Corpus-Based Speech Synthesis
From diphones to units to states

Thierry Dutoit
thierry.dutoit@umons.ac.be

Grenoble, 2014

UMONS - MULTITEL- ACAPELA
-a threesome in Mons, Belgium-

• 1983... UMONS / TCTS Lab
  Applied research on Speech Processing

• 1997... ACAPELA Group
  Spin-off company : Products

• 2000 : MULTITEL ASBL
  Applicative Research Center

⇒ ~25 people on speech in Mons
The numediart Institute for New Media Art Technology of UMONS, is the R&D effort towards Mons2015, EU capital of culture.

Outline

- Corpus-based speech synthesis (a personal view)
  - From diphones to units to states: A personal view of the past 20 years in speech synthesis

- Recent developments around HMM synthesis at UMONS:
  - Synthesizing hypo and hyper articulated speech using HMMs
  - Performative speech synthesis (from vowels to continuous speech using HMMs) ...
  - Applying HMM synthesis to laughter synthesis (hopefully), and to gait synthesis
Prof. Thierry Dutout

Supra-Segmental

Coarticulation !!!
Von Kempelen’s talking machine (1791)

(J.S. Liénard, LIMSI)

(Articulatory synthesis)

- Work by (60’s-70’s):
  K. Stevens, G. Fant, P. Mermelstein, R. Carré (GNUSpeech), Perkell, S. Maeda, J. Shroeter & M. Sondhi...

- More recently (2000-...):
  O. Engwall, S. Fels (ArtiSynth), Birkholz (VocalTractLab) and Kröger, A. Alwan & S. Narayanan (MRI), Badin et al., Hueber et al. (Articulatory-to-Acoustic Mapping), ...
Omer Dudley’s Voder (Bell Labs, 1936)

Noise Source
Oscillator
Resonance Control
Amplifier

Energy switch
Pitch control

Voder Console
Keyboard

0. John Holmes’ formant synthesizer (1964)

Rule-based Synthesis

Haskins Labs (1968)
InfoVox (1983-95)
DecTalk (1983)
1. Diphone concatenation (1977)

1500 diphones
3' of speech
5Mbytes

Cnet-PSOLA (1990) (Phonétiseur : TOPH – PhD V. Aubergé
UMONS – MBROLA (1994)

LP Synthesis (1977)

Olive(1980)
PSOLA (1988)

Cnet (1990) - Phonétiseur TOP (V. Aubergé)

The MBROLA project (1994)

UMONS – MBROLA (1994)
Up from diphones! (90’s)

- **Polyphones** = diphones + some triphones + some quadriphones
  
  (V. Aubergé’s polysons 😊)
  
  Ex: France Telecom 93:
  
  1300 diphones, 750 triphones, 288 quadriphones

- **Demisyllables** (CI-V or V-Cf)
  
  Ex: HADIFIX (HALf-syllables, DIphones, and sufFIXes) for German (2000 units)

- **Context-Oriented Clustering (1994)**
  
  = corpus-based, automatic unit set design algorithm

---

Up From Diphones!

**Context-Oriented Clustering**

*(Flavor 1) (Nakajima, 1994)*

Detect and **explain** allophones?

= “phonological vector quantization”

**OUTPUT = tree** (not the leaves!)
1996: Is there a (bright) future for TTS R&D?

2. Unit selection (1996)

"Take the best, modify the least"

Diphone-based synthesis

1500 diphones
3' of speech
5Mbytes

Prof. Thierry Dutoit
FACULTÉ POLYTECHNIQUE DE MONS
2. Unit selection (1996)

"Take the best, modify the least"

Unit selection-based synthesis

Very Large Corpus

50000 diphones
1h of speech
150Mbytes

(1998)

(ACAPELA, 2005)

2. Automatic unit selection

sent: "To be ...
phonet: _ t U b i: ...
stress: ^ ...
tone: l H ...
dur: 210 40 55 80 198 ...
f0: ...

formants: T A R G E T

sent: "... to bear."
phonet: t U b E@ ...
stress: ^ ...
tone: l L ...
dur: 150 50 85 90 150 ...
f0: ...

