Speech Generation
From Concept and from Text

Martin Jansche

CS 6998

2004-02-11

Components of spoken output systems

Front end: From input to control parameters.

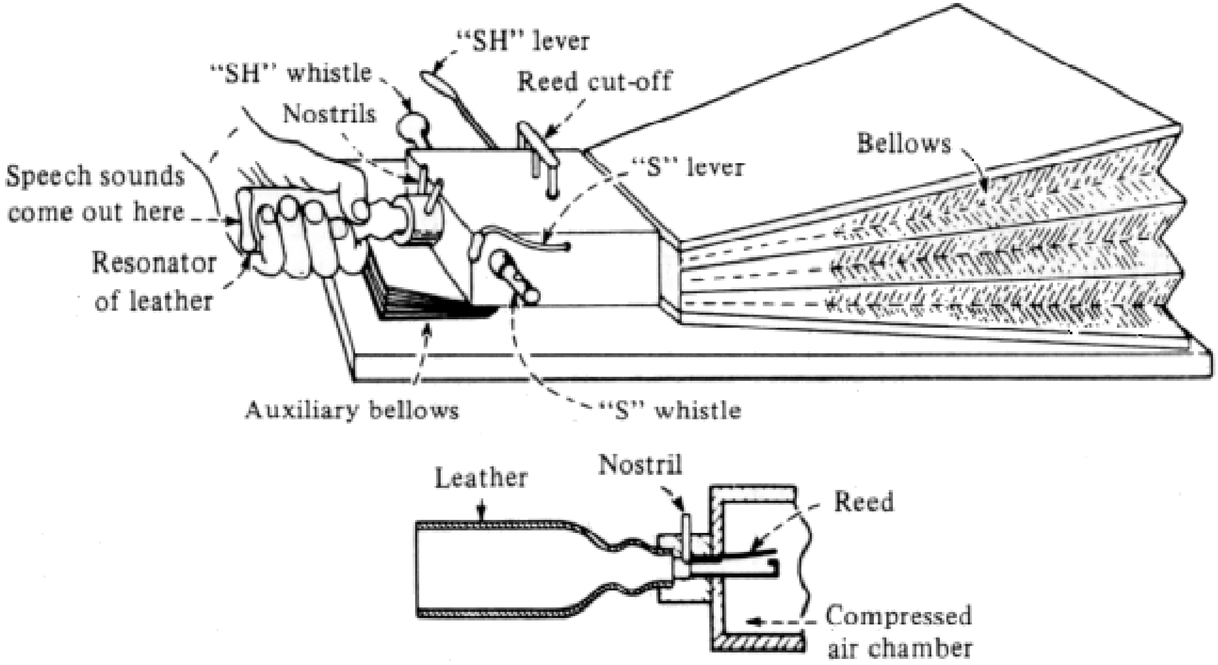
- From naturally occurring text; or
- From constrained mark-up language; or
- From semantic/conceptual representations.

Back end: From control parameters to waveform.

