Speech Synthesis: Then and Now

Julia Hirschberg

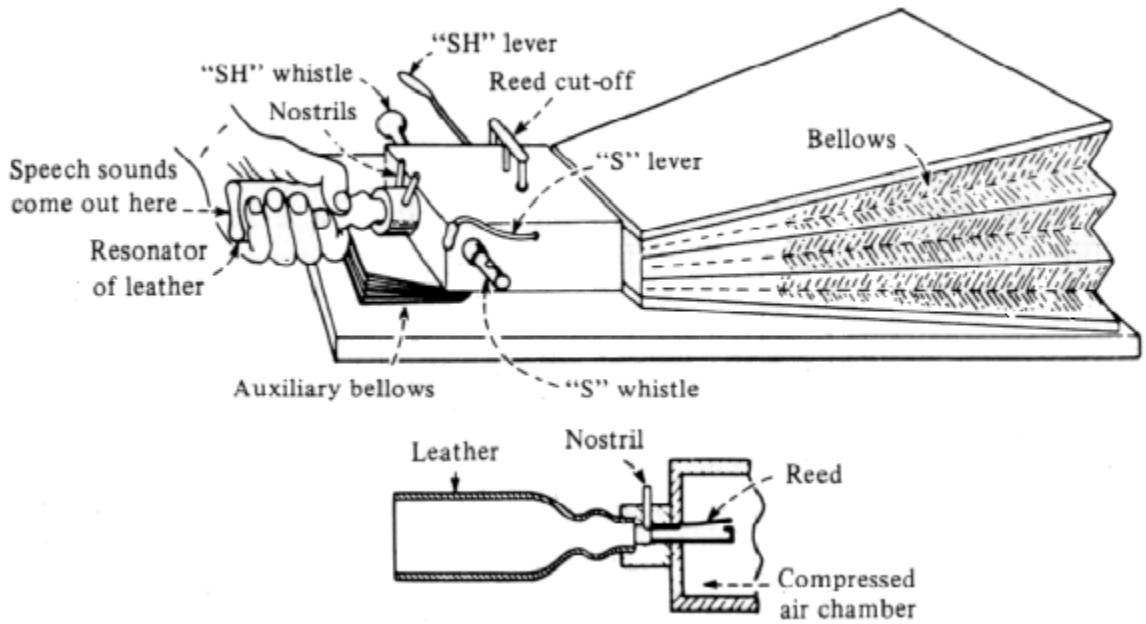
CS 4706

Today

- Then: Early speech synthesizers
- Now: Overview of Modern TTS Systems
- Think about: how do we ***evaluate*** a synthesizer

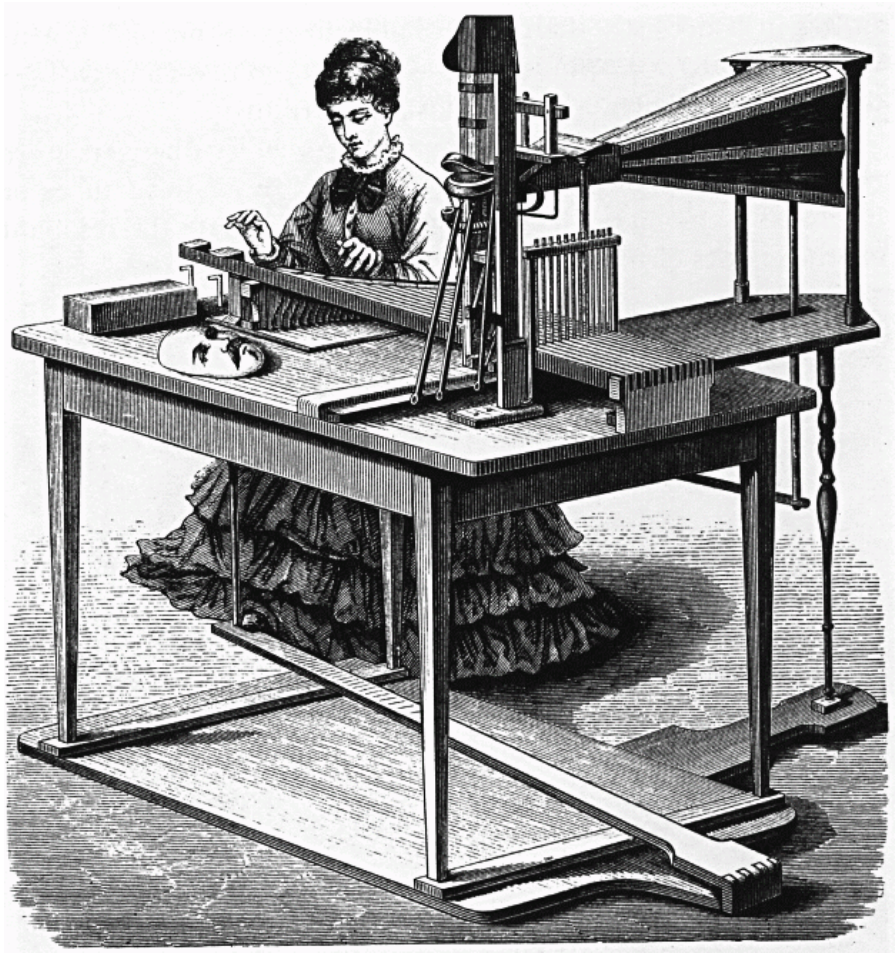
The First ‘Speaking Machine’