Very Large corpus

Formants: T A R G E T

Target j

Target cost tc(j,i)

Unit i
2. Automatic unit selection

- **Unit i**
  - Concatenation cost \( cc(i, i+1) \)
  - Sent: "... to bear."
  - Phonet: \( t \ U \ b \ E@ \ ... \)
  - Stress: \( ^ \ ... \)
  - Tone: \( L ... \)
  - Dur: 150 50 85 90 150 ...
  - F0:

- **Unit I+1**
  - Sent: "... teletubbies ..."
  - Phonet: \( t @ b i: s \ ... \)
  - Stress: \( ^ \ ... \)
  - Tone: \( L ... \)
  - Dur: 80 95 90 130 ...
  - F0:

- **Formants:**

- **Very Large corpus**

2. Automatic unit selection

- **Unit i-1**
  - Concatenation cost \( cc(i-1, i) \)
  - = 0 in case of successive units

- **Unit i**
  - Sent: "... to bear."
  - Phonet: \( t \ U \ b \ E@ \ ... \)
  - Stress: \( ^ \ ... \)
  - Tone: \( L ... \)
  - Dur: 150 50 85 90 150 ...
  - F0:

- **Unit I+1**
  - Sent: "... to bear."
  - Phonet: \( t \ U \ b \ E@ \ ... \)
  - Stress: \( ^ \ ... \)
  - Tone: \( L ... \)
  - Dur: 150 50 85 90 150 ...
  - F0:
Unit Selection: When the dream started to come true

2001, *A Space Odyssey* (HAL voice)  

---

2. Unit selection (1996)

- **Original approach - 1996**
  Searching the whole database for the sequence of units which minimizes target + concatenation costs

- **Context-oriented clustering (flavor 2) - 1998**
  = speed-up method
  Output = tree + leaves!
2005: Is there a (bright) future for TTS R&D?

3. Tokuda’s HTS synthesizer (2002...)

Context-oriented clustering (flavor 3) – keep the tree, model the leaves

Kobayashi et al. 2004
Umons/ACAPELA 2009
HMM speech synthesis: the best of all worlds (?)

Being based on a *parametric* source-filter model, *statistically* trained on a large speech database, HMM synthesis potentially exhibits:

- the **coverage** of unit-selection synthesis
- the **compacity** of rule-based synthesis (<1MB)
- the **flexibility** of LPC-based diphone synthesis

The great return of source/filter modeling


Towards performative speech synthesis

MAGE, Astriniani, d’Alessandro et al, P3S, 2010

The great return of source/filter modeling

A great sandbox for speaker/voice quality conversion

EMIME’s 1000 voices, Yamagishi et al, TSALP, 2009, Hypo-hyper articulation, Picart et al, SSW, 2010

HMM synthesis of anything?


The great return of source/filter modeling

PhD of B. Bozkurt, 2006

IEEE Sig Proc Letters 2005 (344-347)

PhD of Thomas Drugman, 2011

Interspeech, 2009, IEEE TASLP, March 2012 (968-981)
ZZT - Mixed-Phase model of speech
Applying the mixed-phase model to create synthetic signals

ZZT - All-zero representation of speech
Synthetic mixed-phase speech
ZZT - Zero-decomposition for source-tract separation

1. SPECTRUM DATA
2. PCA and voice detection
3. GCI detection
4. GCI synchronous windowing

ZZT

- Z-Transform
- Calculation of zeros

Classification of zeros according to radius

- r < 1: inside the UC
- r ≥ 1: outside the UC

DFT calculation from zeros vocal tract dominated spectrum
DFT calculation from zeros source dominated spectrum

ZZT - Zero-decomposition for source-tract separation

Real speech

Original windowed speech

Original amp. spectrum

Reconstructed glottal excitation

Reconstructed vocal tract response

Reconstructed glottal amp. spectrum

Reconstructed tract transfer function

Copy-Synth

Noise excited tract

Zero-decomposition
ZZT - Zero-decomposition for source-tract separation

- ZZT not tested on HMM synthesis
- J. Cabral’s (CTSR, S. Renals) mitigated results on a similar glottal approach...

