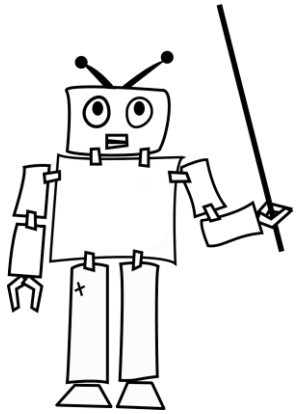
Information structure in speech synthesis



Improving the naturalness of German text-to-speech synthesis by means of a more context-appropriate prosody

Judith Baumann | Rijksuniversiteit Groningen | 71. StuTS + 31. TaCoS Leeuwarden 2022 | j.baumann.1@student.rug.nl

Content

- What is information structure?
- Information structure for speech synthesis
- Theme and focus
- State of the art and hypotheses
- Methodology
  - computational part
  - empirical part
- Expected results and conclusions

What is Information Structure?

- Forming a message in a way that makes it most easily understandable for the addressee in the current context

What does Anna love?

- Anna loves trains.

- #Trains Anna loves. (But: *Anna hates cars but trains Anna loves*)

- Many ways to mark information structural status (syntax, morphology, prosody)

Information structure for speech synthesis

- Prosody is affected by the linguistic (and non-linguistic) context:

Who called you yesterday? *LISA called me yesterday.*
When did Lisa call you? *Lisa called me YESTERDAY.*

In 500 Metern bitte rechts ABBIEGEN. Dann LINKS abbiegen.
In 500 Metern bitte rechts ABBIEGEN. #Dann links ABBIEGEN.
'In 500 meters please turn right. Then turn left.'

- The linguistic context should be taken into account when creating synthetic voices (otherwise the prosody will be off in some contexts)
- Relevant for the synthesis of longer monologues (e.g., audiobooks) or multi-turn conversations

Focus and theme

- Focus: New information/indication of the presence of alternatives

  What does you sister study?

  My sister studies [biology]$_{Focus}$**.**

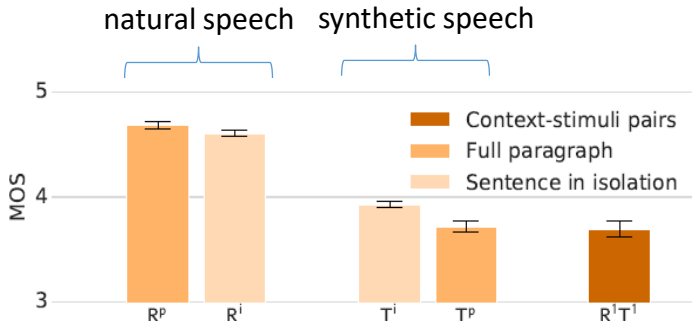
- Theme: What the utterance is about

  What does you sister study?

  [My sister]$_{Theme}$ *studies biology.*

State of the art and hypotheses

- Lower naturalness ratings for paragraphs
  than for the same sentences in isolation (English)
  (Clark et al. 2019)



→ Hypothesis: Considering information structural status in prosody generation can close the gap

- Synthesized isolated sentences from newspaper articles with a prosody modified to match the thematic structure are perceived as more “expressive“ (German) (Domínguez et al. 2018)

→Hypothesis: A German synthetic voice that has a prosody that is sensitive to the focal status of sentence parts is perceived as more natural than a synthetic voice that has a default prosody that does not take the focal status into account

Methodology – computational part

- Training an end-to-end algorithm for speech synthesis: input = text sections and corresponding audio sections → algorithm learns the correlation
- Latif et al. (2021): Training an End-to-End TTS system (*Fastpitch*) with
  - …recordings of read-aloud answers to questions (26.7 h)

Q. What happened ?	A. SARAH CLOSED THE HOUSE.
Q. Did Sarah closed the house ?	A. Sarah closed the house.
Q. Ava closed the house?	A. SARAH closed the house.
Q. Sarah occupies the house?	A. Sarah CLOSED the house.
Q. Sarah closed the parking?	A. Sarah closed the HOUSE.

  - …annotations of these recordings including focus tags on word-level

neutral	sarah closed the house
question	<Q> sarah closed the house
focus subject	<F> sarah closed the house
focus verb	sarah <F> closed the house
focus object	sarah closed the <F> house

Methodology – computational part

- Latif et al.'s result: Variations of f0, intensity, and duration corresponding to focus are similar to those of natural speech

→ „prosodic patterns of contrastive focus can be learned accurately“ (p.544)

- Train the same system with German data that is annotated for focus (challenge: data) and with data that is not annotated for focus

- Synthesize speech with both systems

Methodology – empirical part

- Stimuli:
  - isolated sentences
  - paragraphs (sequences of 3 sentences)

  → 50:50 synthesized by the two systems

- MOS-testing of naturalness (five-point scale from 1 to 5)

Expected results and conclusions

- I expect

  a) the difference in MOS between isolated sentences and paragraphs to be smaller when focus tags are used

  b) both isolated sentences and paragraphs to get a higher MOS when focus tags are used

- If confirmed: support the idea that information-structural categories should be included in the creation of synthetic voices in order to make them sound more natural

References

Clark, R., Silen, H., Kenter, T., & Leith, R. (2019). Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs. arXiv preprint arXiv:1909.03965.

Domínguez Bajo, M., Burga Díaz, A., Farrús, M., & Wanner, L. (2018). Towards expressive prosody generation in TTS for reading aloud applications. IberSpeech 2018; 2018 Nov 21-23; Barcelona, Spain. Baixas, France: ISCA; 2018. p. 40-4.

Drubig, Hans B. & Wolfram Schaffar. 2001. Focus constructions. In Martin Haspelmath (ed.), Language Typology and Language Universals/Sprachtypologie und sprachliche Universalien/La typologie des langues et les universaux linguistiques: 2. Halbband, 1079–1104. Berlin/Boston: De Gruyter Mouton.

Féry, Caroline & Manfred Krifka. 2008. Information structure: Notional distinctions, ways of expression. In Piet van Sterkenburg (ed.), Unity and Diversity of Languages, 123–135. Amsterdam: John Benjamins Publishing

Latif, S., Kim, I., Calapodescu, I., & Besacier, L. (2021, November). Controlling Prosody in End-to-End TTS: A Case Study on Contrastive Focus Generation. In Proceedings of the 25th Conference on Computational Natural Language Learning (pp. 544-551).

Thank you