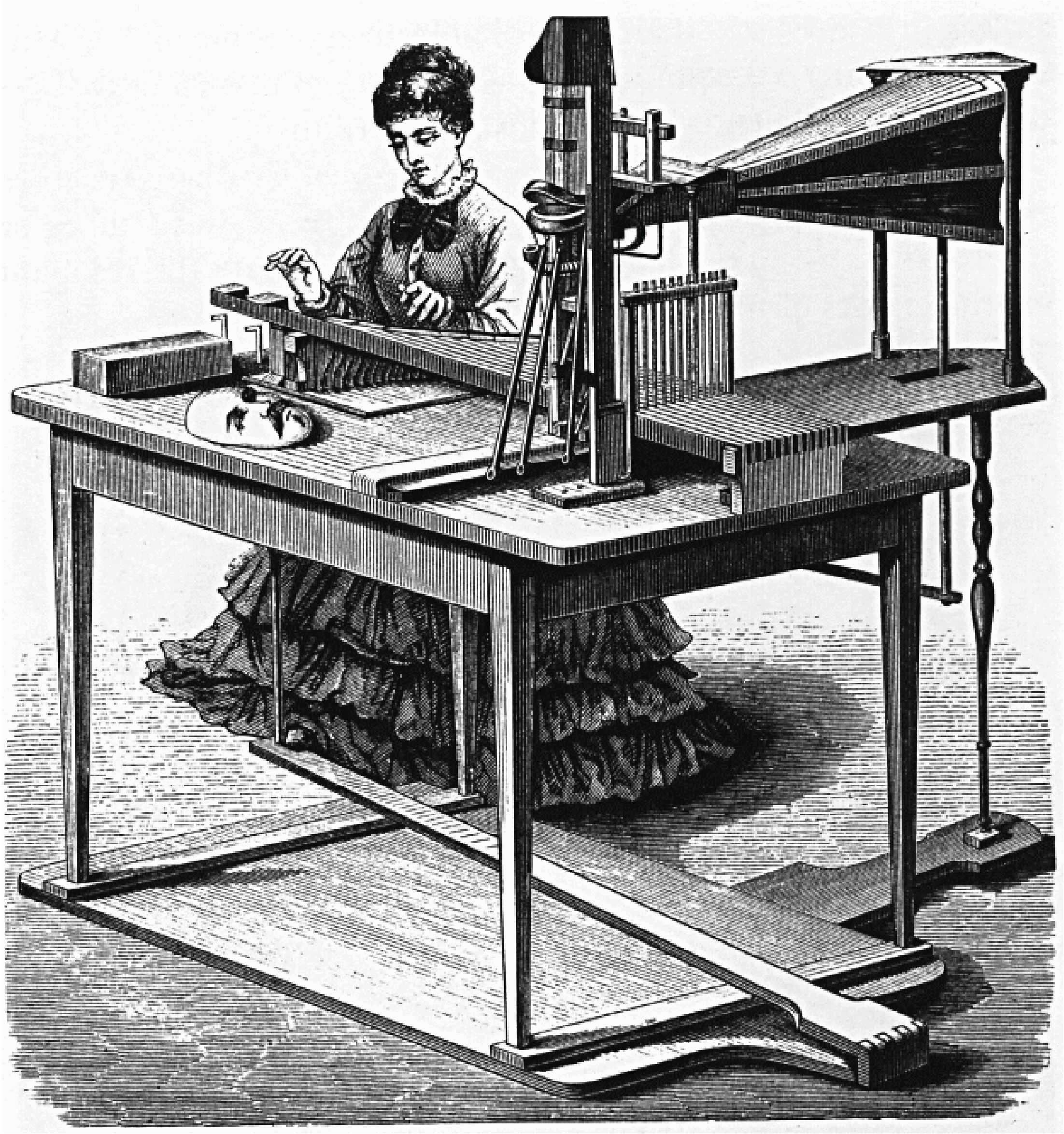
- Articulatory synthesis; or
- Acoustic synthesis:
  - Based predominantly on speech samples; or
  - Using mostly synthetic sources.

Who said anything about computers?

Wolfgang von Kempelen, *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine*, 1791.



