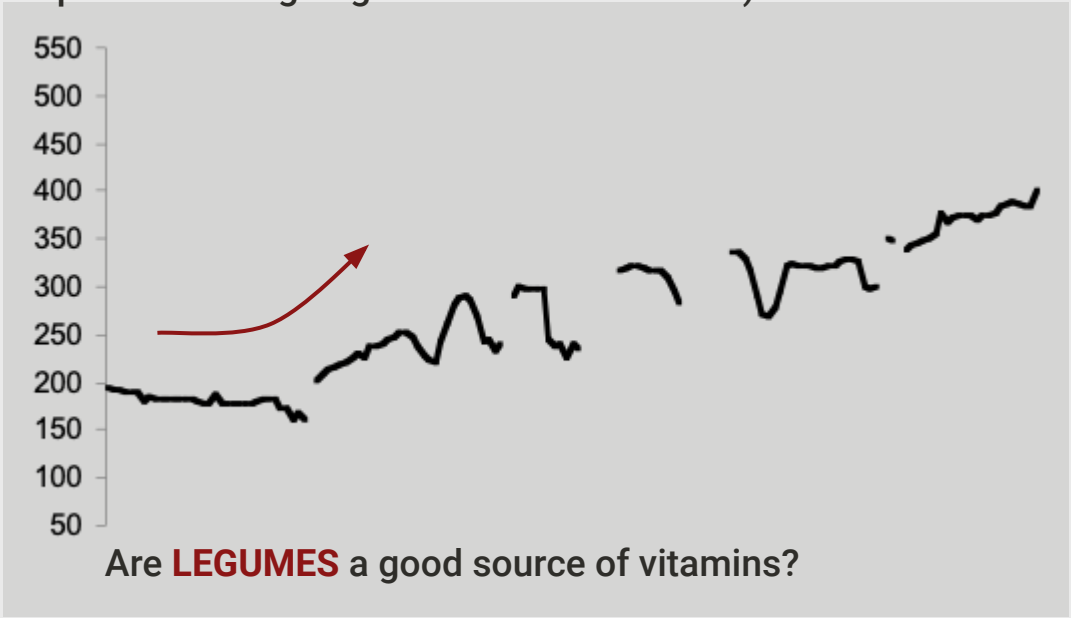
Prosody and Intonation in TTS

- **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

Yes-No Question Tune

- **Rise** from the main accent to the end of the sentence

(This is a pitch track. A plot of the “signing tune” of the utterance)



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

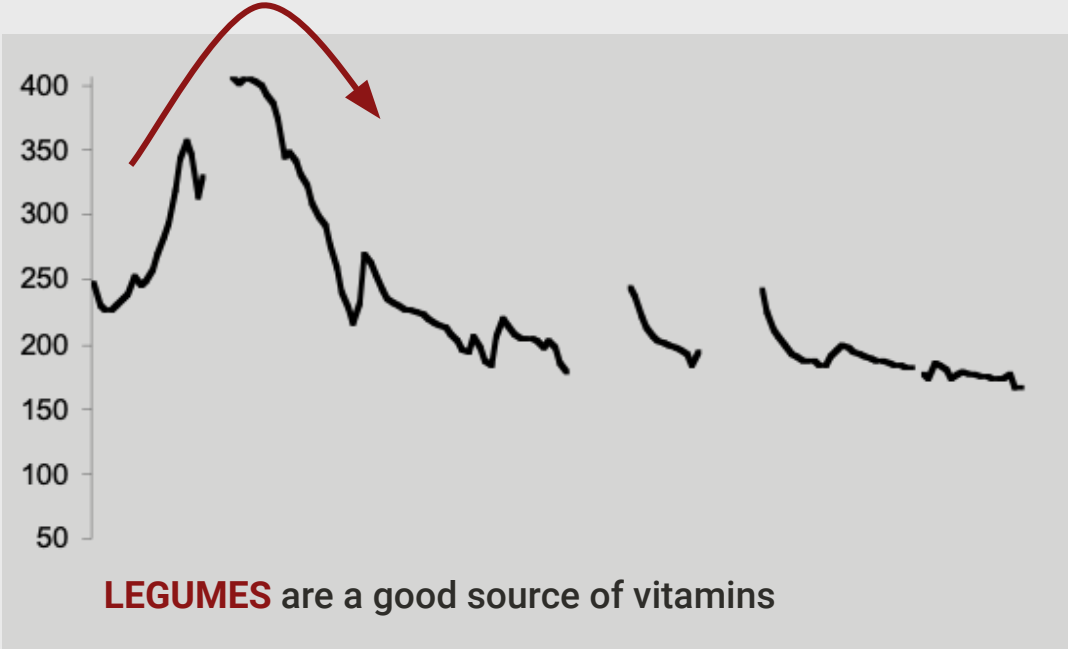
Stress vs. Accent

- **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	x	syllables
vi	ta	mins	Ca	li		for	nia	

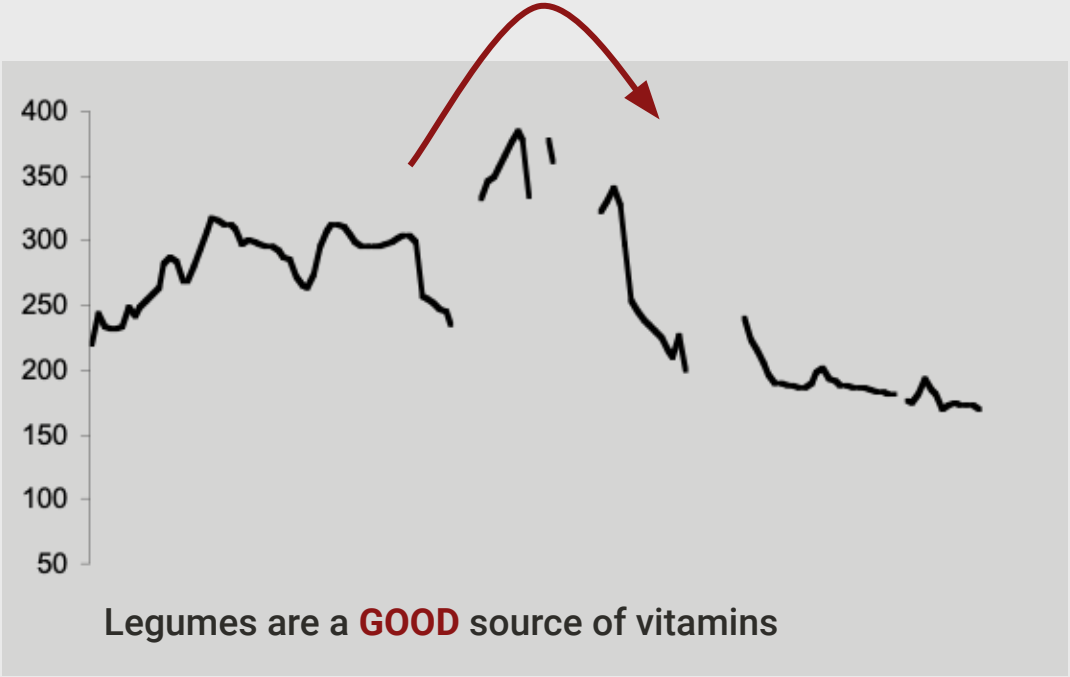
Same ‘Tune’, Different Alignment

- 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
- **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)
- Design choice: How to model prominence. Sometimes use 4 classes:
  - 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.||