- Wolfgang von Kempelen, Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine, 1791 (in Deutsches Museum still and playable)



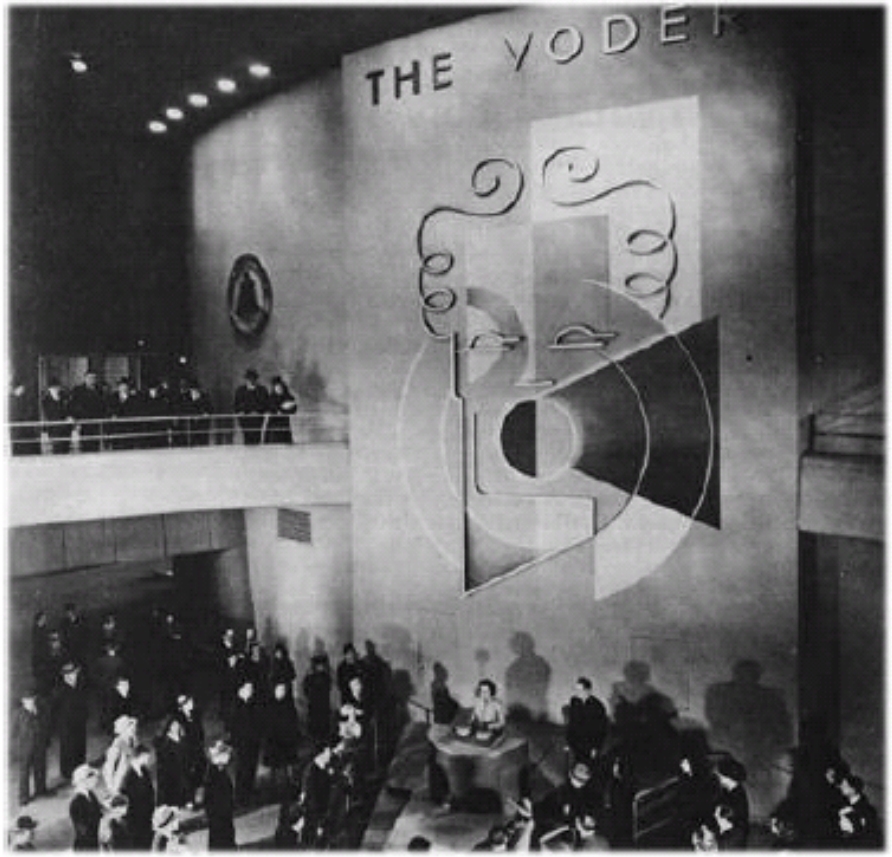
- First to produce whole words, phrases – in many languages

