Evaluating Synthesized Speech: The Cognitive Approach

Gediminas Navickas, Gerda Ana Melnik-Leroy

Institute of Data Science and Digital Technologies
Vilnius University

Background

Modern speech technologies allow the synthesizing of almost perfect synthetic speech. However, despite the increase in computational power, the size of the datasets and the elaboration of the algorithms, this “almost” has still never been surpassed. We propose that one of the major reasons for the impossibility to come up with a 100% naturally sounding synthetic speech is the **lack of appropriate quality evaluation methods**.

Why traditional measures do not work?

- **Subjective measures** (Mean opinion score; intelligibility and comprehension tests) based on **introspection** – not precise. According to research, **humans are not capable of describing the perceptual mechanisms in their mind**; they cannot know what parts of the signal are really important for the processing. Plus, the metacognitive awareness is related to **overconfidence**: participants overestimate their abilities.
- **Objective measures** - suitable to evaluate the physical speech signal and its properties, **but do not reflect how this signal is perceived by the human listener. Even tiny distortions** of the speech signal can have detrimental effects on **speech processing** quality and speed.

What cognitive science can propose

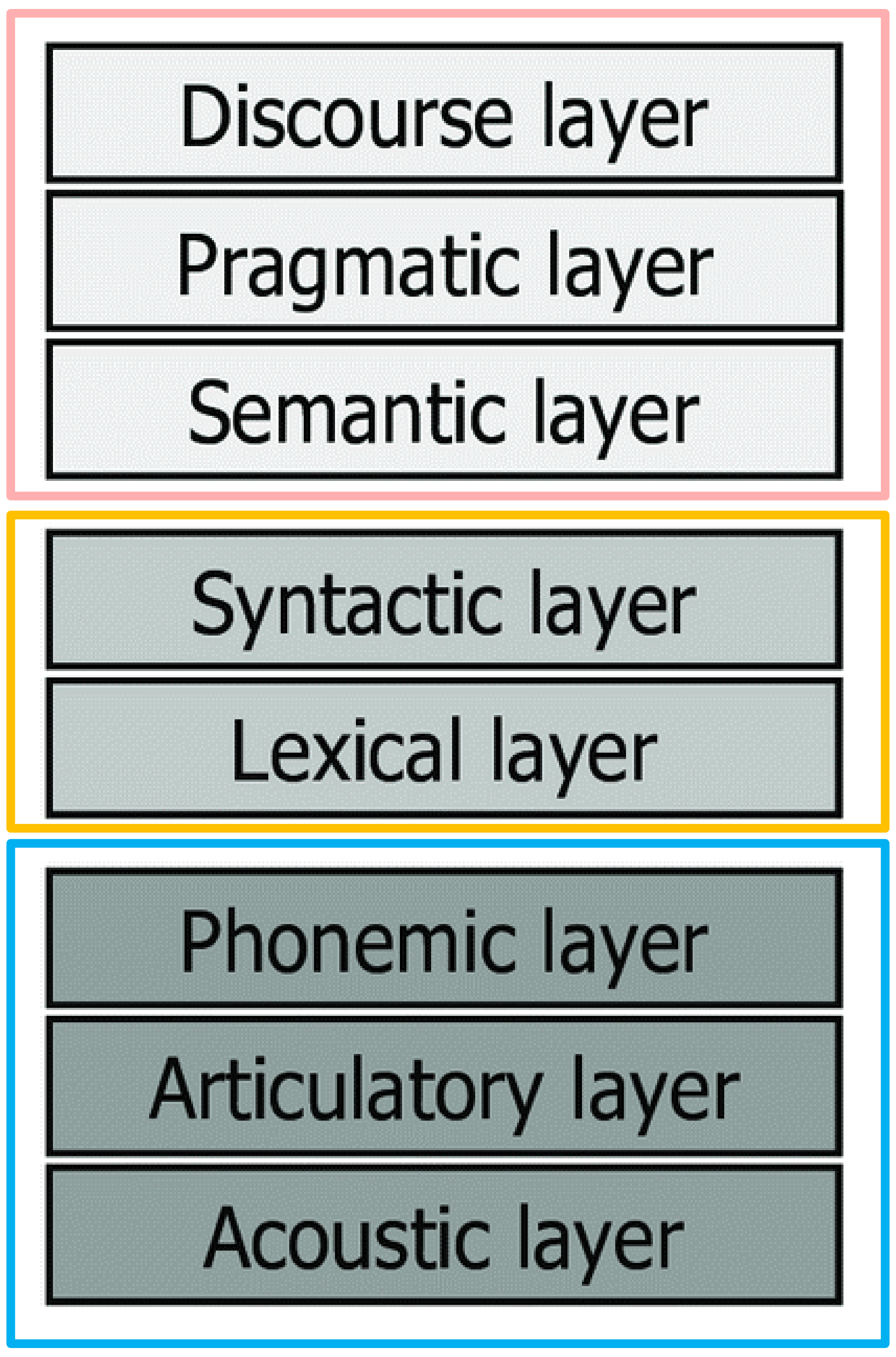
- Speech processing in humans – based on **several stages (processing levels)**.
- **Distortions in each stage affect the comprehension, its speed and effort needed.**
- Using methods from **cognitive psychology** to evaluate the perception/processing at each of these stages + bypass introspection!

Low level processing (acoustic-phonological processing)
e.g.: AX speeded discrimination task
Same or different? Measures: accuracy (might be Reaction times)
Purpose: Identify what listeners can discriminate,
what is their perceptual threshold?
=> Experiment on blind vs. sighted participants; different qualities.

Lexical processing
e.g.: speeded lexical decision task
Words+nonwords; task: press when you hear a real word. Measures: accuracy and Reaction Times.
Purpose: It allows to measure how quickly a lexical item is accessed in the mental lexicon. Slow speed = more effort and errors.

Highest levels of processing
e.g.: Intonation, prosody: Semantic priming;
Semantically or syntactically-ambiguous sentences

Speech Processing Levels



Conclusions and Further Investigations

Since speech perception has been widely studied in cognitive science, we propose that novel approaches to measure synthesized speech quality should be based on this research. Unlike traditional methods which describe only the quality of the signal or the introspective opinions of the listener, **behavioral methods from cognitive science** allow the **evaluation of the perception of the signal**.

Our first results from of a study using **AX speeded discrimination paradigm** show that this task can be used for the **evaluation of synthetic speech quality** and also for **defining the level of perceptual accuracy for particular user groups**.

In further investigations we will therefore explore the prospects of applying **a range of behavioral experimental methods**, that have been already developed and shown to be effective in cognitive science.
Such new methods for synthetic speech evaluation will **take into account the cognitive mechanisms underlying the processing of linguistic phenomena and will be sensitive to the peculiarities of human speech perception**.

Data Analysis Methods for Software Systems, 13th Conference, Druskininkai, Lithuania, December 1–3, 2022

Contact: Gediminas Navickas, gediminas.navickas@mif.vu.lt