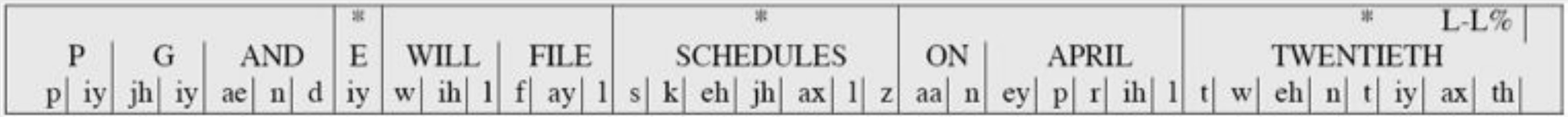
For WBUR, || I’m Margo Melnicove.

Overview of the speech synthesis task

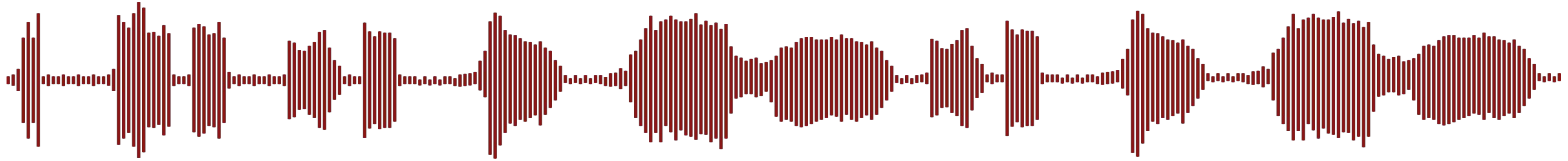
Synthesis in two stages

```
PG&E will file schedules on April 20th
```

1. **Text Analysis:** Text into some intermediate representation. *What to say*



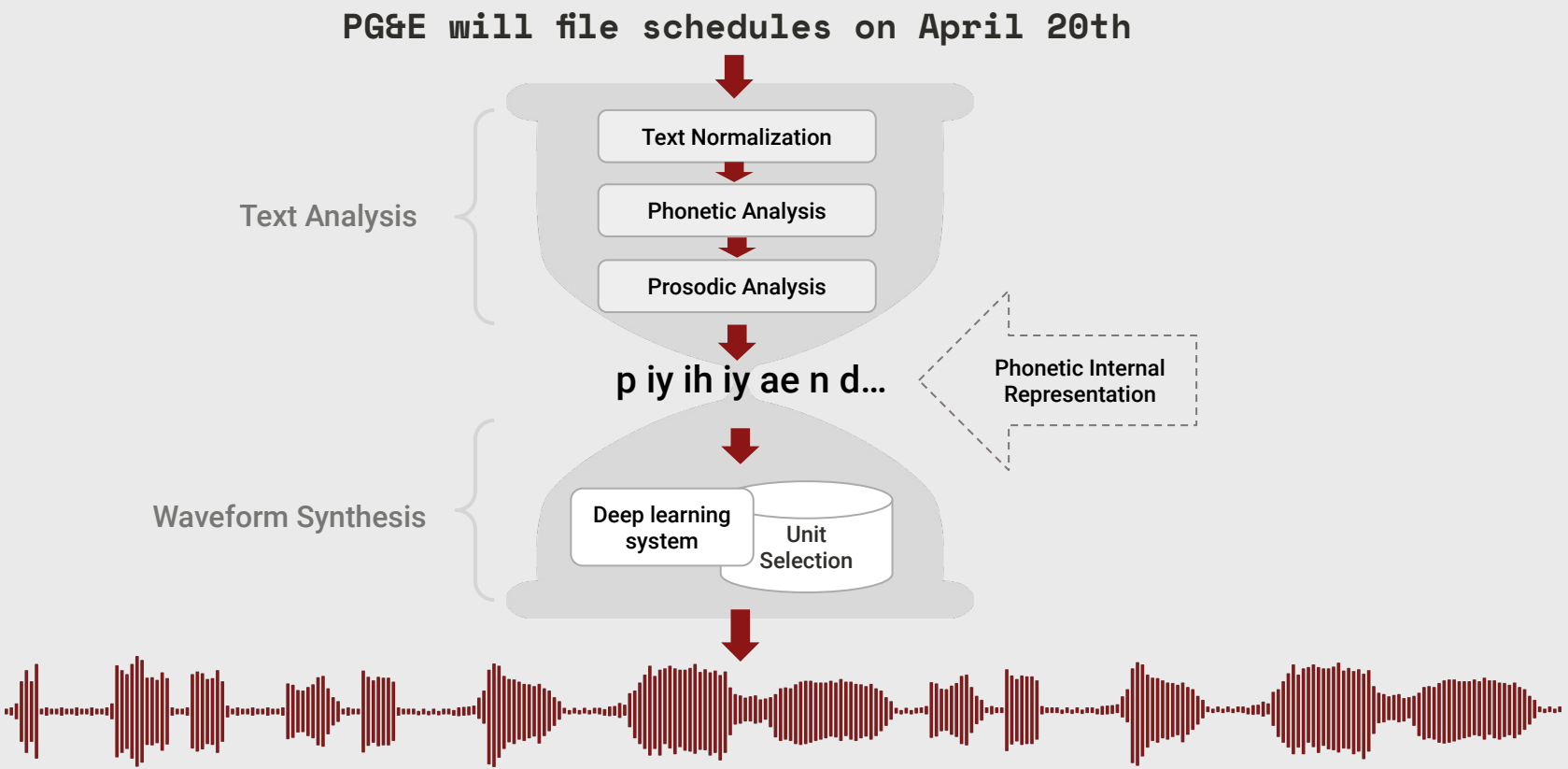
2. **Waveform Synthesis:** From the intermediate representation into audio waveform. *Saying it*



Synthesis in two stages

- **Convert messy text, prosodic commands, or other metadata into normalized text**
  - Synthesis tends to sound more consistent when trained on normalized text data. Even with vast training sets.
  - Public models available. LLMs should be great at this! You can use few-shot LLM prompting to normalize.
  - Some design decisions about how to pronounce some written forms. And how to enable prosody controls
- **Key interface: Intermediate representation output by text conversion. Speech audio generates audio from this representation**
- **Waveform generation. Create audio from text + prosodic information**
  - Modern systems: One or more deep learning modules.
  - Older ideas: Unit selection. Stitch together curated examples of actual speech to form new utterances
  - Text -> Mel spectrogram + voicing -> waveform is a somewhat standard set of steps
  - End-to-end is possible – from text directly to waveform (or at least spectrogram + separate vocoder)