- ZZT remains an original source-filter decomposition method
- Shown to be equivalent to Complex Cepstrum liftering (Drugman et al., SPECOM 2011)

DSM - Deterministic/Stochastic model of LPC residual

Deterministic component of the excitation

Stochastic component of the excitation
DSM in HMM Synthesis

<table>
<thead>
<tr>
<th>Speech Experts</th>
<th>Male Speaker</th>
<th>Female Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSM vs Pulse</td>
<td>1.205 ± 0.198</td>
<td>1.241 ± 0.209</td>
</tr>
<tr>
<td>DSM vs STRAIGHT</td>
<td>0.167 ± 0.217</td>
<td>0.037 ± 0.197</td>
</tr>
<tr>
<td>Naive listeners</td>
<td>Male Speaker</td>
<td>Female Speaker</td>
</tr>
<tr>
<td>DSM vs Pulse</td>
<td>0.75 ± 0.176</td>
<td>0.722 ± 0.188</td>
</tr>
<tr>
<td>DSM vs STRAIGHT</td>
<td>-0.010 ± 0.164</td>
<td>-0.072 ± 0.201</td>
</tr>
</tbody>
</table>

Table 12.2 - Average CMOS scores together with their 95% confidence intervals, for both speech experts and naive listeners.

A great sandbox for speaker/voice quality conversion

PhD of Th. Drugman, 2011

PhD of B. Picart, 2013
Speech synthesis workshop SSW7, 2010
Interspeech 2011, pp. 1797-1800
**DSM and speaker identity**

Two glottal signatures from the DSM

![Graph showing identification rate](image)

<table>
<thead>
<tr>
<th>Method</th>
<th>168 speakers</th>
<th>630 speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDGF</td>
<td>28.65</td>
<td>/</td>
</tr>
<tr>
<td>MCGF</td>
<td>4.70</td>
<td>/</td>
</tr>
<tr>
<td>VSCC</td>
<td>5.06</td>
<td>12.95</td>
</tr>
<tr>
<td>Using only the eigenresidual</td>
<td>8.76</td>
<td>17.14</td>
</tr>
<tr>
<td>Using only the energy envelope</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Using both glottal signatures</td>
<td>1.98</td>
<td>3.65</td>
</tr>
</tbody>
</table>

**DSM and voice quality**

![Graphs showing tilt spectral and maximum voiced frequency](image)

Data: de7 (Marc Schroeder)
DSM and voice quality

Modification of 3 modal voices, using HMM synthesis (2 men, 1 woman)
9 sentences each, with automatic conversion to « tense » and « soft »
Subjective tests (10 experts listeners)

<table>
<thead>
<tr>
<th></th>
<th>Effort vocal</th>
<th>Scores MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modal vers Doux</td>
<td>36.11 ± 2.60</td>
<td>3.189 ± 0.145</td>
</tr>
<tr>
<td>Modal</td>
<td>52.89 ± 2.82</td>
<td>3.017 ± 0.147</td>
</tr>
<tr>
<td>Modal vers Tendu</td>
<td>72.11 ± 2.60</td>
<td>2.606 ± 0.146</td>
</tr>
</tbody>
</table>

Results (average, 95% confidence)

How does a speaker accommodate to speaking conditions?

1359 sentences, by a male actor

Cf LISTA (Listening Talker) EU/FET project
How does a speaker accommodate to speaking conditions?

Vocal tract  Source

---

2. Adapting a normal voice using a subset of hypo-hyper sentences

<table>
<thead>
<tr>
<th>Bases de données</th>
<th>Coefficients MGC</th>
<th>Intra-speaker adaptation (CMLLR)</th>
<th>Hyperarticulated HMM-based speech synthesizer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyper-articulated (1220)</td>
<td>α = 0.42, γ = 0, Ordre de l’analyse MGC = 24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral (1220)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypo-articulated (1220)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coefficients MGC
α = 0.42, γ = 0
Ordre de l’analyse MGC = 24

Quality of synthetic speech using a variable number of adaptation sentences (50 to 1200)?
### 3. Continuous control of the degree of articulation

#### Bases de données

- **Hyperarticulated** (1220)
- **Neutral** (1220)
- **Hypoarticulated** (1220)

#### Continuous Control (slider)

<table>
<thead>
<tr>
<th>Hypoarticulation</th>
<th>Neutre</th>
</tr>
</thead>
<tbody>
<tr>
<td>- 1.5</td>
<td>- 1.25</td>
</tr>
<tr>
<td>Exemple 1</td>
<td></td>
</tr>
<tr>
<td>Exemple 2</td>
<td><img src="example_icon.png" alt="example_icon" /></td>
</tr>
</tbody>
</table>