Charles Wheatstone's reconstruction of von Kempelen's machine

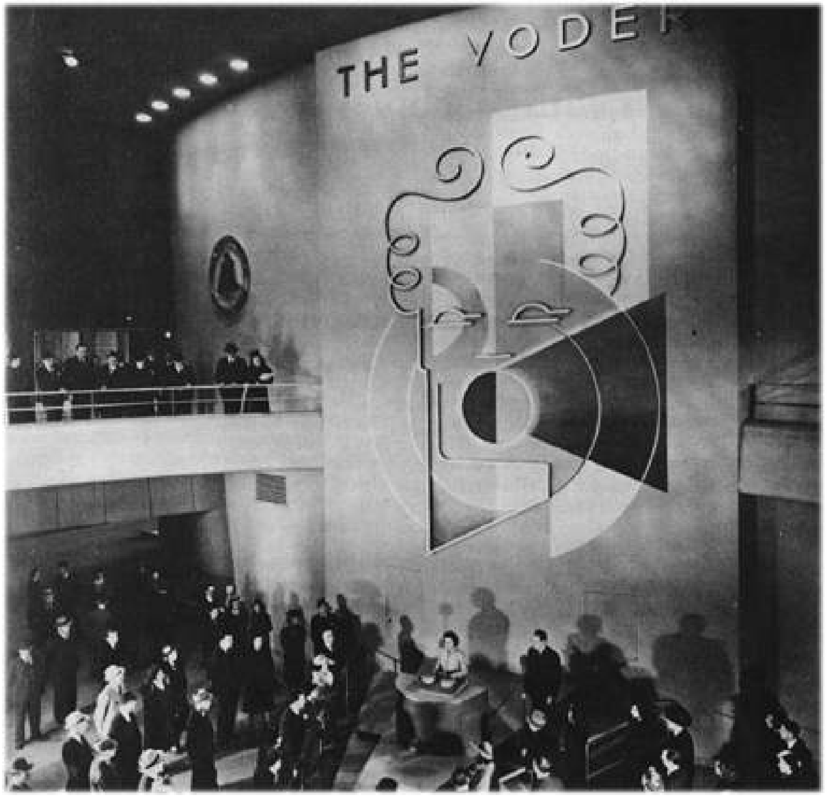


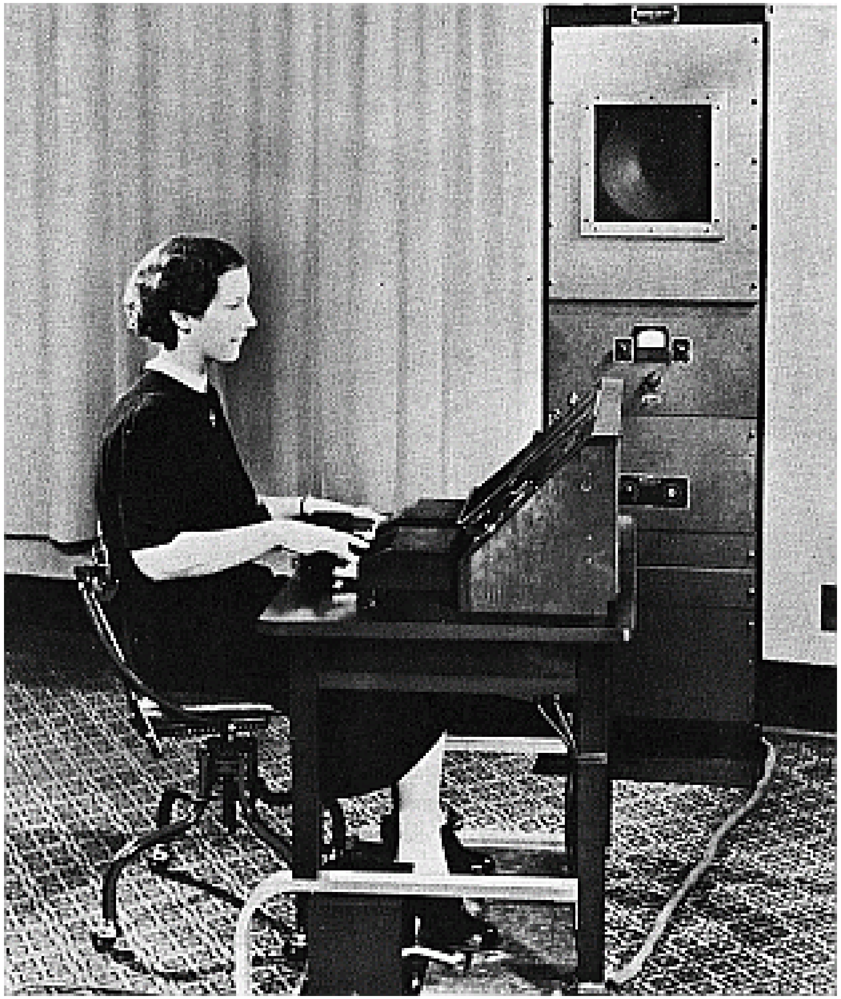
Joseph Faber's *Euphonia*, 1846

Modern articulatory synthesis

- Output produced by an articulatory synthesizer from Dennis Klatt's review article (JASA 1987)
- Praat demo
- Overview at Haskins Laboratories (Yale)

