Speech recognition and synthesis

1. Automatic Text-To-Speech synthesis
   - Introduction
   - Computer Speech
   - Text preprocessing
   - Grapheme to Phoneme conversion
   - Morphological decomposition
   - Lexical stress and sentence accent
   - Duration
   - Intonation
   - Acoustic realization, PSOLA, MBROLA
   - Controlling TTS systems
   - Assignment
   - Bibliography

Introduction

Uses of speech synthesis by computer

- Read aloud existing text, eg, news, email and stories
- Communicate volatile data as speech, eg, weather reports, query results
- The computer part of interactive dialogs

The building block is a Text-to-Speech system that can handle standard text with a Speech Synthesis (XML) markup. The TTS system has to be able to generate acceptable speech from plain text, but can improve the quality using the markup tags

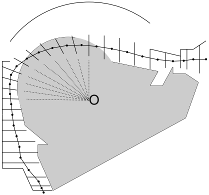
Computer Speech: Generating the sound

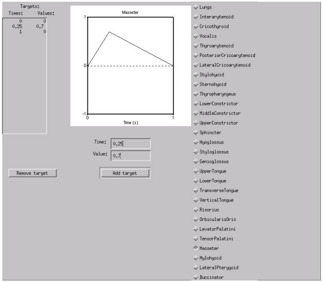
Speech Synthesizers can be classified on the way they generate speech sounds. This determines the type, and amount, of data that have to be collected.

Speech Synthesis

- Articulatory models
- Rules (formant synthesis)
- Diphone concatenation
- Unit selection

Computer Speech: Articulatory models

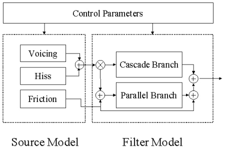


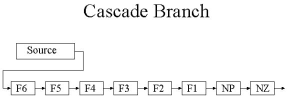


Characteristics (/ɛrə/ from Praat) [Boersma(1998)]

- Quantitative Source-Filter model of vocal tract
- Solve Navier-Stokes equations for air-flow
- Needs hard-to-get articulatory data

Computer Speech: Rule, or formant, based synthesis



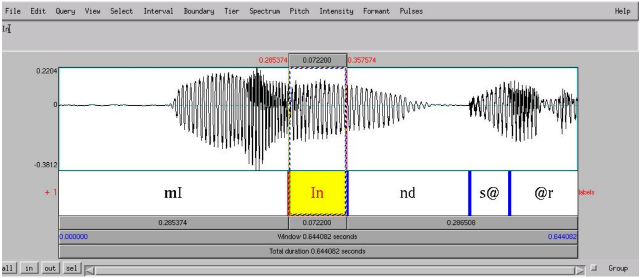


Klatt synthesizer [Sproat(), SRL()]

Characteristics (YorkTalk [Möhler(2005)])

- Recreate sounds using source and resonances
- Model formant tracks by rules
- Endless tuning, no data driven modelling possible

Computer Speech: Diphone synthesis



Characteristics (Spengi, Philips/IPO [Möhler(2005)])

- Concatenative synthesis: Glue phoneme-phoneme transitions
- Good quality, but requires all phoneme combinations to be present
- Sound encoding must allow intonation changes

Computer Speech: Nextens diphone synthesis

'Nederlandse Extensie voor Tekst naar Spraak' or 'Dutch Extension for Text to Speech' example

Nextens runs on top of Festival [Nextens(2003), Festvox(2005)]

- New Dutch voices in Festival
- Nintens GUI (io, commandline in Festival)
- Available for non-commercial use (*not* Open Source)
- Developed at the Radboud University and the University of Tilburg (Joop Kerkhof, Erwin Marsi, and others)

Computer Speech: Non-uniform unit selection

Generalize diphone synthesis to use larger, non-uniform, units like: diphones, multiphones (clusters), demi-syllables, syllables, words, and short phrases