#### Hyperarticulation

<table>
<thead>
<tr>
<th>Hyperarticulation</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>1</th>
<th>1.25</th>
<th>1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exemple 1</td>
<td><img src="example_icon.png" alt="example_icon" /></td>
<td><img src="example_icon.png" alt="example_icon" /></td>
<td><img src="example_icon.png" alt="example_icon" /></td>
<td><img src="example_icon.png" alt="example_icon" /></td>
<td><img src="example_icon.png" alt="example_icon" /></td>
<td><img src="example_icon.png" alt="example_icon" /></td>
</tr>
<tr>
<td>Exemple 2</td>
<td><img src="example_icon.png" alt="example_icon" /></td>
<td><img src="example_icon.png" alt="example_icon" /></td>
<td><img src="example_icon.png" alt="example_icon" /></td>
<td><img src="example_icon.png" alt="example_icon" /></td>
<td><img src="example_icon.png" alt="example_icon" /></td>
<td><img src="example_icon.png" alt="example_icon" /></td>
</tr>
</tbody>
</table>
Performative Speech Synthesis

PhD of N. d’Alessandro, 2009

PhD of M. Astrinaki, 2014

P3S Workshop, Vancouver, 2011
LISTA Workshop, 2012

Omer Dudley’s Voder
(Bell Labs, 1936)
The HandSketch (2007)
HTS vs. pHTS (performative HTS)

Synthesis part in HTS
- Targeted Text to be converted into speech
- All labels containing all contextual information
- Decision Trees
  - Prosody
  - Excitation Signal Generation
- Synthesis Filter (MLSA)
- Synthesized Speech

Synthesis part in pHTS
- Label 1 → Label 2 → Label 3 → ... → N
- Decision Trees
- Generate parameters from HMMs
- Excitation Signal Generation
- Synthesis Filter (MLSA)
- Silence, samples 1 → samples 2 → ... → N

Testing pHTS with limited phoneme lookahead

Imposed F0 and durations

<table>
<thead>
<tr>
<th>mel-cepstral distortion in dB</th>
<th>spectral distortion in dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>male voice</td>
<td>female voice</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>3.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Prof. Thierry Dutoit

Faculté Polytechnique de Mons
Mage 2.0 :: current applications

MAGE & FaceOSC
application for reactive
voice manipulation through
facial expression

MAGE & Guitar
reactive control of prosody and
context of a synthetic voice
based on guitar playing
techniques

MAGE & HandSketch
reactive control of pitch, speed, duration
and quality of a voice through hand
gestures

MAGE & Kinect
reactive prosody control of
synthetic voices through body
gestures

MAGE & Accents
application for reactively
controlling, combining and
manipulating regional
specificities in speech
synthesis, here accent,
through simple map

EU/FET ILHAIRE PROJECT
http://www.ilhaire.eu/

UMONS (lead), ParisTech, UZH, Supelec Metz, University of
Belfast, U. Augsburg, UCL London, U Genova,
La Cantoche SA
PhD of J. Urbain (ongoing)
Structure of laughter

Laughter synthesis

- Only limited research so far

- [2004] Trouvain & Schröder
- [2006] Campbell
- [2007] Lasarcyk & Trouvain
- [2007] Sundaram & Narayanan
- [2007] Lasarcyk & Trouvain
- [2008] Beller
- [2013] Thati et al.
- [2013] Urbain et al.
From labels to laughter

AVL C

hahahahaha

HMM synthesis

Manila laughter workshop - HMM-based laughter synthesis

From intensity to laughter

Data base

Ehahahahaha

Generation

HMM synthesis

Manila laughter workshop - HMM-based laughter synthesis

POLYTECHNIQUE DE MONS
Visual laughter synthesis

PhD of H. Cakmak (ongoing)
And now, something completely different
Explicit Mapping: Interactive gait synthesis

What HMMs *cannot* (easily) do

- Produce non-linear modifications in the HMM parameter space
  - Ex: speech-smiled speech-laughter

- Synthetic speech quality is still limited...
  - Refine excitation?
2013: Is there a (bright) future for TTS R&D?

Deep Neural Nets?