Joseph Faber's Euphonia, 1846



- Constructed 1835 w/pedal and keyboard control
  - Whispered and ordinary speech
  - Model of tongue, pharyngeal cavity with manipulable shape
  - Singing too: “God Save the Queen”
- Riesz’s 1937 synthesizer with almost natural vocal tract shape
- Forerunners of Modern Articulatory Synthesis: George Rosen’s DAVO synthesizer (1958) at MIT

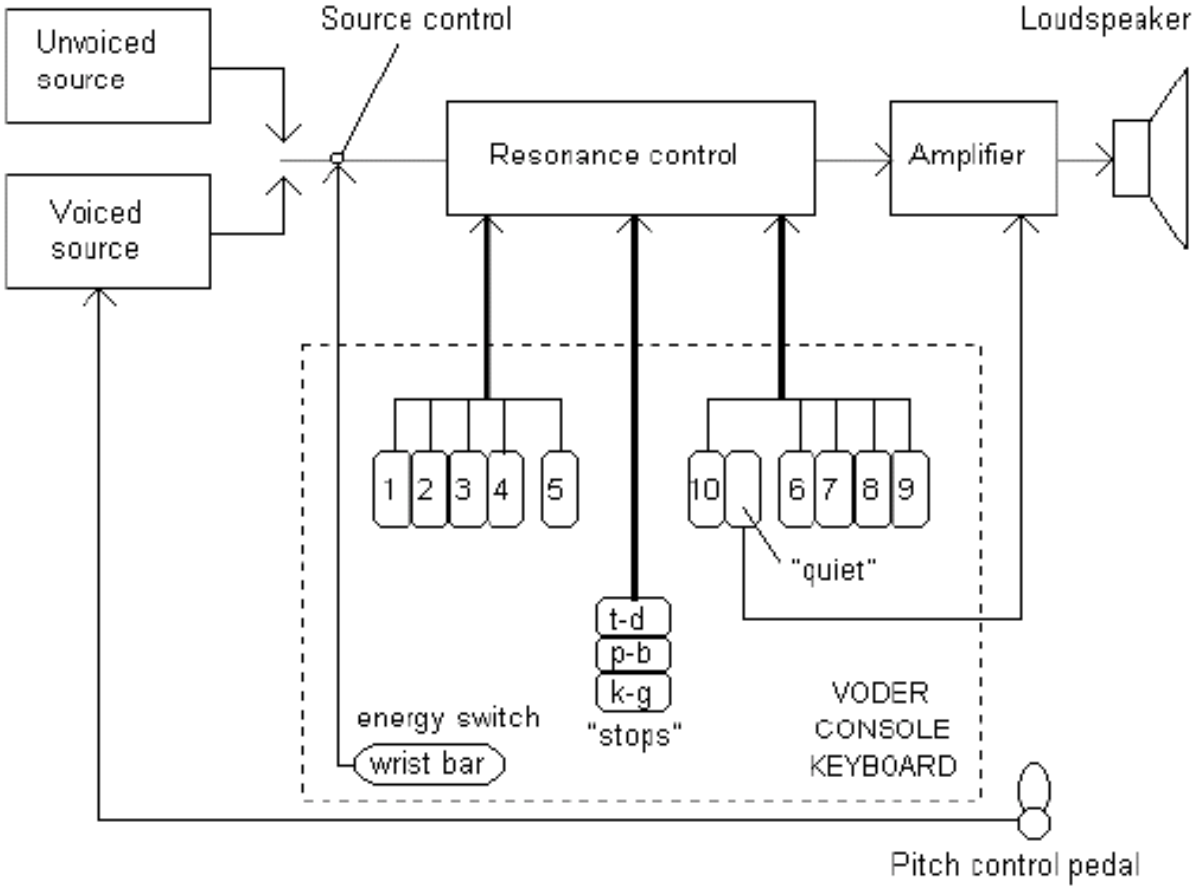
The Voder ...