Characteristics (Festival [Black and Lenzo(2003a)])

- Requires large annotated speech corpora ($\sim$ GByte range)
- Corpus must be well annotated and searchable
- Efficient statistical search algorithms to optimize unit selection based on prosody and concatenation costs
- More speech in corpus $\Rightarrow$ Better synthesis
- But also $\Rightarrow$ More work to find the best combination

Computer Speech: Text-to-Speech

Text in Speech out: Processing "steps"

- Text normalization
- Grapheme Phoneme conversion
- Accent placement
- Duration generation
- Intonation generation
- Speech Generation

Text preprocessing: Normalize texts

Text should contain only pronounceable tokens

- Abbreviations
- Dates
- Times
- Telephone numbers
- Money
- Street Addresses
- General numbers
- Special characters

Join Kerry Stratton & his guest chamber orchestra as they bring the music of the Italian Maestro to life on our stage. Tickets $46.00

5 Easy Ways to Order Tickets

A Visit our Box Office (map) Mon through Sat, 11:00 a.m. to 6:00 p.m. Summer Hours: July 4 to Sept 2, 2005 - 11:00 a.m. to 4:30 p.m.

B Call our Box Office at 905-305-SHOW (7469) or Toll Free at 1-866-768-8801 (not available in 416/647 area codes).

C Fax your order form to 905-415-7538.

D Return your completed order form with payment to: Markham Theatre, 171 Town Centre Blvd., Markham, ON, L3R 8G5.

E Online ticket sales are currently only available for Single Tickets beginning September 13, 2005.

Grapheme to Phoneme conversion: By dictionary and rules

Tokenize the text and look up the words in a pronunciation dictionary. If not found, use rules

- Dictionary entries: `("dictionary" nil (d ih1 k sh ax n eh1 r iy0))`
- Rules: `( LC [ alpha ] RC => beta )`
  - `( # [ c h ] r => k )` "ch" word initially in English
  - `( # [ c h ] => ch )` "ch" word initially in English
  - `( [ c ] => k )` default rule for "c"

After all words have been converted, there is a second pass to catch changes at word boundaries and general effects of running speech

Morphological decomposition: Out-of-Vocabulary words

Compound words and other words not in the dictionary are common

- Compound words are common in many languages, eg, German, Dutch, Finnish, Turkish
- Compound word consist of lexical words that are connected with infixes, eg, *-s-* and surrounded by affixes, eg, *a-*, *in-*, *-ed*
- Compounding or affixes can change the pronunciation and orthography of a word component, eg, *Kunst* → *Künst+ler*)
- Parse complex words with a statistical weighted finite-state transducer (WFST) [Möbius(1998)]

Morphological decomposition: German examples

Unerfindlichkeitsunterstellung
“allegation of incomprehensibility”

WFST states: **START PREFIX ROOT INFIX SUFFIX END**

German decompositions [Möbius(1998)]

- *gener+ator* “generator”
- *honor+ar* “fee”
- *Schwind+sucht* “consumption”
- *Arbeit+s+amt* “employment agency”
- *Sonne+n+schein* “sunshine”
- *Un+er+find+lich+keit+s+unter+stel+lung* “allegation of incomprehensibility”