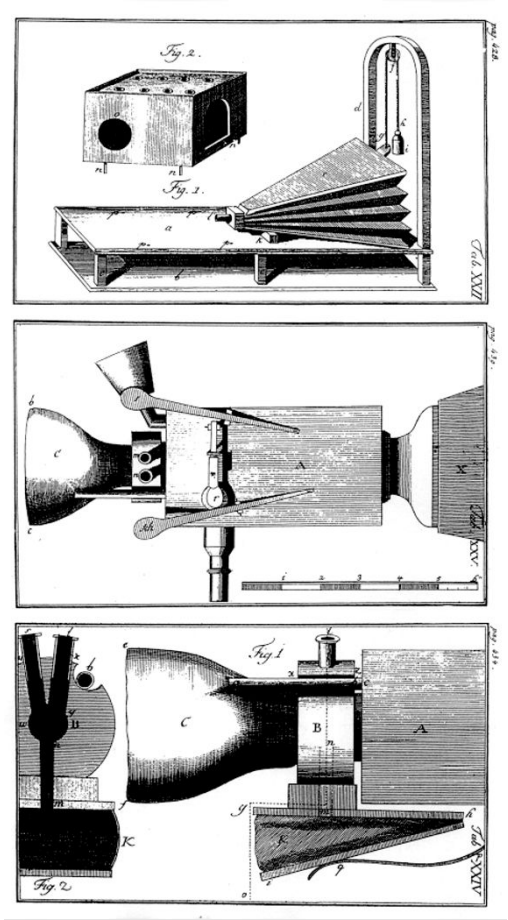
Two-stage approach allows separation of data collection



TTS Modeling History and Overview

Synthesis in 1780 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 website

Homer Dudley 1939 VODER

- Manually controlled through complex keyboard
- Operator training was a problem
- A demo trick:
  Telling the audience what they are about to hear improves understanding

Gunnar Fant's OVE Synthesizer

- Of the Royal Institute of Technology, Stockholm
- Operator training was a problem

Pre-modern TTS

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

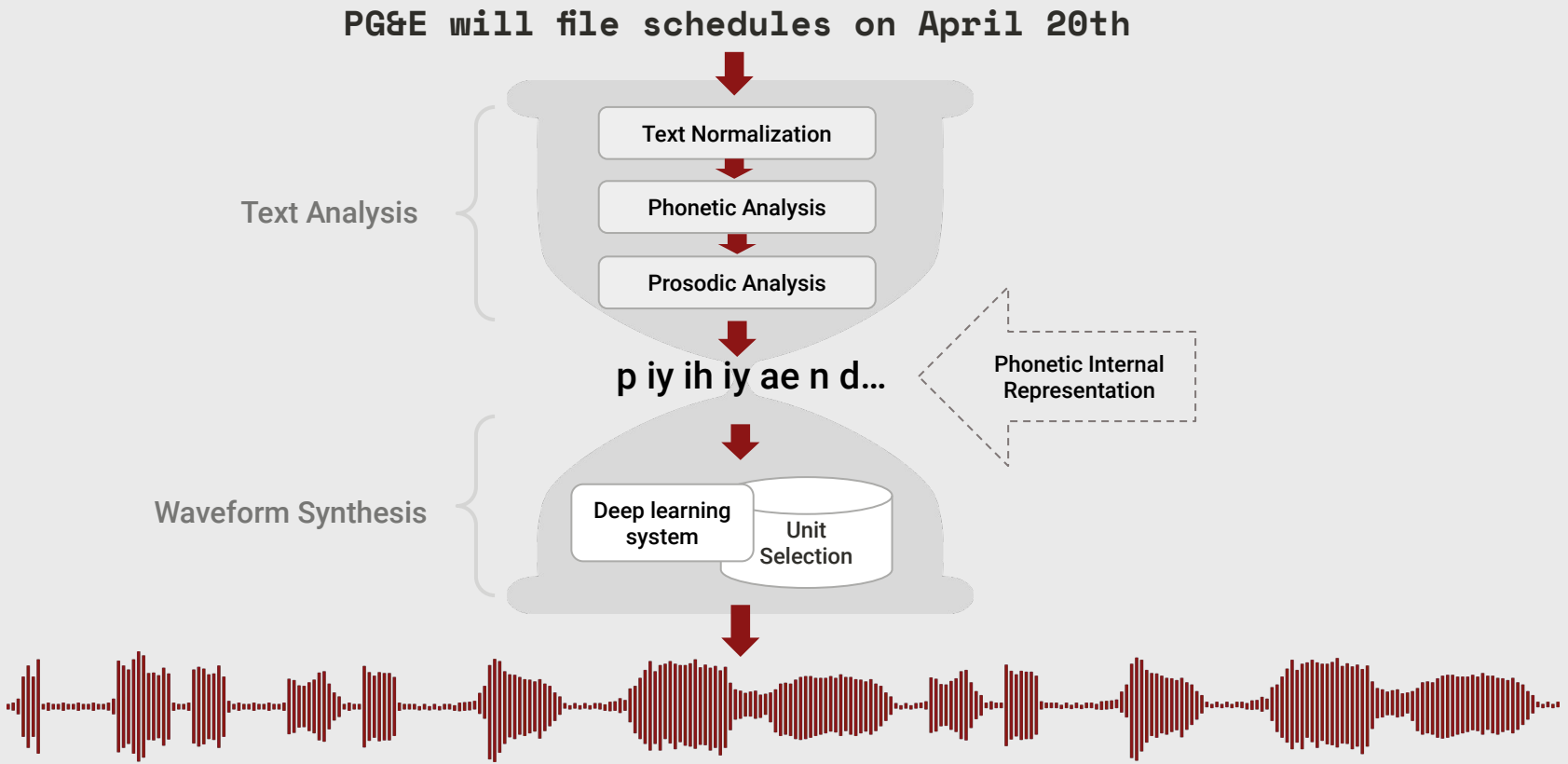
1990s - ~2015: Improving Unit Selection Synthesis

- **Units are diphones or larger word chunks**
    - As computes improved, number of units went from thousands to millions (including specific phrases)
- **Paste them together and modify prosody.**
    - No need to simulate low-level aspects of voice, just use recordings!
    - Generating realistic prosody is hard
    - Systems built on single speaker. No clear way to train on many speakers and generate interpolated or novel voices
- **Small errors occur where units join. Some special heuristics to help smooth joins ("join errors")**
- **Deep learning systems for parametric synthesis improved in early 2010's. Then deep learning took over**
    - Neural nets for TTS go back to at least the 1970's. Dan Jurafsky did a postdoc on this in the early 90's
    - Parametric synthesis simply didn't have good enough function approximation models until 2010's

