Evaluation of Information Structure in Speech Synthesis: The Case of Product Recommender Systems

Frank Kügler, Bernadett Smolibocki, Manfred Stede

Dept. of Linguistics/EB Cognitive Science & SFB 632 “Information structure”,
University of Potsdam, Karl-Liebknecht-Straße 24-25, 14476 Potsdam
Email: {kuegler,smoliboc,stede}@uni-potsdam.de
Web: www.sfb632.uni-potsdam.de/projects_t2eng.html

Abstract

Speech synthesis nowadays is of acceptable quality for many purposes. Nonetheless there are applications where contextual and other pragmatic factors play an important role, which cannot be accounted for by straightforward text-to-speech (TTS) systems. This is the case for systems giving product comparisons and recommendations: For instance, an appropriate intonation is required that signals contrasting entities, and in longer discourse there is a need to distinguish between given and new entities prosodically. That is, the linguistic notion of *information structure* (IS) should be considered in the synthesis. In our project, we are extending an existing text generator for product comparison/recommendation with a speech synthesis component, and we are aiming at integrating information structure in a systematic way.

Our paper describes the architecture of our system (as currently being built) and the results of two perception experiments that we have conducted in order to verify that listeners do indeed perceive the difference between "standard" TTS and IS-enriched synthesis. The results show that there is a benefit of the IS-enriched synthesis for the listeners.

1 Introduction

Any TTS system uses a minimal syntactic text analysis (part-of-speech tagging). For many purposes, this is sufficient, but more complex discourse tasks require some information structure (IS) marking to highlight particular constituents of a sentence. A TTS system cannot determine different information structure categories by surface-based text analysis methods. Some more elaborate systems for handling the different information status of “given” and “new”, have been proposed several years ago, e.g. the SYNPHONIC project for German [1].

Our goal is to generate an IS-tailored intonation with the TTS system MARY [2] so that the information statuses of the constituents of a sentence are signaled. To this end we use a separate text generating system, a product recommender system for mobile phones, which provides IS-attributes/categories. Hence, the output of this system is a text with additional IS-tags for the individual constituents, specified in MARYXML (the interface description language of MARY).

The translation of the IS-attributes into appropriate intonation patterns is conducted by the synthesizer which we have upgraded with additional phonetic realizations of pitch accent assignment rules and phonetic details of pitch accent realization. The basic idea is similar to the multi-level flight information system FLIGHTS [3].

1.1 Information structure

The appropriate information packaging of an utterance depends on features of the discourse. In particular, constituents refer to different IS categories, namely *new/given, topic/comment, focus/background* [4, 5]. In order to signalize these IS categories appropriately to the listener, certain prosodic (e.g. different prominence relations) or syntactic means (or a combination of both) are used in intonation languages such German [6-9]. For this paper, we concentrate on the IS dimension *corrective focus*, which is defined as a correction of a previously mentioned fact that from the semantic point of view restricts possible alternatives. We assume that corrective focus is a subtype of contrastive focus [10, 11]. Relevant phonetic cues of corrective focus concern the horizontal and vertical alignment of the pitch peak. For example, corrective focus is realized by an enhanced pitch register [7, 12].

1.2 Architecture of the generation system

Product recommendation, as it has become commonplace in recent years, means automatically selecting "possibly interesting" products to be offered to a user on the basis of some representation of his or her consumer preferences. One well-known example is the recommendation function that amazon.com offers to their registered customers by merely listing books or other items that "could be of interest". In contrast, in the context of NL dialog systems, the recommendation scenario is much more elaborate: The user explicitly states his or her needs, and then the system engages in a dialog that suggests products, compares and possibly actively recommends them.

The setting that we have developed in *Polibox* [13] is in-between these two extremes: There is no language *understanding* component, because the user enters the desired product features via a menu. The system determines suitable products, and then the user selects those products that s/he wants to have described and compared. The system output, however, is in natural language (as well as in a markup language which serves as input for the speech synthesis component: MARYXML).

The current product domain is mobile phones. When the user has given a target description, Polibox finds the best matches in its database and then generates short texts, which describe a phone, and compares them when appropriate. Polibox also gives recommendations as to the relative suitability of a phone with respect to the query. Our speech synthesis front-end produces acoustic output, and our central research goal here is to account for the information structure phenomena (IS) in a principled way. The IS categories we handle are *information status* (given versus new), *topic/comment*, and *focus/background*. When the generator plans the text, it keeps track of the entities being talked about in its discourse memory, and it thus "knows" the values of these parameters at any point in the generation process. The precise mapping from discourse features to IS and prosody is our current work in progress.