Morphological decomposition: Decomposition

noun forming prefixes					noun forming suffixes				
	N	Ftyp	n1	P		N	Ftyp	n1	P
*schwind-	1	1	1	1	-chen	1140	255	42	0.0368
vor-	104	14	2	0.0192	-ling	278	20	3	0.0108
be-	600	6	1	0.0017	-heit	604	7	2	0.0033
ge-	8125	164	10	0.0012	-schaft	11109	171	15	0.0014
semi-	12	3	0	0.0000	-ett	51	1	0	0.0000
adjective forming prefixes					adjective forming suffixes				
	N	Ftyp	n1	P		N	Ftyp	n1	P
*wiss-	1	1	1	1	-haft	1107	102	14	0.0126
ur-	108	10	1	0.0093	-voll	132	6	1	0.0076
un-	10010	601	64	0.0064	-är	502	17	1	0.0020
in-	219	49	1	0.0046	-lich	32168	569	51	0.0016
aller-	42	2	0	0.0000	-ig	3966	40	3	0.0008
verb forming prefixes					verb forming suffixes				
	N	Ftyp	n1	P		N	Ftyp	n1	P
weit-	94	11	3	0.0318	-er	65	24	5	0.0769
vor-	1401	31	4	0.0029	-el	1197	86	11	0.0092
ent-	13007	200	18	0.0014	-isier	1019	75	7	0.0069
ver-	53899	930	71	0.0013					
dar-	1071	6	1	0.0009					

Use a dictionary and include a morphological compound list with pronunciations. [Möbius(1998)]

Lexical stress and sentence accent: Prominence

Some words are more prominent than others. They are:

- Accented, i.e. carry a pitch movement
- Longer
- Louder
- Less reduced

Prominence is determined by

- Word type, function words are almost never prominent
- Word frequency, rare words are prominent more often
- New information is prominent, given is not
- Not too many prominent words in a row

There are rules for assigning prominence, but they need good POS tagging. Just accenting every content words works too

Lexical stress and sentence accent: Syllable stress

Some syllables are more prominent than others. They are:

- Longer
- Louder
- Less reduced

Syllable stress is determined by

- The lexicon or language (lexical/fixed stress positions)
- Syllable weight, "heavy" syllables are stressed
- No stressed syllables in a row
- Informative syllables are stressed

Mostly, you can get away with either the lexicon, or fixed positions. Syllable stress shifts in compound words. Morphological decomposition gives rules for these shifts

Lexical stress and sentence accent: Phrase boundaries

Intonation covers utterances of a few words at a time (around 5-7). Breaking up sentences at acceptable places is difficult

- Use punctuation
- Guess boundaries on POS tags (HMM style)
- Do a partial syntactic parse and use phrases

In general, it is difficult to go beyond punctuation and some simple heuristics without syntactic parsing

Duration

Phoneme duration is determined by:

- Phoneme identity
- Surrounding phonemes
- Sentence accent/prominence
- Syllable stress
- Syllable length and position (Onset, Coda)
- Word length
- Phrase/sentence boundary position
- ...

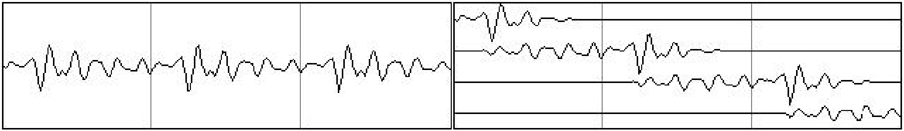
These factors are used to construct statistical models from annotated speech corpora. Golden standard is Correlation and Regression Trees (CART). But many other statistical methods are used

Intonation

With the durations known, the pitch contour can be calculated

- Speaker and style determine the pitch range
- Give each accent a pitch movement shape and size
- Assign each vowel its target F_0 value
- Interpolate the values into a valid contour
- Assign each phoneme it's F_0 values

Acoustic realization, PSOLA, MBROLA



Multi Band Excitation (Time Domain) Pitch Synchronous Overlap Add [MBROLA(2005)]

- Mark all pitch periods (blue pulses in *Praat*)
- Fixed periods for voiceless speech
- Window speech around each mark
- To lengthen/shorten a sound, reduplicate/delete periods
- To increase/decrease F_0, shorten/lengthen times between periods
- Synthesize sound by summing windowed periods at their correct time position

Controlling TTS systems: XML standards for speech synthesis

VoiceXML: Control of web based dialog applications

- SRGS: Speech Recognition Grammar Specification
- SSML: Speech Synthesis ML
- CCXML: Call Control XML
- NLSML: Natural Language Semantics ML for the Speech Interface Framework
- SISR: Semantic Interpretation for Speech Recognition
- SCXML: State Chart XML, State Machine Notation for Control Abstraction
- PLS: Pronunciation Lexicon Specification
- ECMAScript/JavaScript