Parametric Synthesis

- **Train a statistical model on large amounts of data.**
- **Learn text -> sound mapping.**
    - Possibly use phonemes as an intermediate representation.
    - Deep learning architectures have recently explored different ways to split up this task
- **Previously associated with HMM Synthesis**
    - A reverse of HMM-based speech recognition. Lots of hand-coded assumptions.
- **Deep learning approaches (Wavenet, Tacotron) are the latest generation of parametric**
    - Early success of modern deep learning approaches relied on careful audio engineering
    - Use neural net to predict parameters of a hand-engineered vocoder (parameters -> waveform)

Text Normalization


PG&E will file schedules on April 20th
Text Analysis
Text Normalization
Phonetic Analysis
Prosodic Analysis
p iy ih iy ae n d...
Phonetic Internal Representation
Waveform Synthesis
Deep learning system
Unit Selection

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 Rough Steps

1. Identify tokens in text
2. Chunk tokens
3. Identify types of tokens
4. Convert tokens to spoken word forms

Identify Tokens and Chunk

- Whitespace can be viewed as separators
- Punctuation can be separated from the raw tokens
- For example, Festival (older research TTS system) 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

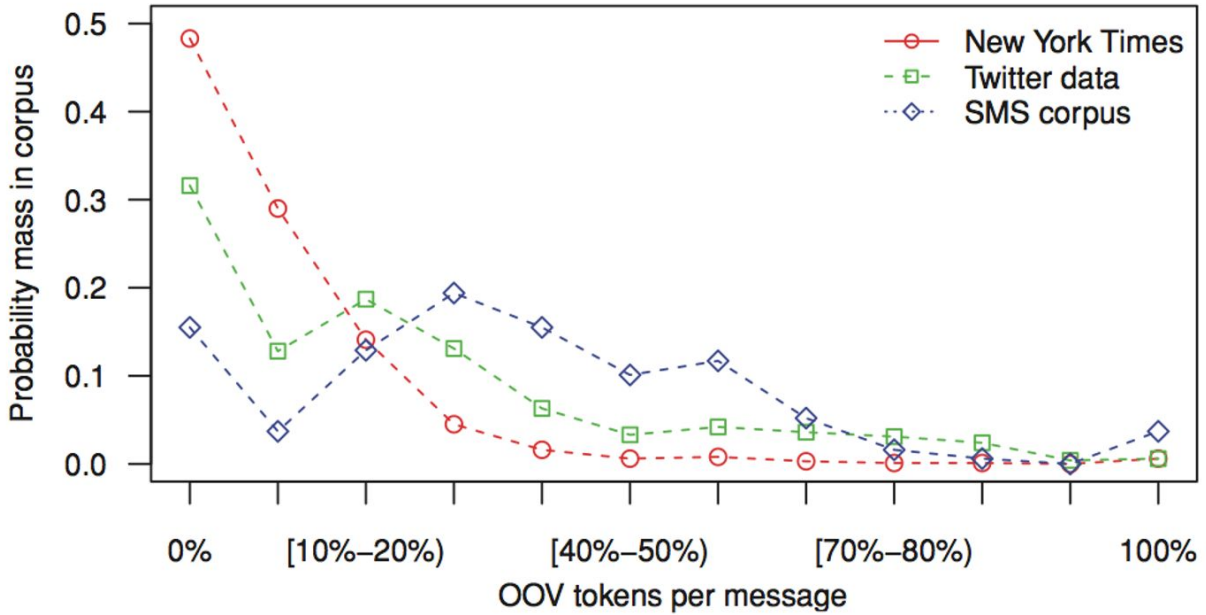
- Relatively simple if utterance ends in ?!
- But what about ambiguity of “.”?
- Ambiguous between end-of-utterance, end-of-abbreviation, and both
  - **My place on Main St. is around the corner.**
  - **I live at 123 Main St.**
  - (Not “**I live at 151 Main St.**”)

Identify Types of Tokens, Convert Tokens to Spoken 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

How Common are Non-standard Words (NSWs)?

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



- 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
- How to build: Pick your favorite machine learning tool and go for it
  - There are increasingly many public models trained on large datasets which offer a great default choice
- Modern deep learning methods trained on large datasets remain state-of-the-art
  - Earlier work: (Black et al. 1998) Two steps: alignment and (CART-based) rule-induction
  - Modern DL approaches might handle tokenization and G2P jointly as a single text normalization module
- Choose your G2P system in the context of the overall TTS architecture you're using
  - General tip: You probably don't need to train your own G2P when building a baseline system

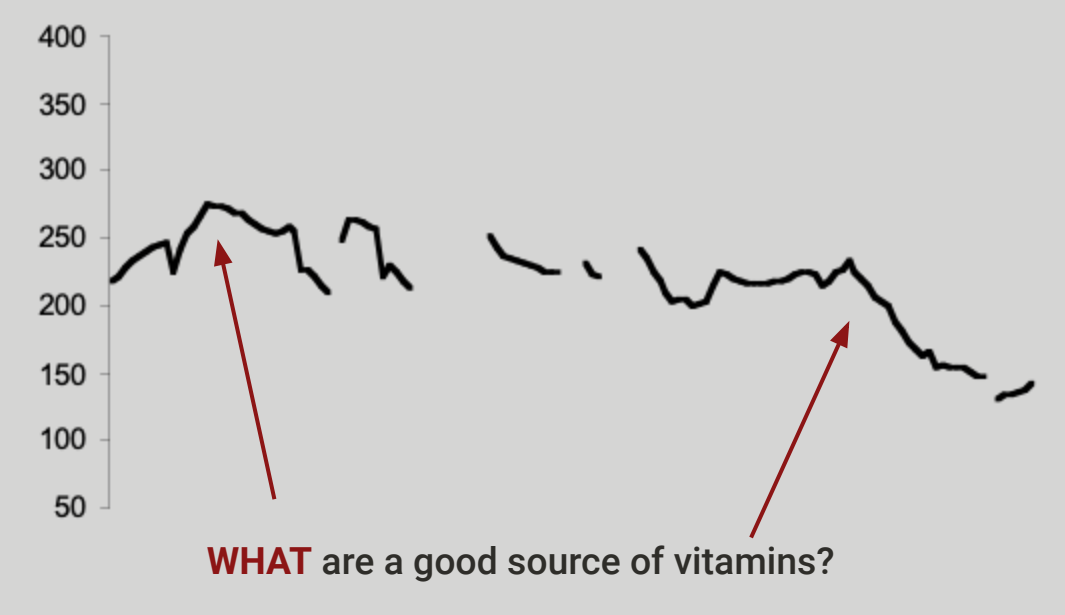
Thank You

More Prosodic Tunes

WH-question Tune

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

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



Broad Focus

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

"Tell me something about the world"