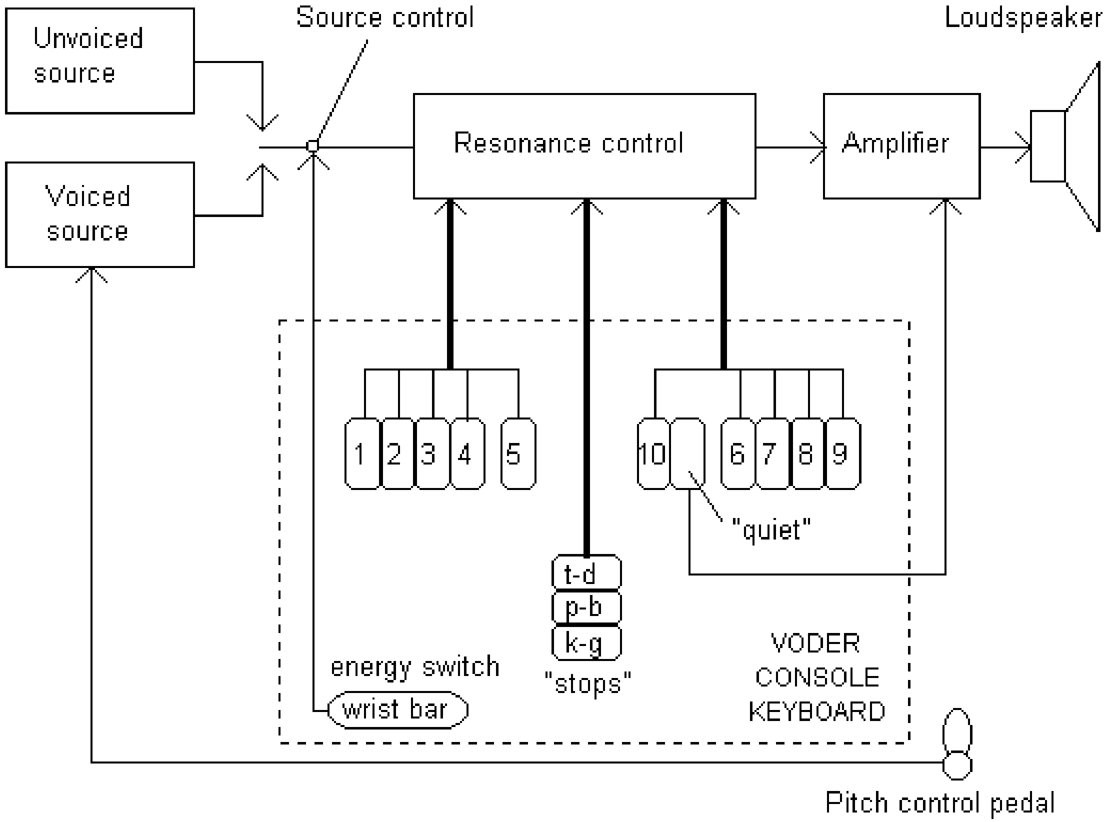
The Voder ...