Developed by Homer Dudley at Bell Telephone Laboratories, 1939

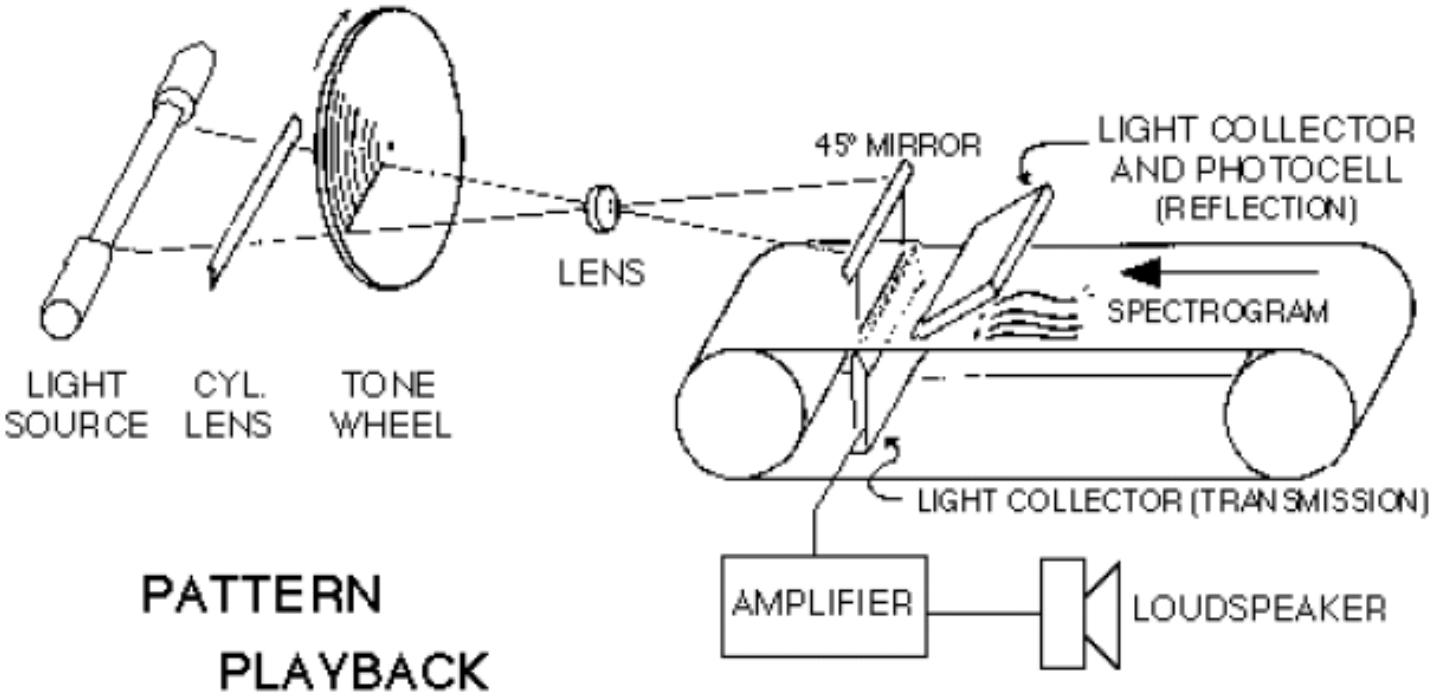
- World’s Fair in NY, 1939
- Requires much training to ‘play’
- Purpose: coding/compression
  - Reduce bandwidth needed to transmit speech, so many phone calls can be sent over single line

... an acoustic synthesizer



Architectural blueprint for the Voder

The Pattern Playback



Developed by Franklin Cooper at Haskins Laboratories, 1951

- Answers:
  - These days a chicken leg is a rare dish.
  - It’s easy to tell the depth of a well.
  - Four hours of steady work faced us.
- ‘Automatic’ synthesis from spectrogram – but can also use hand-painted spectrograms as input
- Purpose: understand perceptual effect of spectral details

Formant/Resonance/Acoustic Synthesis

- Parametric or resonance synthesis
  - Specify minimal parameters, e.g. f0 and first 3 formants
  - Pass electronic source signal thru filter
    - Harmonic tone for voiced sounds
    - Aperiodic noise for unvoiced
    - Filter simulates the different resonances of the vocal tract
- E.g.
  - Walter Lawrence's Parametric Artificial Talker (1953) for vowels and consonants
  - Gunnar Fant's Orator Verbis Electris (1953) for vowels
  - Formant synthesis download (M$demo)

Synthesis by Computer

- Beginnings ~1960; dominant from 1970—

Ignatius Mattingly, 1974: "The advantage of a simulation [by computer] is that it can be completely reliable and accurate, and the design of the synthesizer can be readily modified; the disadvantage is that an extremely powerful computer is required and such computers are too expensive to permit extended real-time operation."