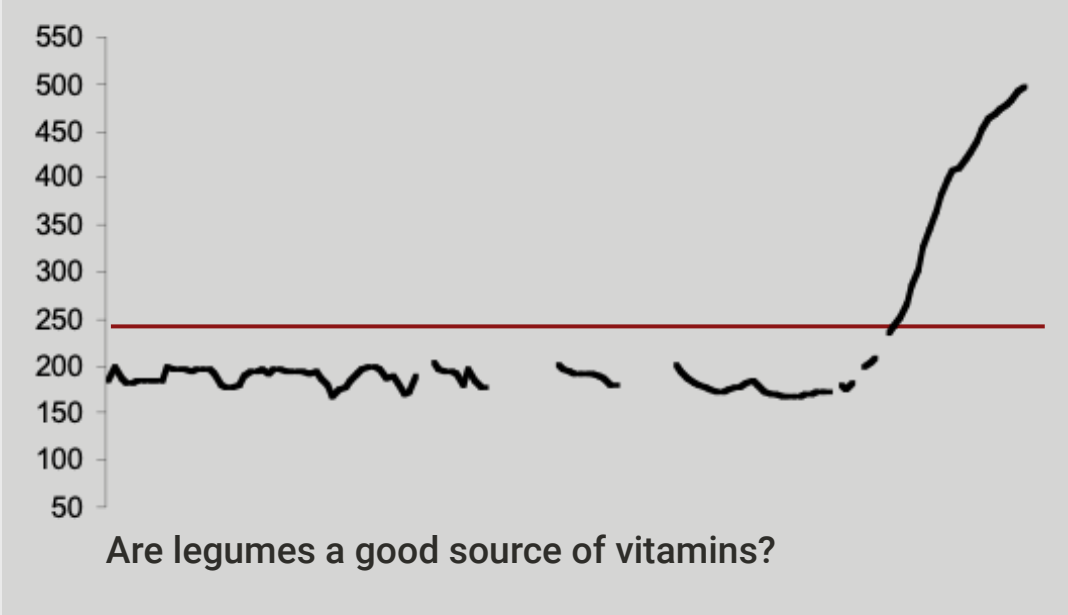
Rising Statements

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

“Tell me something I didn’t know”

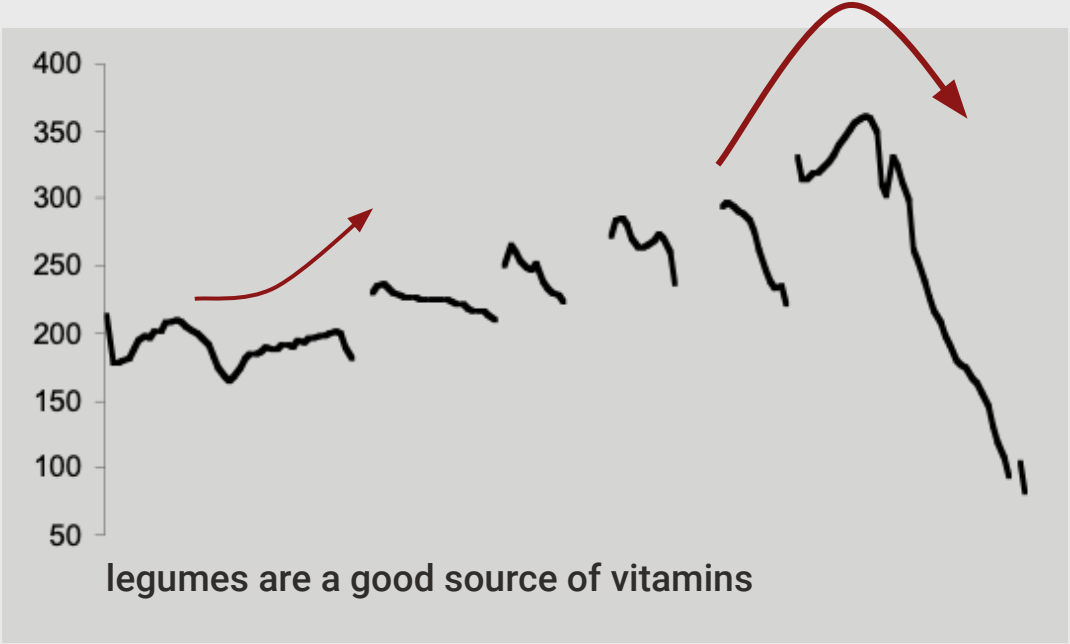
Yes-No question

- **Rise** from the main accent to the end of the sentence



‘Surprise-redundancy’ Tune

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



‘Contradiction’ Tune

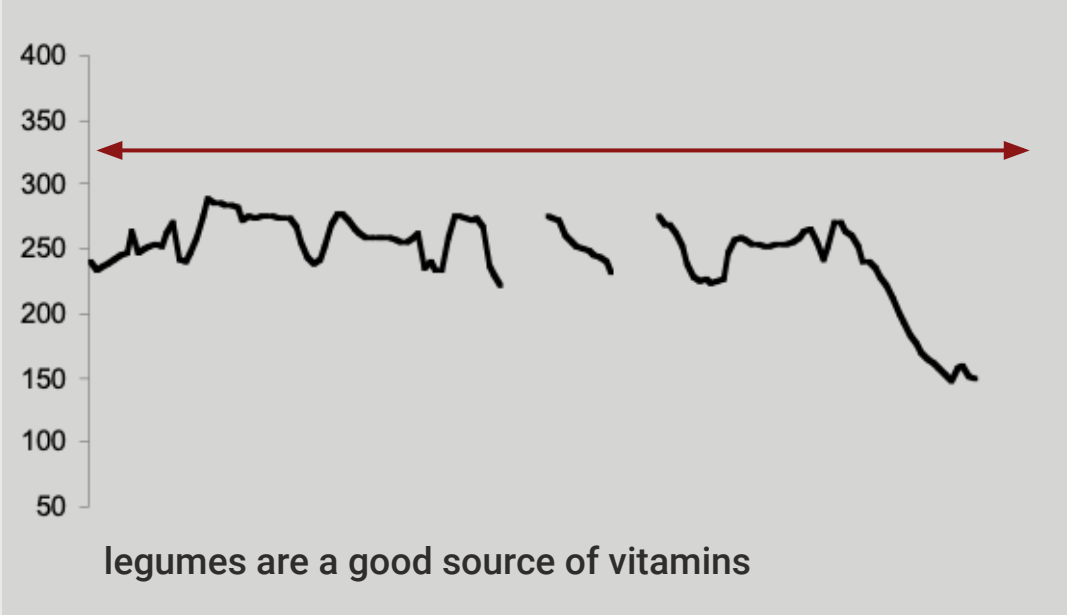
- **Sharp fall** at the beginning, **flat and low**, then **rising** at the end

“I’ve heard that linguini is a good source of vitamins”

[... how could you think that?]

