Speech Breathing Synthesis for Empathic Virtual Agents

Nicolò Loddo

Main concepts to decide on.

01

Voice Synthesis

How will the synthesis work? What models will be employed?
Will it be dynamic or work with static key-points?
Will it be pre-rendered or in real time?

02

Agent Features

Will the agent be an embodied agent or only a voice? If it is embodied, will it be anthropomorphic, theriomorphic, object?
If it is not embodied, will it feature some form of visual cues?

03

Study Design

After concretizing the voice and agent designs, how will we assess the research question? What will be the exact research question?
How will we involve users in the evaluation?

03 - let's start from the end

Study Design:

Research Question definition

Research Question Definition

The interest behind this study:
The communicative power of <u>the unsaid</u>.

- Enhancing **empathy** towards Virtual Agents
- The role of **breathing and pauses** in emotive communication
- Speech-breathing synthesization for Virtual Agents with **AI generative models**
- The possibility of combining the speech with an **emotional animation** of the character

My proposed Research Question

“Does Spontaneous Speech Synthesis with breathing noises improve empathy towards Virtual Agents, in respect to non-spontaneous speech methods?”

Post meeting note: more focus on the effect of breathing noises, with both conditions using spontaneous speech

Spontaneous and Non-Spontaneous

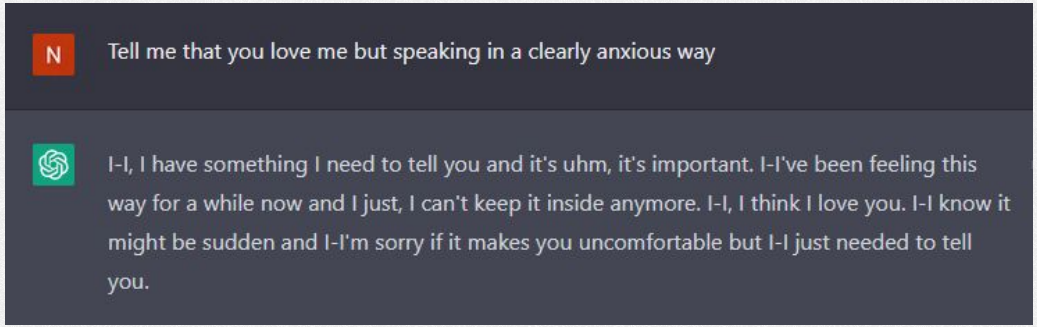
Different sets of Phonemes

Non-Spontaneous

The script is clear as written language. The prosody is dictated only by the style of the reading voice, but in general is comparable to a well pre-studied discussion.

Spontaneous

Breathing noises, “uh”s, “ehm”s, and other possible means of pause from high cognitive load are included in the phonemes during the speech model training.



Any considerations?

02 - Agent Embodiment Features

What agent type would make the study more objectively and straightforwardly give an answer to the RQ?

Humanoid	This can mean a human, robot or animal character with human features
Non-Humanoid	Non-anthropomorphic character: this can vary widely
Object Embodiment **(with few emotional cues)**	Something like Amazon Alexa cylinders, or a purposefully designed object with ad-hoc emotional cues
Invisible Human	Fake talking to a human without showing the human: for example, show the media of communication instead, or show the lamp in which a genie is trapped. Visual emotional cues can still be provided by the communication media.
Abstract Visualization	For example spectrograms or geometric lines
Only voice perhaps?	

02 - Agent Features

What agent type would make the study more objectively and straightforwardly give an answer to the RQ?

My considerations on this:

02 - Agent Features

What agent type would make the study more objectively and straightforwardly give an answer to the RQ?

Humanoid	MOST DIFFICULT
Non-Humanoid	DIFFICULT, LESS UNCANNY VALLEY PRONE
Object Embodiment (with few emotional cues)	NOT COMPATIBLE WITH THE VOICE REALISM, MAY LEAD TO UNCANNY VALLEY (Weis, 2017), AND NOT MUCH LIKED IN GENERAL (Wang et al., 2019)
Invisible Human	NOT EMBODIED, BUT CONGRUENT REALISM AND AVOIDS INTERACTION EFFECTS: E.G. breathing animation sync problem
Abstract Visualization	NOT EMBODIED, BUT CERTAINLY AVOIDS INTERACTION EFFECTS
Only voice perhaps?	NOT EMBODIED, BUT CERTAINLY AVOIDS INTERACTION EFFECTS

Continuum

Humanoid -—— Non Human

Human

- Prone to uncanny valley;

- Congruent with voice realism, but it probably depends on the stylization of the agent

Robot

- When kept at low human appearance is the least prone to uncanny valley

- Less congruent with realism of the voice and emotional arousal

Animal/Alien

- Attenuation of uncanny valley, but not always (Iannizzotto, 2018)

- Something similar to animated characters may work!

About the stylization: we can often see spontaneous speech in 3D animation movies without uncanny valley effects

Embodied or not embodied?
Human or not human?

Post meeting note: feasibility pushes towards not embodied, better assessment is needed and will be done along the way.

01 - Voice Synthesis

- **Fit cognitive model** to understand where spontaneous speech phonemes should be placed to produce spontaneous text: this has been done by Székely et al. (2020). Differently than in their paper, we will need to do it **depending on arousal and valence**
- **Fit Flowtron** with added phonemes as in Székely et al. (2019)
- Given normal text, arousal and valence, **infer an Emotional Spontaneous Speech Recording** with the fitted Flowtron.

Backup plan idea

Do not use Flowtron: after generation, use Speech Synthesis Markup Language and then manually add the breathing noises to obtain few recordings to ultimately test empathy.

P.S. - I still did not see the database: this procedure might change.

00 - thank you!

Thank you for the attention and time!

Do you have any final suggestions?
Do you have any questions?