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

Judith Baumann

Abstract

The term 'information structure' refers to a way of forming a message in a way that makes it most easily understandable for the hearer against the background of a certain context (Féry & Krifka 2008). Two examples for information-structural statuses are theme and focus. The theme of an utterance can be defined as what the utterance is about while the focus is often defined as the new information of the sentence (e.g., by Halliday 1967 or Vallduví & Engdahl 1996). Recent studies have shown for various languages that the naturalness of synthetic voices improves when their prosodic structure matches the thematicity of the sentence parts (see, e.g., Meurers et al. 2011, and Domínguez et al. 2018 for German, or Vanrell et al. 2013 for Catalan). Against the background of the well-established finding that information structure is reflected in prosody in many languages (Halliday 1967; Chafe 1976; Vallduví 1993), this is not surprising and there is reason to assume that the same holds for other information-structural categories.

To test this assumption, I will investigate whether the naturalness of a German synthetic voice can profit from an intonation that matches the focus structure of the sentence. Based on Domínguez et al. (2018) who have shown that the mean opinion score (MOS) for a synthetic voice improves when the prosody is automatically or manually modified in accordance with thematicity, I hypothesize that the MOS for a German synthetic voice will also improve when the prosody is modified to match the focus structure.

To that aim, I am planning to synthesize two German voices - one with and one without a focus-matching prosody - and compare their MOS's. To create these voices, I will either train an existing statistical-parametric-speech-synthesis (SPSS) based TTS system once with focus/non-focus labels and once without focus/non-focus labels or I will adopt the strategy of Domínguez et al. (2018) who used an open-source TTS software and enriched the default synthesized speech with prosodic information according to the thematicity of the corresponding sentence parts.

A result that confirms my hypothesis would suggest that information-structural categories should be included in the creation of synthetic voices in order to make them sound more natural.

Keywords *– speech synthesis, prosody, information structure, naturalness*

References

Chafe, L. W. 1976. Givenness, contrastiveness, definiteness, subjects, topics and point of view. In Charles N. Li (ed.), *Subject and Topic*, 25–56. New York: Academic Press.

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.

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.

Halliday, Michael A. K. 1967. Notes on transitivity and theme in English: Part 2. *Journal of Linguistics* 3. 177–274.

Meurers, Detmar, Ramon Ziai, Niels Ott & Janina Kopp. 2011. Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In Proceedings of the TextInfer 2011 workshop on textual entailment TIWTE '11. Association for Computational Linguistics, Stroudsburg, PA, USA.

Vallduví, Enric. 1993. *The informational component*. Doctoral Dissertation. Philadelphia: University of Pennsylvania.

Vallduví, Enric & Elisabet Engdahl. 1996. The linguistic realization of information packaging. *Linguistics* 34(3). 459–519.

Vanrell, Maria, Ignasi Mascaró, Francesc Torres-Tamarit & Pilar Prieto. 2013. Intonation as an encoder of speaker certainty: Information and confirmation yes-no questions in Catalan. Language and Speech 56. 163–190.