▸ XML standards

Controlling TTS systems: SSML

Speech Synthesis Markup Language

```
<speak version="1.0" xml:lang="en-GB">
          Hello, how are you?
        <prosody rate ="x-fast" >
              This sentence is spoken fast
        </prosody>
        <prosody pitch = "x-low">
               This sentence is spoken low pitch
        </prosody>
        <prosody pitch = "medium">
               This sentence is spoken medium pitch
        </prosody>
        <prosody pitch = "x-high">
                This sentence is spoken high pitch
         </prosody>
         <prosody rate = "fast">
                    This sentence is spoken fast
         </prosody>
         <emphasis level = "strong">
             This sentence is spoken with stress
         </emphasis>
</speak>
```

Controlling TTS systems: eSpeak formant synthesis

eSpeak can be used both for stand-alone formant synthesis and as a front end for Mbrola voices

- *espeak 'text to say' -w test.wav* ⇒ standard example
- *espeak -v mb-en1 'text to say'* |
  mbrola -e /usr/share/mbrola/en1 - test.wav ⇒ Mbrola example
- Free Software (GPL)
- Supports SSML (partially, eg, not <emphasis>)
- Many languages, eg, Dutch, Latin, Mandarin, and Cantonese

Assignment: Week 5 TTS

Introduction to eSpeak

- Install eSpeak from `http://espeak.sourceforge.net/`
- Try out short texts using several voices and languages
- Inspect phoneme conversions with *espeak -x*
- Try to improve synthesis by hand-crafting phoneme input using, eg, *espeak -v en "[[D,Is Iz sVm f@n'EtIk t'Ekst 'InpUt]]"*
- Try out SSML on eSpeak using, eg, *espeak -m -f example.ssml -w example.wav*
- Describe the differences in quality
- More on Blackboard...

Further Reading I

Christina L. Bennett.
Large Scale Evaluation of Corpus-based Synthesizers: Results and Lessons from the Blizzard Challenge 2005.
In *Proceedings of Interspeech 2005, Lisboa, Portugal*, September 2005.
URL http://festvox.org/blizzard/bc2005/IS052023.PDF.

Alan W. Black and Kevin A. Lenzo.
Festvox.
Web, 2003a.
URL http://festvox.org/.
Festival speech synthesis.

Alan W. Black and Kevin A. Lenzo.
Building Synthetic Voices.
Festvox, 2 January 2003b.
URL http://festvox.org/bsv/.
Published on the festvox website.

Alan W. Black and Keiichi Tokuda.
The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common datasets.
In *Proceedings of Interspeech 2005, Lisboa, Portugal*, September 2005.
URL http://festvox.org/blizzard/bc2005/IS051946.PDF.

P. Boersma.
Praat, a system for doing phonetics by computer.
Glot International, 5:341–345, 2001.
URL http://www.Praat.org/.

Further Reading II

P. Boersma and D. Weenink.
Praat 4.2: doing phonetics by computer.
Computer program: http://www.Praat.org/, 2004.
URL http://www.Praat.org/.

Paulus Petrus Gerardus Boersma.
Functional Phonology: Formalizing the Interactions between Articulatory and Perceptual Drives.
PhD thesis, University of Amsterdam, September 1998.
URL http://www.fon.hum.uva.nl/paul/papers/funphon.pdf.

Murtaza Bulut, Shrikanth S. Narayanan, and Ann K. Syrdal.
Expressive speech synthesis using a concatenative synthesizer.
In *Proceedings of ICSLP 2002, Denver, COLORADO*, September 2002.
URL http://www.research.att.com/projects/tts/papers/2002_ICSLP/expressive.pdf.

Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors.
Survey of the State of the Art in Human Language Technology.
Cambridge University Press, 1996.
URL http://cslu.cse.ogi.edu/HLTsurvey/.
ISBN 0-521-59277-1.

Festvox.
Festvox.
Web, 2005.
URL http://www.festvox.org/.

Further Reading III

FSF.
GNU General Public License.
Web, June 1991.
URL http://www.gnu.org/licenses/gpl.html.

MBROLA.
The MBROLA Project.
Web, 2005.
URL http://tcts.fpms.ac.be/synthesis/.
Synthesis.

Bernd Möbius.
word and syllable models for german text-to-speech synthesis.
In Mike Edgington, editor, *Third ESCA/COCOSDA Workshop on SPEECH SYNTHESIS*, 26 November 1998.
URL http://www.slt.atr.co.jp/cocosda/jenolan/Proc/r06/r06.pdf.

Gregor Möhler.
Examples of Synthesized Speech.
Web, 2005.
URL http://www.ims.uni-stuttgart.de/~moehler/synthspeech/.
Good web-site with many examples.

Nextens.
NeXTeNS: Open Source Text-to-Speech for Dutch.
Web, 2003.
URL http://nextens.uvt.nl/index.html.

Further Reading IV

Louis C.W. Pols, Jan P.H. van Santen, Masanobu Abe, Alan Black, David House, Mark Liberman, and Zhibiao Wu.
Easy access via a TTS website to mono- and multilingual text-to-speech systems.
In *Proceedings of the Third ESCA/COCOSDA Workshop on SPEECH SYNTHESIS*, November 1998.

Project Gutenberg.
Project gutenberg free ebook library.
Web, 2005.
URL http://www.gutenberg.org/.

Richard Sproat.
ECE 598: Sp eech Synthesis.
Web.
URL http://catarina.ai.uiuc.edu/ECE598/Lectures/klattlpc.pdf.

SRL.
Synthesis of Speech.
Web.
URL http://wagstaff.asel.udel.edu/speech/tutorials/synthesis/.
Speech Research Lab, A.I. duPont hospital for children and University of Delaware.

Appendix A: XML standards in Speech Technology

XML standards in Speech Technology

VoiceXML: Control of web based dialog applications

- SRGS: Speech Recognition Grammar Specification
- SSML: Speech Synthesis ML
- CCXML: Call Control XML
- NLSML: Natural Language Semantics ML for the Speech Interface Framework
- SISR: Semantic Interpretation for Speech Recognition
- SCXML: State Chart XML, State Machine Notation for Control Abstraction
- PLS: Pronunciation Lexicon Specification
- ECMAScript/JavaScript

Back to TTS control

XML standards in Speech Technology: VoiceXML

Application independent spoken dialog control

```
<?xml version="1.0"?>
<vxml version="2.0">
<menu>
  <prompt>
    Say one of: <enumerate/>
  </prompt>
  <choice next="http://www.sports.example/start.vxml">
     Sports
  </choice>
  <choice next="http://www.weather.example/intro.vxml">
     Weather
  </choice>
  <choice next="http://www.news.example/news.vxml">
     News
  </choice>
  <noinput>Please say one of <enumerate/></noinput>
</menu>
</vxml>
```

Back to TTS control

XML standards in Speech Technology: SRGS

Speech Recognition Grammar Specification

```
<grammar root="buyShirt" xml:lang="en-US">
    <rule id="buyShirt" scope="public">
        <item>
           Get me a <ruleref uri="\#ruleColors" />
           shirt and a <ruleref uri="\#ruleColors"/>
           tie</item>
    </rule>

    <rule id="ruleColors" scope="public">
         <one-of>
            <item>red</item>
            <item>white</item>
            <item>green</item>
        </one-of>
    </rule>
</grammar>
```