2 Perception

To evaluate the sensitivity of listeners towards the presence or absence of IS in speech synthesis, we conducted two perception experiments with listeners judging the quality of two types of synthesis: (i) the original default synthesis by MARY and (ii) IS-enriched synthesis with prosodic parameters adjusted according to our expectations of how the information structure categories should be realized. The experiments differ according to how the IS-enriched synthesis was built. The first perception experiment used manually manipulated stimuli (section 2.1.2) while the second used stimuli that were synthesized by the prosody module of MARY with respect to the pitch accent realization for contrastive constituents (section 2.2.2).

Both experiments rely on the method of a forced-choice semantic congruency task [14] where listeners are confronted with a context-target sentence dialog. In this perception paradigm the task is to evaluate whether the target sentence matches the context or not. The context in this experiment corresponds to the user-model input of the product recommender system. The following example illustrates a user request with a corresponding text output of the text generation system; two hits of mobile phones are matched.

User request: looking for a mobile phone, the user selects the feature *'diagonale Displaygröße von mindestens 2 Zoll'* (diagonal measurement of the screen at least 2 inch)

Output of the text generation system:
Wie gewünscht hat das Samsung i8510 eine diagonale Displaygröße von 2,8 Zoll. Im Gegensatz zu dem Samsung hat das iPhone 3GS eine diagonale Displaygröße von $(3,5)_{cf}$ *Zoll.*
(As requested, the Samsung i8510 display measures 2,8 inches diagonally. In comparison, the iPhone 3GS display measures $(3,5)_{cf}$ inches diagonally.)

Both mobile phones match the user request. But the second one has the bigger screen. Thus, $(3,5)_{cf}$ corrects the previous constituent (2,8). For the perception tasks, we use less complex target sentences. But basically, the dialogs used for the experiments correspond to the structure of this example dialog.

2.1 First experiment

2.1.1 Material

In a forced-choice semantic congruency task, pairs of question and answers were presented. The questions were produced by a human speaker eliciting wide focus, or contrast on either of the arguments or the verb of the target sentence. Hence in contexts of contrast the corrected constituent represents new information while the other constituents represent given information. The syntactic structure of the target sentences was S Aux O V. Corrective focus was systematically varied as a function of position in the sentence. Hence, subject focus refers to sentence-initial focus, object focus to sentence-medial, and verb focus to sentence-final focus. In order to vary the semantic content of the target sentences we chose the objects from the domain of mobile phones (and a few others). Tab. 1 shows the complete set of questions with the matching target sentences of the domain 'mobile phones'.

Context condition	Questions-Answer
wide focus	Erzähl mal, was ist passiert? (What happened?) *Martin hat sich ein iPhone gekauft.* (Martin has bought an iPhone.)
initial contrast	Hat sich Mona ein iPhone gekauft? (Did Mona buy an iPhone?) *[Martin]$_{cf}$ hat sich ein iPhone gekauft.*
medial contrast	Hat sich Martin ein Samsung Wave gekauft? (Did Martin buy a Samsung Wave?) *Martin hat sich ein [iPhone]$_{cf}$ gekauft.*
medial contrast	Hat sich Martin ein iPhone geliehen? (Did Martin borrow an iPhone)? *Martin hat sich ein iPhone [gekauft]$_{cf}$.*

Table 1: Example of speech material for the perception tasks. Corrective focus is marked with $[--]_{cf}$.

In addition, cross-spliced versions (mismatching pairs) of the four sets were created: All of the four context questions were combined with the target sentences from the other three remaining contexts. The task was arranged with the MFC software of PRAAT [15], and the short dialogs were presented auditorily via headphones. Listeners had to estimate whether a target sentence (answer) matches the previously presented context question or not. Therefore, two possibilities for answering were given on the screen; (i) *match* or (ii) *no-match*. The participants had to click either on the *match*-button or on *no-match* button after every stimulus pair. The test started after a short introduction by the experimenter and three training dialogs; one cycle lasted about 15 minutes. Overall, 60 stimuli pairs were presented (4 original stimuli, 4 IS-enriched (see section 2.1.2), 12 cross-spliced and three repetitions of each), and 22 participants attended. No hearing deficits of the participants were reported.

2.1.2 Development of the IS-enriched stimuli

MARY provides no built-in handling of information structure. For the original MARY synthesis, the identical target sentence was thus used in combination with the different context questions. Fig. 1 presents an example of a wide-focus realization with a characteristic downstep pattern [7]. The pitch contour starts at 106 Hz with the rising pitch accent L+H* for the subject *Martin* and a pitch peak about 158 Hz. The object *iphone* receives an H* pitch accent with a pitch peak about 139 Hz. Tone labels are based on GToBi [16].

