From speech synthesis to virtual speaker: LAIPTTS → Mimic → Poser

Eric Keller, LAIP
IMM Lettres
University of Lausanne

Demos available at:
http://www.unil.ch/jimm/page22063.html
(scroll to **Work in Progress** at the bottom)

This is a Pilot Project…

…to examine

- the mechanics
- the quality attainable
- the difficulties
- the limits

of the LAIPTTS ➔Mimic ➔Poser link

Methodology – 1
Character construction

- **Poser 6** (CuriousLabs), the **Michael figure** (DAZ) and **hair by Kozaburo** were used to build a male character.
- The **character's default facial features** were manipulated, based on external user input, to simulate a *character who could be a computer science student* at the University of Lausanne.

Methodology – 2
Character Concretization

- The **completed character** was given a…
  - *name:* Giulio Lorenzo
  - *origin:* the Ticino canton in Switzerland
  - *age:* 23
  - *level of advancement:* third-year student
  - *marital status:* in a steady relationship with a Ticino woman who cuts his hair every once in awhile (but not often enough)
- The **situation** where he was "video-taped" was identified:
  - **Giulio was working away at his virtual reality project when he was asked to be video taped for a short promotional message.** He is thus *in a bit in a hurry*. He's *being nice* to the audience, but he really wants to get back to his projet.

Methodology – 3
Psychological and Situational Specification

- Giulio's **steady personality traits** are thus:
  - Pleasant, steady, confident, yet a bit tense (a go-getter, a Type I character)
- The **situational constraints** for the speech performance are thus:
  - "Look straight into the camera during the whole message"
  - "Don't move your head too much"
  - "Smile, but don't overdo it"
  - "Show determination"

Methodology – 4
Speech generation and enhancement

- **LAIPTTS-F** (1998) and **Mbrola** (2005) were used to *generate from French written text*:
  - A 2-second sentence:

    “Bonjour, et bienvenue à l’IMM”
  - A 15-second paragraph:

    "Bonjour! Je m'appelle Giulio Lorenzo, et j'étudie l'informatique dans la section IMM de l'Université de Lausanne. Je suis en plein milieu de travaux en statistique et en réalité virtuelle. Cela me demande beaucoup d'efforts, mais cela vaut *I1 bien *I0 la peine. Venez à *I1 Lausanne *I0 et voyez par *I2 vous *S2 mêmes *S0 *I0 !"

    [*I1 = minor increase in intensity]
    [*I2 = major increase in intensity]
    [*S2 = major slowdown]
- The **Sony “Enhance” filter** available in **Sound Forge 7** was used to *increase the sound quality* of the Mbrola output

Methodology – 5
Facial gesture construction

- **Mimic 3.1** was used to *generate facial gestures*.
- A small **Java program** was written to *convert the Mbrola Sampa phonetics* into Mimic-style US English approximations to French sounds.
- *Gestures were automatically aligned* by Mimic, but *manual verification of alignment* was necessary.

Methodology – 6
Animation and Rendering

- **Poser 6** was used to **animate** the Michael figure. This step needed no external intervention.
- The **Poser 6 firefly engine** was used to render the figure at 25 fps in AVI-PAL (720x576 pixels) uncompressed. *Render time:* about 2 seconds (50 frames) per hour on an AMD 2600+.
- **Windows Movie Maker** was used to convert the AVI video into **wmv** (Windows Media Video) at 512 Mbps and 1500 Mbps.

Results – the Good

- The whole project took approximately **one month, part-time**.
- Generally, **user response is positive**. The quality attainable seems reasonable for the effort.
- However, there are several areas that **need much improvement** ➔ *next slide*

Results: the Bad

- *The mechanics:*
  - too much manipulation, particularly with Mimic
- *The quality attainable:*
  - known limitations of LAIPTTS prosody and the voice quality of Mbrola concatenative output
  - Mimic synchronization is still relatively poor
- *The difficulties:*
  - Unclear linking of parameters in the Michael figure (e.g. “open lips”-“m-gesture”) prevents credible work when combining features like “happy” (open lips) and an /m/-gesture (closed lips)
  - The Poser gesture track editor is very poor and renders manual editing very difficult. There is insufficient support for working on sound and gesture at the same time.
- *The limits:*
  - Taken at surface value, programs like Mimic and Poser are insufficiently transparent in their definition of the facial and gestural parameter space. Their underlying physics and mechanics are barely documented.

Conclusion: Priorities

1. **LAIPTTS:** Improve prosody
2. Replace **Mbrola** with an engine that permits voice quality manipulation
3. Create a **new linking program** between TTS and Poser.