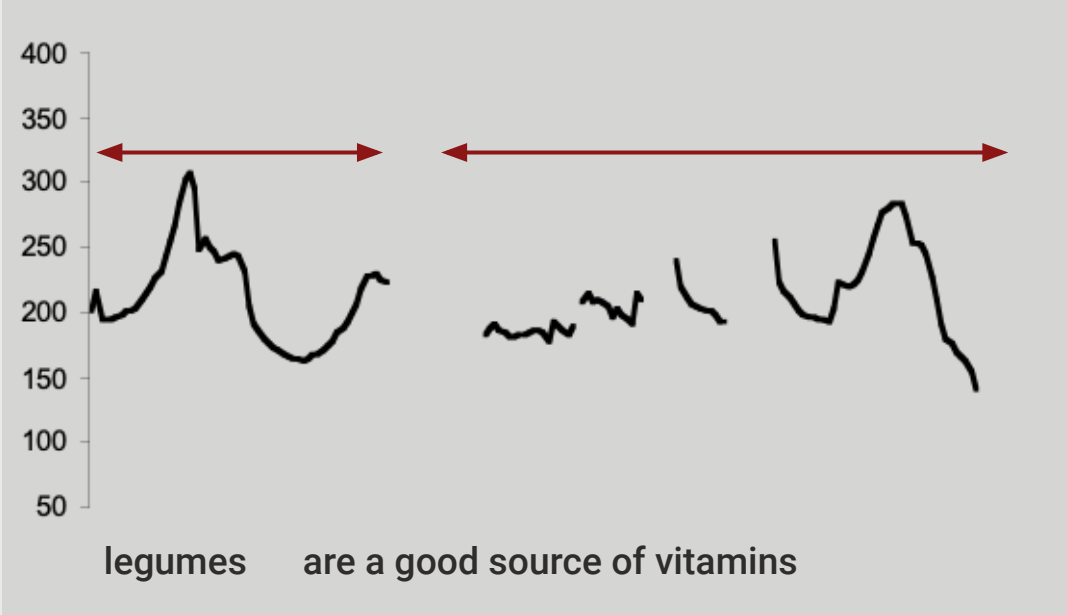
A Single Intonation Phrase

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



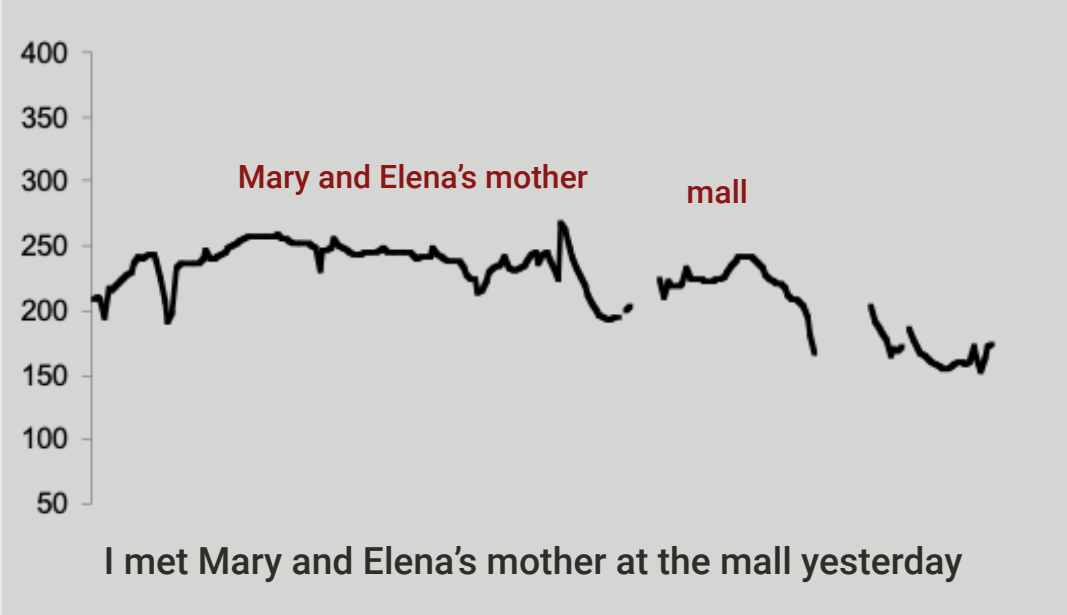
Multiple Phrases

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



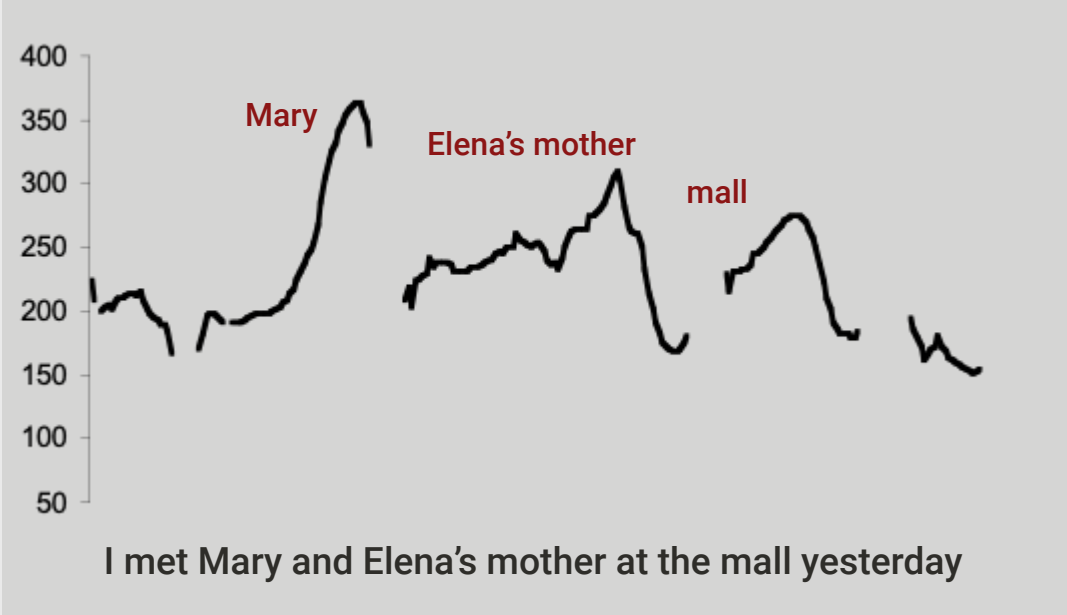
Phrasing Sometimes Helps Disambiguate

- One intonation phrase with relatively flat overall pitch range



Phrasing Sometimes Helps Disambiguate