On the basis of the literature on the phonetic realization of focus in German we assume that the original MARY synthesis needs to be adjusted prosodically according to the context in which the sentence appears. Hence, to produce the IS-enriched stimuli, the original synthesis by MARY was manually manipulated using the overlap-add resynthesis method in PRAAT. Fig. 2 represents an example with a manipulated correctively-focused constituent. The existing pitch accent was adjusted according to the phonetic findings on vertical scaling and horizontal alignment in German [7, 12]. In consideration of these results, the pitch peak of the object contrast was increased by around 20 percent up to approximately 156 Hz and realized at the end of the stressed syllable of the correctively focused constituent [17]. The accentual fall began immediately after the stressed syllable and extended to the following word at about 90 Hz. The other constituents, representing given information, were also adjusted to mirror deaccentuation [18].

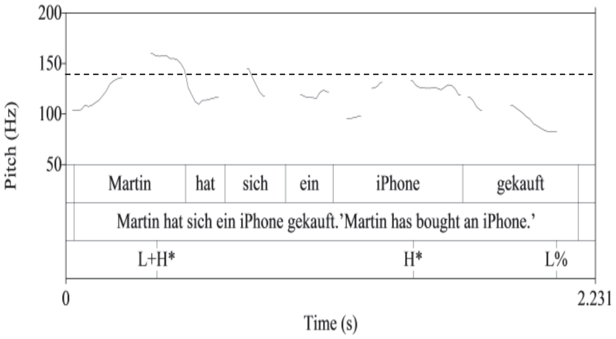

Figure 1: F0 contour of the original synthesis by MARY in a wide-focus context; GToBI labels represent the pitch accents. The dotted line represents the pitch level for the pitch peak of the nuclear accent H* about 130 Hz.

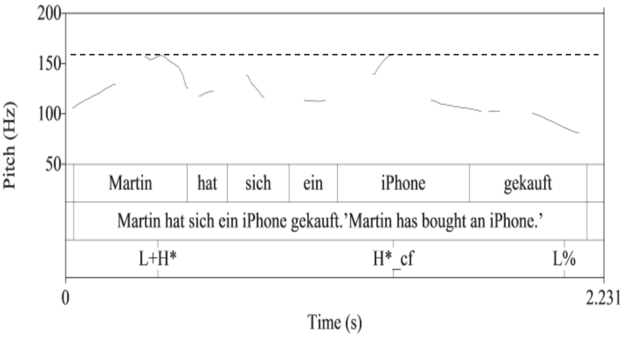

Figure 2: F0 contour of the IS-enriched synthesis of the same sentence as in Fig. 1 produced in an object contrast context. GToBI labels represent the pitch accents. The dotted line represents the pitch level for the pitch peak of the nuclear accent H*_cf about 156 Hz.

2.1.3 Results

Fig. 3 displays the congruency ratings for the matching contexts as a function of focus position. The results show that manipulated stimuli were rated as more congruent than the original unaltered realizations by MARY across all information structure conditions except for wide focus. In this case, listeners rated the original and manipulated versions equally good. Presumably, listeners rated the relative prominence relations in the wide focus sentences as good enough to function as an all-new sentence. This means that the default synthesis of MARY corresponds with the wide focus condition and a more fine-grained phonetically adjusted synthesis does not add any further benefit to this case.

Fitting a linear mixed model [19] with manipulation as fixed factor, and speaker and item as random factors confirms the hypothesis that manipulated stimuli are rated as significantly more congruent than the original MARY synthesis (SE 0.2120, z = -8.035, p < 0.000).

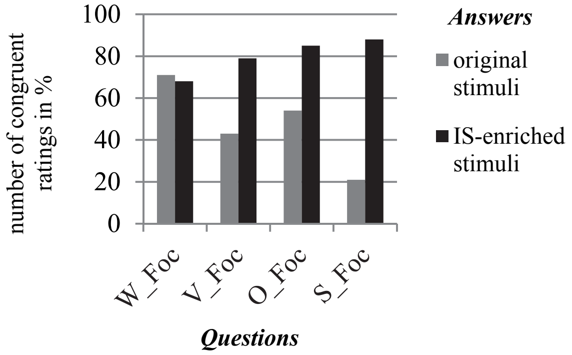

Figure 3: Comparison of original and IS-enriched stimuli in matching contexts. W_Foc= wide focus, V_foc=verb focus, O_Foc=object focus, S_Foc=subject focus, n=66

Fig. 4 shows the results for the mismatching contexts. All of the stimuli were rated equally bad except for the wide focus answers in the contrast focus condition of the object. In wide focus both the subject and the object of the sentence carry a pitch accent, hence indicating some prominence. The high congruency rating of wide focus answers in case of object focus results from the perceptual impression of prominence due to a pitch accent on the object. In all other cases shown in Fig. 4 where there is no pitch accent on the subject or verb, hence no prominence, the listeners identified the absence of prominence as a mismatch.

