Text-to-Speech Synthesis

Bernd Möbius

Language Science and Technology
Saarland University

Lecture 1
May 7, 2020

Introduction:
Synthesis methods
Speech synthesis: Ambition and dilemma

- **Ambition** of speech synthesis:
  - modeling the production side of the most complex human cognitive ability

- **Dilemma** of speech synthesis:
  - emulate a human speaker or reader, without
    - world knowledge
    - language comprehension
    - speech organs
  - achieve optimal intelligibility and naturalness

- **Speech synthesis**: an impossible task!?
Human-machine dialog (1)
End-to-end synthesis (TACOTRON)

**Tacotron 2** Generating Human-like Speech from Text

**Tacotron 2** Audio samples
“The new automated ordering system has really speeded up our business. We’re losing customers faster than ever.”
Course details

- Offered for:
  - M.Sc. Language Science and Technology, LCT
  - B.Sc. Computerlinguistik
  - M.Sc./B.Sc. Computer- und Kommunikationstechnik
  - M.Sc./B.Sc. Computer Science

- Coordinates, contact:
  - Lecture, Thu 10-12, C7.4/1.17, 2 SWS, 3 LP/ECTS,
  - LSF #121407
  - http://www.coli.uni-saarland.de/~moebius/ → Teaching
  - moebius@lst.uni-saarland.de
"Speaking" statues

Devices designed by Heron of Alexandria
(1st cent. BC)

Colossi of Memnon,
Theban, Egypt
(cf. Terra X, ZDF, 6-2-2011)
Wolfgang von Kempelen (1791): speaking machine
https://www.youtube.com/watch?v=k_YUB_S6Gpo
Mechanical systems

Wolfgang von Kempelen (1770)
Mechanical systems

Kratzenstein (1779): isolated sounds

Wheatstone (1838): connected sounds

Fig. 9. Kratzenstein’s resonators for synthesis of vowel sounds. The resonators are actuated by blowing through a free, vibrating reed into the lower end. The sound is produced simply by blowing into the lower pipe without a reed.\(^1\)
Electrical systems

Dudley (1939): the Voder

Fig. 11. Schematic diagram of the electrical speaking machine Voder.14
Formant synthesis

Gunnar Fant (1953): OVE I, serial filters

John Holmes (1973): parallel filters
Formant synthesis

- Acoustic-parametric synthesis
- modeling the acoustic properties of speech sounds
Formant synthesis

- http://www.youtube.com/watch?v=wlrOKpQ6UBI

Prof. Stephen Hawking † and speech synthesizer (DECTalk DTC01)
Articulatory synthesis

- Articulatory synthesis
  - modeling components of the speech production system
    - voice source, articulators, 3D vocal tract, etc.

IP Köln (1995)

Vocal Tract Lab (2007)
http://www.vocaltractlab.de/
Synthesis methods

- Acoustic-parametric synthesis
  - a.k.a. formant synthesis
  - modeling the acoustic properties of speech sounds

- Articulatory synthesis
  - modeling components of the speech production system
    - voice source, articulators, 3D vocal tract, etc.

- Concatenative synthesis
  - uses segments of natural speech, concatenated and resequenced to synthesize the intended utterance
    - e.g. diphone synthesis, unit selection synthesis
Concatenative synthesis

- Data-based, concatenative synthesis
  - *offline*: extraction of units from recordings of natural speech
  - *online*: selection and sequential concatenation of units
- Which units are appropriate?
  - allophones? [Ger: 45]
Allophone synthesis
Concatenative synthesis

- Data-based, concatenative synthesis
  - **offline**: extraction of units from recordings of natural speech
  - **online**: selection and sequential concatenation of units
- Which units are appropriate?
  - allophones? [Ger: 45]
  - diphones? [Ger: 2025]
Diphone synthesis

Hadifax

Festival

SVOX

Bell Labs
Concatenative synthesis

- Data-based, concatenative synthesis
  - *offline*: extraction of units from recordings of natural speech
  - *online*: selection and sequential concatenation of units
- Which units are appropriate?
  - (allo)phones? [Ger: 45]
  - diphones? [Ger: 2,025]
  - triphones? [Ger: 91,125]
  - syllables? [Ger: 12,500+]
Concatenative synthesis

- Unit Selection: dynamic selection of units at synthesis run-time
  - "The best solution to the synthesizer problem is to avoid it." [Carlson & Granström, 1991]
- sound inventory: large, phonetically rich speech database
- selection of the smallest number of the longest units from a large corpus (2–10+) of recorded natural speech
- variable unit size (phones, syllables, words, ...)