Developed by Homer Dudley at Bell Telephone Laboratories, 1939

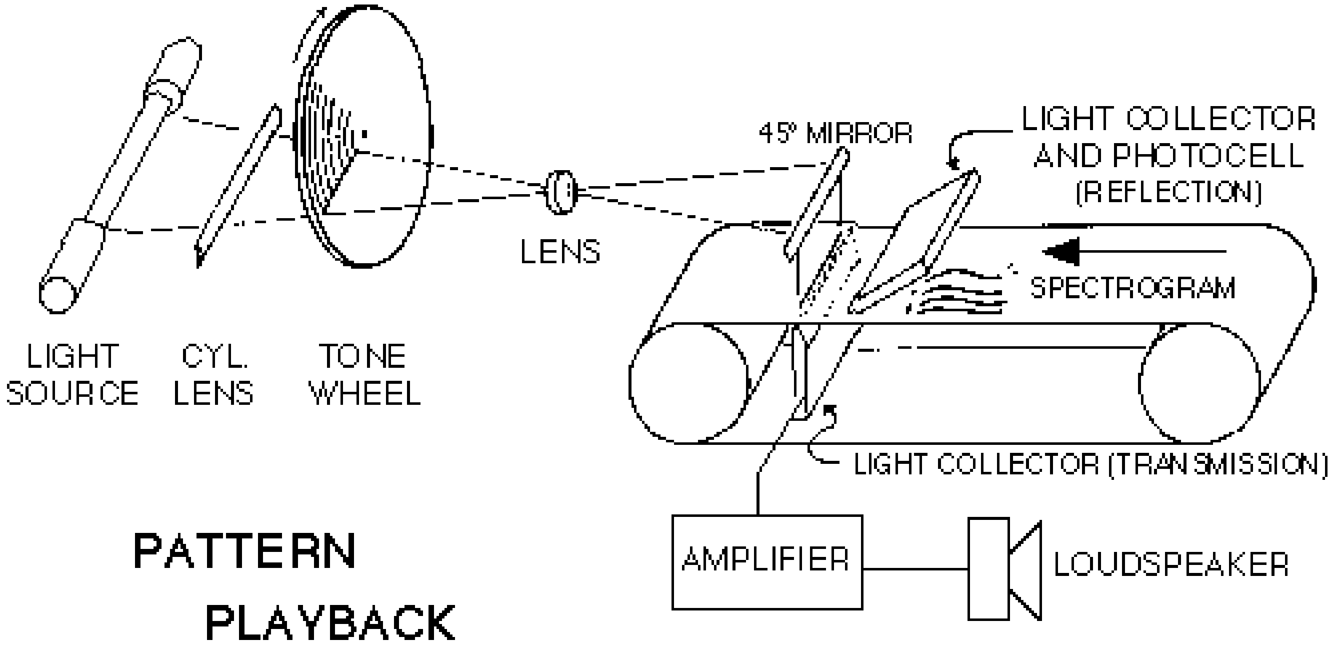
… an acoustic synthesizer



Architectural blueprint for the Voder

Output produced by the Voder

The Pattern Playback



Developed by Franklin Cooper at Haskins Laboratories, 1951

No human operator required. Machine plays back previously drawn spectrogram (spectrograph invented a few years earlier).

Can you understand what it says?

Output produced by the *Pattern Playback*.

Can you understand what it says?

Output produced by the *Pattern Playback*.

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.

Synthesis-by-rule

- Realization that spectrograph and *Pattern Playback* are really only recording and playback devices. Within limitations, they can record and resynthesize any sound, not just speech.

- What are the spectral characteristics of speech?

- Formulation of instructions for painting spectrograms.

- Draw the spectrogram of the utterance directly, following written instructions.

- This briefly reintroduces the human operator.

Parametric synthesizer

- Also known as *formant synthesizer* or *resonance synthesizer*.
- Only a few parameter need to be specified – typically the central frequencies of the fundamental frequency and the location of the first three formants.
- Walter Lawrence's *Parametric Artificial Talker*, 1953. (Output)
- Gunnar Fant's *Orator Verbis Electris*, 1953. (Output)
- Formant synthesis demo.

Speech synthesis by computer

- Starting around 1960, ubiquitous from 1970 onward.

- Computer technology was a limiting factor then.

  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.”

- Increase in commercial applications starting in mid 1970s.

Concatenative synthesis

- Splice together speech snippets.

- “Probably the first practical application of speech synthesis was in 1936 when the UK Telephone Company introduced a speaking clock. It used optical storage for the phrases, words and part-words which were appropriately concatenated to form complete sentences.”

- Steady rise since 1970s, dominant paradigm today.

- A contemporary example.

- A, uhm, “different” example.

Variants of concatenative synthesis

The size and shape of the inventory of synthesis units vary:

- Diphone synthesis (e.g. Festival).
- Microsegment synthesis.
- So-called unit-selection synthesis.

Basic issues:

- What is the acoustic quality of a unit in a given context, after possible modifications?
- How well do the units fit together?

The interface to the front end

In order to synthesize a waveform, we need to have:

- A specification of the segments (parameters of the filter); and
- A specification of so-called suprasegmentals like duration, fundamental frequency, amplitude, etc. (parameters of the source).