Concatenative Synthesis

- Most common type today
- First practical application in 1936: British Phone company's Talking Clock
  - Optical storage for words, part-words, phrases
  - Concatenated to tell time
- E.g.
- And a 'similar' example from Radio Free Vestibule (1994)
- Bell Labs TTS (1977) (1985)

Variants of Concatenative Synthesis

- Inventory units
  - Diphone synthesis (e.g. Festival)
  - Microsegment synthesis
  - “Unit Selection” – large, variable units
- Issues
  - How well do units fit together?
  - What is the perceived acoustic quality of the concatenated units?
  - Is post-processing on the output possible, to improve quality?

Overview: Synthesizer I/O

- Front end: From input to control parameters
  - Acoustic/phonetic representations, naturally occurring text, constrained mark-up language, semantic/conceptual representations
- Back end: From control parameters to waveform
  - Articulatory, formant/acoustic, concatenative, (diphone, unit-selection/corpus, HMM) synthesis

TTS Production Levels

Knowledge

- World Knowledge
- Syntax, semantics, lexicon
- Phonetics/phonology
- Acoustics/signal processing

Task

- Text Normalization
- Pronunciation, intonation assignment
- Duration, f0, durations
- Waveform production

Text Normalization Issues

- Reading is what W. hates most.
- Reading is what Wilde hated most.
- The NAACP just elected a new president.
- In 1996 she sold 2010 shares and deposited $42 in her 401(k).
- The duck dove supply.
- Homographs, numbers, abbreviations

Pronunciation Issues

- Rules for disambiguation in context: bass
- Lexicon: comb, tomb, Punxsutawney Phil
  - Letter-to-Sound Rules
    - Hand built
    - Learned from data (pronunciation dictionary)
    - Hard to get good accuracy and coverage – many exceptions
  - Dictionary of pronunciations
    - More accurate
    - New (Out-of-Vocabulary) words a problem

Intonation Assignment Issues: Phrasing

- Traditional: hand-built rules
  - Use punctuation: 234-5682, New York, NY
  - Context/function word: no breaks after function word: He went to dinner. He came to and went to dinner.
  - Syntax: She favors the nuts and bolts approach. She went home and Dave stayed.
- Current: machine learning on large labeled corpus

Intonation Assignment Issues: Accent

- Hand-built rules
  - Function/content distinction He went out the back door/He threw out the trash
  - Complex nominals:
    - Main Street/Park Avenue
    - city hall parking lot (stress shift)
- Statistical procedures trained on large corpora
  - Need lots of data
  - Why learn what you already know?

Intonation Assignment Issues: Contours

- ***Simple*** rules
  - '.' = declarative contour
  - '?' = yes-no-question contour *unless wh*-word present at/near front of sentence
    - Well then, how did he do it? And what do you know?
- Pretty monotonous in long stretches of speech
- Problem: no one knows how to assign other contours from text

Phonological Specification Issues

- Task is to produce a phonological representation from phones and intonational assignment
  - Align phones and f0 contour
  - Specify durations and intensity
- Select/create appropriate acoustic realization from this specification:
  - Acoustic transformation
  - Concatenation: diphone, unit selection
  - HMM

Not Quite There

- Festival concatenative:
- Acuvoice concatenative:
- HMM synthesis (Rob Donovan):
- Rhetorical unit selection
  - (acquired by Nuance)
- AT&T Labs Naturally Speaking

Next Class

- Project Phase I assigned: building a TTS System
- Introduction to Festival TTS