B Möbius
TTS: Introduction
Unit Selection: units=words

- Target utterance: *I have time on Monday.*
- Step 1: list all candidate words for target sentence

```
I  have  time  on  Monday
I  have  time  on  Monday
I  have  on  Monday
I  have  on  Monday
I  on  Monday
```
Unit Selection: units=words

- Target utterance: *I have time on Monday.*
- Step 2: connect all units

```
I have time on Monday
```
Target utterance: I have time on Monday.

Step 3: selection of units along optimal path
Unit Selection synthesis

- best path minimizes 2 cost functions
  - **target costs:**
    how similar to target unit is the candidate unit?
  - **concatenation costs:**
    how smoothly does the unit connect to its neighbors?
Unit Selection: variable-size units

Im Gloria Kino

läuft "Der Mann ohne Vergangenheit"

... der
phrase init.
sent. med.
word single

Mann
rising accent
phrase med.
sent. med.
word single

... Vergangenheit
falling accent
low phrase boundary
phrase final
sentence final

'\textipa{d}\textipa{e}\textipa{:}\textipa{6}'}
phrase init.
sent. med.
word single

'man'
rising accent
phrase med.
sent. med.
word single

... f6
phrase med.
sent. med.
word init.

'\textipa{g}\textipa{a}'}
falling accent
phrase med.
sent. med.
word med.

N@n
phrase med.
sent. final
word med.

haIt
low phrase boundary
phrase fin.
sent. fin.
word fin.

... de:6 man
Unit Selection: demos

- example speech output from several systems:
  - CHATR (1996)
  - AT&T (2001)
  - Festival (2004)
  - SmartKom (2005)
  - Loquendo (2010)
  - BOSS (pol., 2009)
Statistical Parametric synthesis

Offline stage

- Ac. param. extraction
- Desc. feat. extraction
- Training stage
- Models
- Par. to Speech

Online stage

- Desc. feat. extraction
- Generation stage
DNN synthesis (Wavenet)
End-to-end synthesis (Tacotron)

Diagram:
- Corpus
- Signal
- Text
- Training stage
- Models
- Online
- Generation stage
- Text
- Speaker
## TTS: Audio demos

<table>
<thead>
<tr>
<th>System</th>
<th>Method</th>
<th>Interactive</th>
<th>Lang.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DECTalk</td>
<td>formant</td>
<td>no</td>
<td>Eng</td>
</tr>
<tr>
<td>Infovox</td>
<td>formant</td>
<td>no</td>
<td>Ger</td>
</tr>
<tr>
<td>IP Köln</td>
<td>articulatory</td>
<td>no</td>
<td>Ger</td>
</tr>
<tr>
<td>Hadifix</td>
<td>diphones</td>
<td>yes</td>
<td>Ger</td>
</tr>
<tr>
<td>SVOX</td>
<td>diphones</td>
<td>yes</td>
<td>Ger</td>
</tr>
<tr>
<td>Bell Labs</td>
<td>diphones</td>
<td>yes</td>
<td>Ger</td>
</tr>
<tr>
<td>Festival</td>
<td>diphones</td>
<td>yes</td>
<td>Ger</td>
</tr>
<tr>
<td>AT&amp;T</td>
<td>unit selection</td>
<td>yes</td>
<td>Eng</td>
</tr>
</tbody>
</table>

"Welcome to the Cocosda / LDC interactive TTS comparison site."

"Willkommen auf der interaktiven Seite von Cocosda und LDC für den Vergleich von Sprachsynthesesystemen."
Essential content

Speech synthesis methods

- expert systems, rule-based approaches
  - formant synthesis
  - articulatory synthesis
- concatenative approaches
  - diphone synthesis
  - unit selection synthesis
- statistical approaches
  - statistical-parametric (HMM) synthesis
  - neural network based synthesis
The tone of voice

“This phone has a special voice filter. It makes you sound honest when you discuss business, sincere when you apologize, and terminal when you call in sick.”