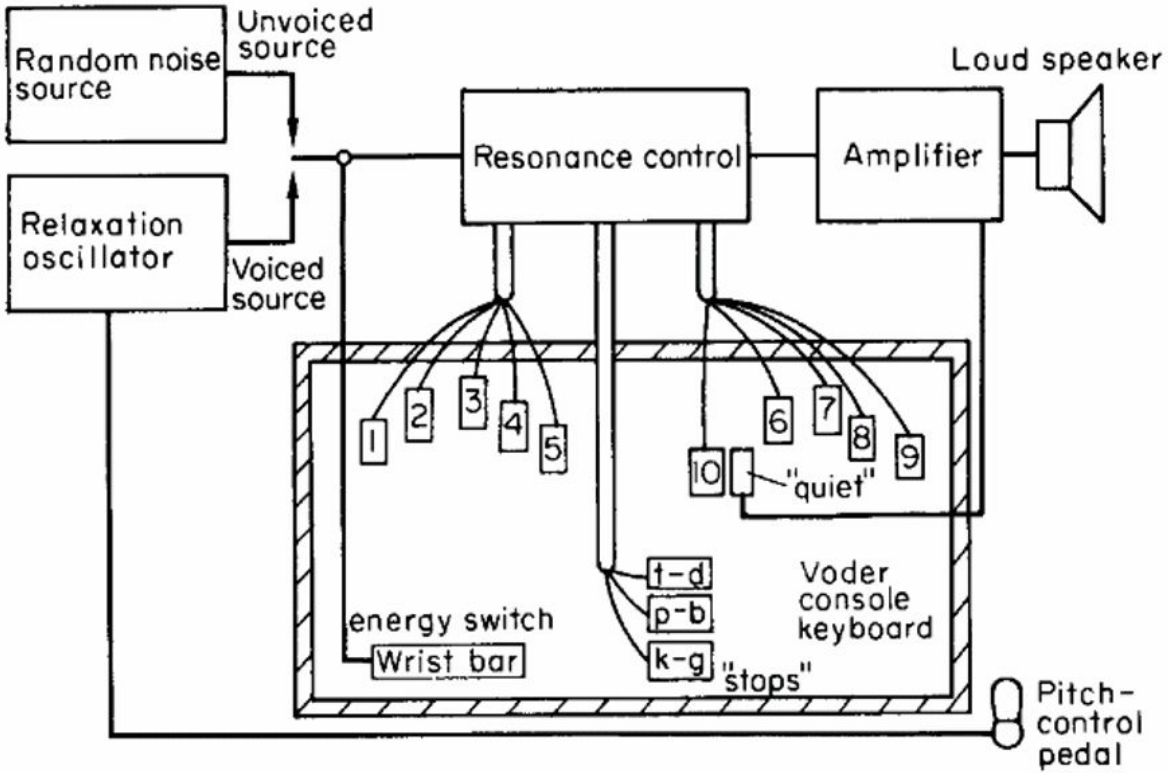
- Separate phrases, with expanded pitch movements



Appendix

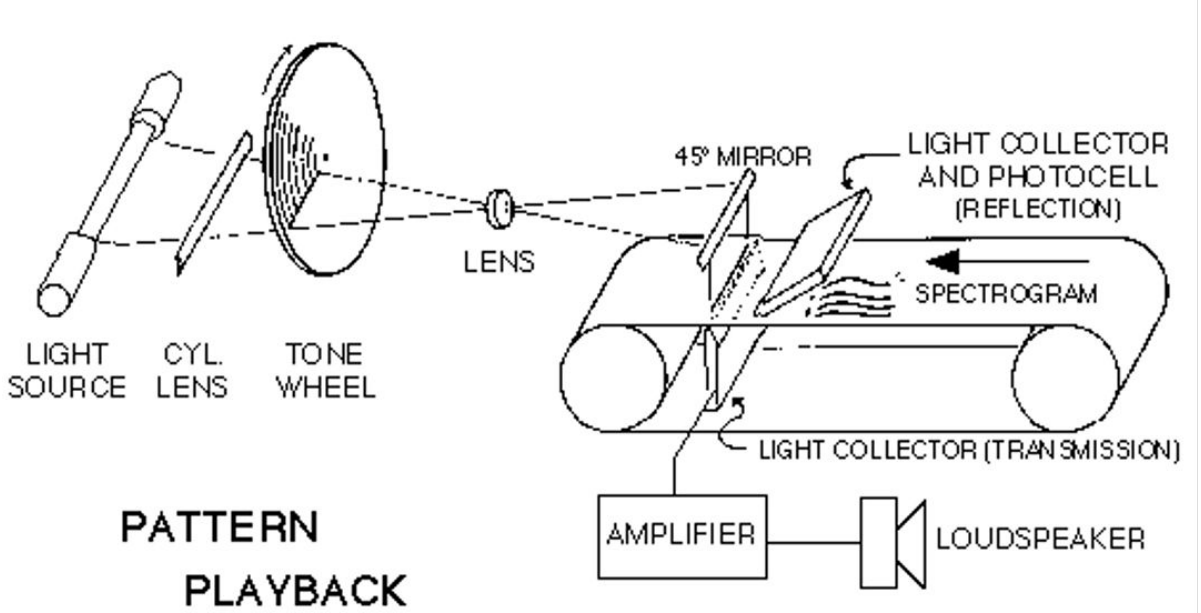
Homer Dudley 1939 VODER

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



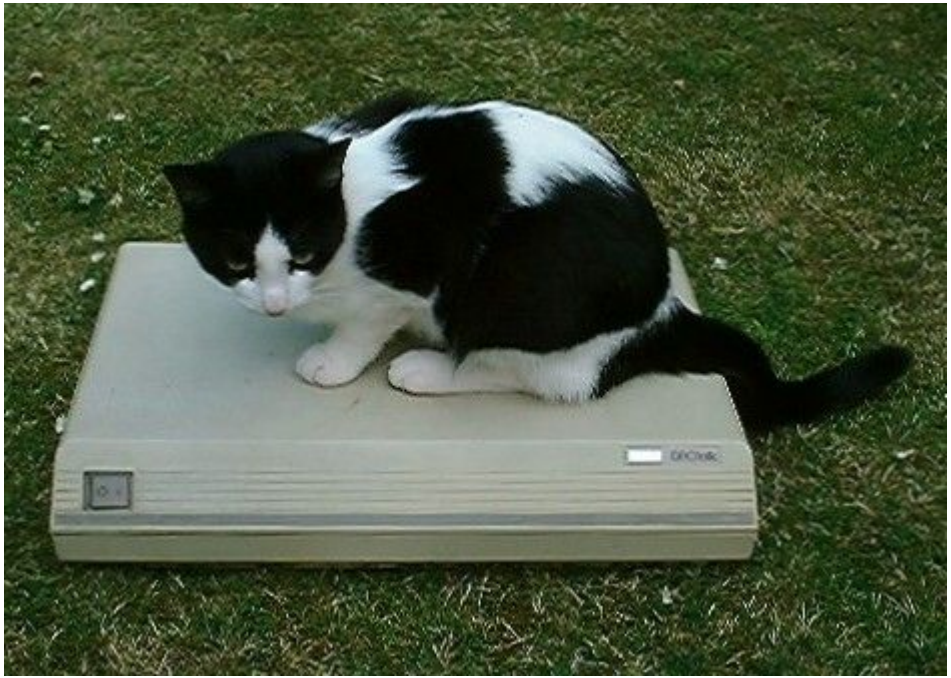
Cooper's Pattern Playback

- 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



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”