Back to TTS control

XML standards in Speech Technology: SSML

Speech Synthesis Markup Language

```
<speak version="1.0" xml:lang="en-GB">
           Hello, how are you?
         <prosody rate ="x-fast" >
               This sentence is spoken fast
         </prosody>
         <prosody pitch = "x-low">
                This sentence is spoken low pitch
         </prosody>
         <prosody pitch = "medium">
                This sentence is spoken medium pitch
         </prosody>
         <prosody pitch = "x-high">
                 This sentence is spoken high pitch
          </prosody>
          <prosody rate = "fast">
                    This sentence is spoken fast
          </prosody>
          <emphasis level = "strong">
              This sentence is spoken with stress
          </emphasis>
</speak>
```

XML standards in Speech Technology: PLS

Pronunciation Lexicon Specification

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0"
         xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
      alphabet="ipa" xml:lang="en-US">
  <lexeme>
    <grapheme>lead</grapheme>
    <alias>led</alias>
    <phoneme prefer="true">li:d</phoneme>
  </lexeme>
  <lexeme>
    <grapheme>lead</grapheme>
    <phoneme prefer="true">led</phoneme>
    <phoneme>li:d</phoneme>
  </lexeme>
</lexicon>
```

Back to TTS control

XML standards in Speech Technology: CCXML

Voice Browser Call Control

```
<ccxml version="1.0">
  <eventhandler>
    <transition event="connection.CONNECTION_ALERTING"
                name="evt">
      <log expr="'The caller ID is ' + evt.callerid + '.'"/>
      <if cond="evt.callerid == '8315551234'">
        <accept/>
      <else/>
        <reject/>
      </if>
    </transition>
    <transition event="connection.CONNECTION_CONNECTED">
      <log expr=
           "'Call was answered. We are going to start a dialog.'"/>
      <dialogstart src="'start.vxml'"/>
    </transition>
  </eventhandler>
</ccxml>
```

▸ Back to TTS control

XML standards in Speech Technology: NLSML

Natural Language Semantics Markup Language for the Speech Interface Framework

```
<interpretation grammar="http://generalCommandsGrammar"
 xmlns:xf="http://www.w3.org/2000/xforms">
  <xf:model>
    <group name="command"/>
      <string name="action"/>
      <string name="doer"/>
    </group>
  </xf:model>
  <xf:instance>
    <myApp:command>
    <action>reduce speech rate</action>
    <doer>system</doer>
    </myApp:command>
  </xf:instance>
  <input mode="speech">slow down</input>
</interpretation>
```

Back to TTS control

XML standards in Speech Technology: SISR

Semantic Interpretation for Speech Recognition

```
<rule id="sub_hundred_thousand">
    <ruleref uri="#sub_hundred"/>
    <tag>out = (1000 * rules.sub_hundred)</tag>

    thousand
    <item repeat="0-1">
        <item repeat="0-1">and</item>
        <ruleref uri="#sub_thousand"/>
        <tag>out += rules.sub_thousand;</tag>
    </item>

</rule>
```

▸ Back to TTS control

XML standards in Speech Technology: SCXML

State Machine Notation for Control Abstraction

```
<scxml xmlns="http://www.w3.org/2005/07/scxml" version="1.0"
       initalstate="S1">
  <state id="S1">
    <datamodel>
      <data name="rand">
    </datamodel>
    <onentry>
      <assign name="rand" expr="Math.random()"/>
    </onentry>
    <transition event="E1" cond="rand <= 0.3" target="S2"/>
    <transition event="E1" cond="rand > 0.3" target="S3">
  </state>
  <state id="S2"/>
  <state id="S3"/>
</scxml>
```

▸ Back to TTS control

XML standards in Speech Technology: ECMAScript/JavaScript

JavaScript is the procedural language of VoiceXML

```
<script>
   var n = 0;
   for (var i = 0; i < 3; i++) {
   n += i;
   <prompt> You have <value expr="n"/> copies</prompt>
}
</script>
```

▸ Back to TTS control

Copyright License