It's the job of the front end to provide these specifications.

Some phenomena of textual input

Reading is what W. hates most.

Reading is what Wilde hated most.

Have the students read the questions

Dr. Smith lives on Elm St., but St. John lives on Oak Dr.

Dr. North lives on Maple Dr. South.

In 1996 she sold 1995 shares and deposited $42 in her 401(k).

The 5tet features Dave Holland on lead bass.

What do you know about scales on a bass?

The duck dove supply

RTFM, IMHO, AFAIK, OTOH, ANFSCD

scuba, laser, radar

Tcl (tickle), PNG (ping), DLX (deluxe), SCSI (scuzzy)

UFO, NAACL

USBancorp

Knowledge that comes into play

- Part of speech (noun, verb, adjective): dove, multiply, coax
- Subject matter (marine biology, music): bass
- Text category (personal email, recipe, classified ad, software license): drm, IMO
- Origin (Spanish, French, Greek): mole, resume, attaches
- Conventions for numbers and symbols: 1995, 278, 5tet, $5, $5 bill, &c, i18n, eva1u8, 1337 5p34k
- Syntactic units/properties: *wanna* contraction, *liaison* in French

From letters to phonemes

punge; frobnicate; jall, blooth

droog, viddy, yarbles; bool, greath, swot, rull; brillig, mimsy, frumious, manxome, uffish, tulgey

Xerox, Kodak, Paradil, Gerbadigm

breath, thigh, ether; breathe, thy, either; anthill, Thomas, asthma

physics, Phish; Stephen; shepherd, loophole, haphazard, upholster

Phoenix, amoeba; Oedipus; does; shoe, canoe; hoe, aloe; Chloe; coed; coexist; Citroen/Citroën, Goethe, Roentgen/Röntgen

Knowledge required below the word level

- Part of speech: use, close
- Origin: phoenix, shoe
- Partial morphemic analysis: frobnicate
- Analogical relation to existing words: yarbles
- Letter/sound correspondences: oo, th, sh, qu, “magic E”

Automatic letter-to-phoneme conversion

- Dictionary of frequent and/or exceptional words.
- Disambiguation components.
- Fall-back strategies for out-of-vocabulary words:
  - System of hand-written rewrite rules that encode knowledge about letter/sound correspondences; or
  - Machine learning from pronunciation dictionary.

Above the word level

- Part of speech ambiguities (*convict*, *supply*).
- Word sense ambiguities (*bass*).
- Commonly solved using shallow text processing techniques (context windows, bags of words, etc.).
- Stress placement ambiguities (*Fifth Stree/Avenue*).
- Local phrasing/syntactic ambiguities difficult for shallow processing (*Pat bites down on the fish and chips a tooth.*).

Suprasegmentals

- Global intonational contours.
- Intonational phrasing.
- Placement of pitch accents.
- Assign segmental duration.

Only the last point can be modeled reliably. Predicting the placement of pitch accents from text is still a hard (and therefore interesting) problem.

Concept to speech

- Don't start from text, since that ultimately requires inferring the writer's intentions.

- For many applications, the intentions are known. This is especially true for conversational dialogue systems:
  - Status of utterance is known (response to query, clarification question, etc.).
  - Discourse context is known.
    I'm sorry – I can't book you on a flight to La Guardia. But I can get you on a (flight/bus) to (La Guardia/Newark).

Current issues

- Technology is no longer a limiting factor. It is our understanding of articulation, speech acoustics, etc. which has to improve.
- Has concatenative synthesis reached a dead end? There has been renewed interest in parametric synthesis recently, this time from an empirical perspective.
- The debate about the use of machine learning continues. Machine learning allows rapid development and porting of synthesizers to new languages and/or domains, but some argue that the quality of automatically trained components is not on a par with hand-crafted components.

- Many aspects of the prosody–meaning connection are still unclear. Robust, high-quality prosodic taggers that can predict pitch accents etc. are still far off.

- Synthetic speech generally lacks in expressiveness. Contrast the widespread use of realistic looking computer graphics for movies and video games with the almost nonexistent use of realistic sounding speech synthesis.