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

Slide: Richard Sproat

Festival as Software

- http://festvox.org/festival/
- General system for multilingual 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

Slide: 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

Slide: Richard Sproat

Dictionaries

- FCMU 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/

Slide: Richard Sproat

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%
classifieds	17.9%

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, Jimenez, Rajan, Raghavan, Sondhi, Xu, Hsu, Zhang, Chang, Nguyen
  - Company/Brand names: Infinit, Kmart, Cytyc, Medamicus, Inforte, Aaon, Idexx Labs, Bebe

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

Part of Speech Tagging for Homograph Disambiguation

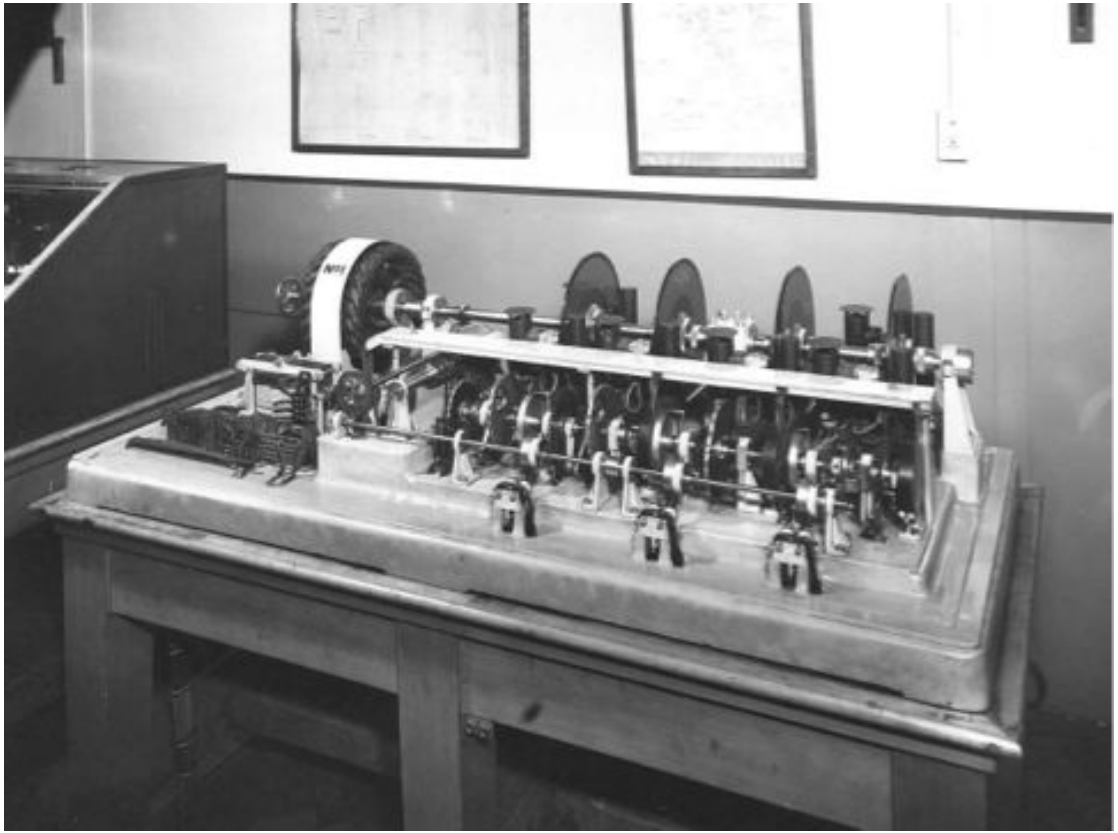
- Many homographs can be distinguished by POS

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

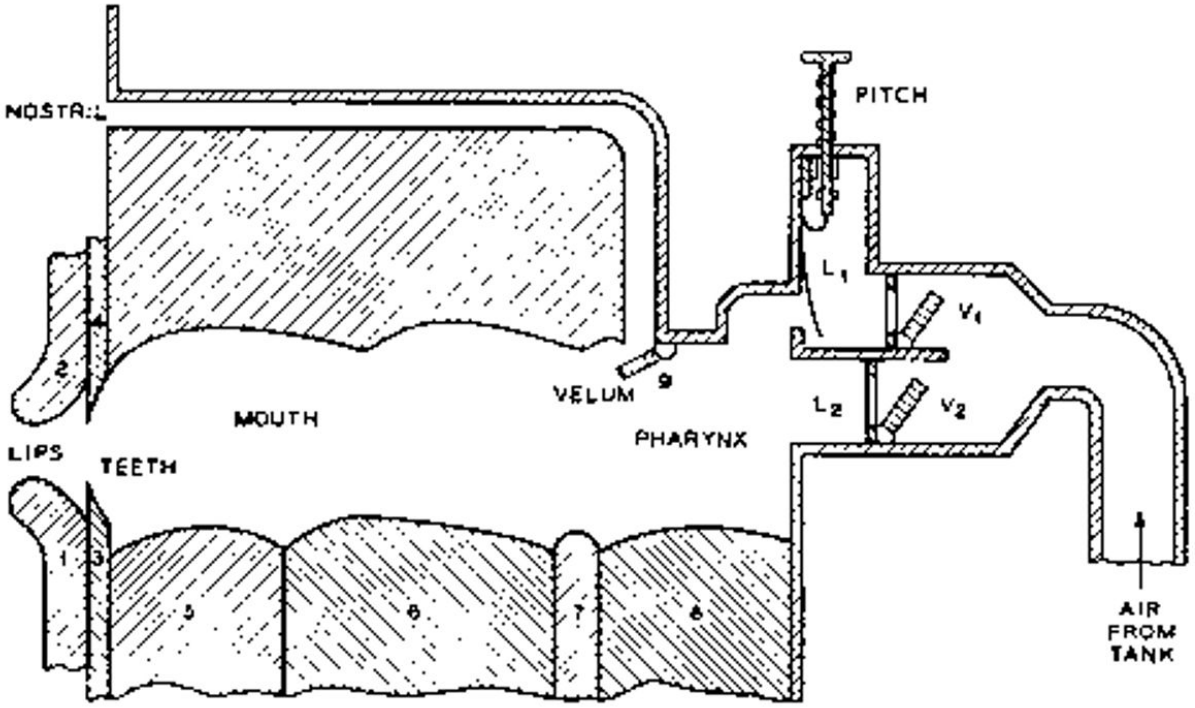
- July 24, 1936
- Photographic storage on 4 glass disks
- 2 disks for minutes, 1 for hour, one for seconds.
- Other words in sentence distributed across 4 disks, so all 4 used at once.
- Voice of “Miss J. Cain”



The 1936 UK Speaking Clock

- July 24, 1936
- A technician adjusts the amplifiers of the first speaking clock

Closer to a Natural Vocal Tract: Riesz 1937



Some More Modern Examples

Sample 1

Sample 2

Sample 3

Sample 2