Fitting a linear mixed model with cross-splicing as fixed factor, and speaker and item as random factors confirms the hypothesis that any mismatch stimulus is rated as incongruent to the corresponding context question (SE 0.1270, z = -12.991, p < 0.000).

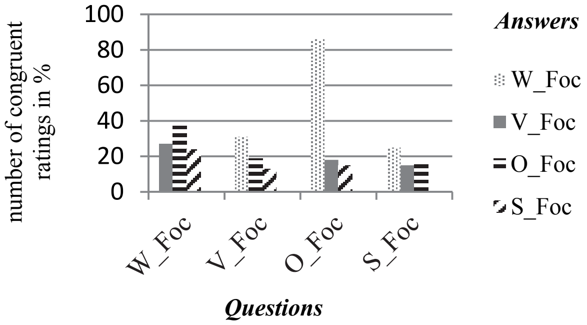

Figure 4: Comparison of IS-enriched stimuli in mismatching contexts. W_Foc=wide focus, V_foc=verb focus, O_Foc=object focus, S_Foc=subject focus, n=66

2.2 Second perception experiment

In comparison, the results of the IS-enriched synthesis of the first perception task indicated that there is a benefit for the listeners with an appropriate prosodic marking of contrastive focus. Therefore, for the second perception run we improved the MARY code with respect to its pitch accent realization in order to generate an appropriate intonation for a sentence with a constituent marked with contrast.

2.2.1 Material

16 participants took part in this perception task by using the same experiment set-up as the first one (see section 2.1). Furthermore, the same stimulus material was reused for the subject, object and verb contrast, but the prosodic IS-adjustment of the target sentences according to the context was different. The wide-focus condition and cross-spliced sets were not investigated due to the results of the first perception task. In total, 18 question-answer pairs were presented (3 original stimuli, 3 IS-enriched (see section 2.2.2) and three repetitions of each).

2.2.2 IS-enrichment in MARY

The idea of the IS-enrichment in MARY is based on an appropriate pitch accent assignment for the individual constituents depending on their particular information status. For this, the text generator provides IS-tags. Currently the constituents of a sentence receive an additional *contrast*-attribute which is considered for the following pitch accent assignments. In MARY 4.3.0, it is possible to control the prosody with an additional *prosody element* such as *rate, pitch* and *contour* or by using predefined GToBI accents [20, 21].

The prosodic calculations and adjustments are executed by the prosody module in MARY. Due to the fact that no specific accent configuration for a contrastive element exists, a specific accent configuration for contrast in relation with a H* pitch accent was added. This new accent was created as a string of tuples, where the first figure indicated the temporal level and the second the Hertz level according to the phonetic realizations of the intonation contour of contrastive accents. Any additional IS configuration such as deaccentuation of the particular elements transferring given information is not realized in MARY yet.

2.2.3 Results

Fig. 5 shows the results of this perception task. As in the previous experiment the IS-enriched stimuli were rated as more congruent than the original synthesis by MARY.

Fitting a linear mixed model with manipulation as fixed factor, and speaker and item as random factors confirms the hypothesis that manipulated stimuli are rated as significantly more congruent that the original MARY synthesis (SE 0.3882, z = -4.293, p < 0.000). Compared with the results of the first experiment the scores of the congruency rating were lower, especially in the case of the initial subject contrast. The particular role of initial subject focus may be due to the fact that in the remainder of the sentence some prominence of the object interferes with the impression of initial focus. Note that object prominence was not adjusted.

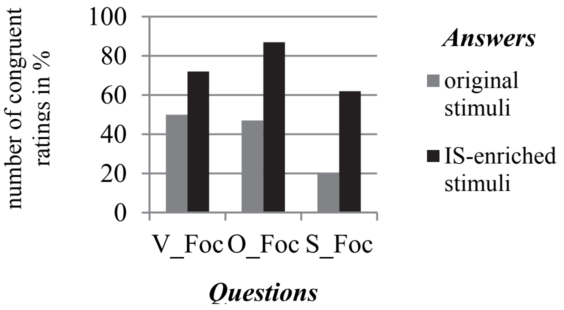

Figure 5: Comparison of original and IS-enriched stimuli in matching contexts. V_foc=verb focus, O_Foc=object focus, S_Foc=subject focus, n=48

3 Conclusions

Two perception experiments were conducted aiming to investigate the benefit of IS-enriched synthesis for the listeners. The results show higher congruency ratings for the IS-enriched stimuli across all contrast conditions in matching contexts. That means that listeners prefer the IS-synthesis over the original realization by MARY. In addition to the case of subject contrast, the results of the second perception experiment indicate that it is also necessary to consider the intonation of a whole sentence. The use of one IS category e.g. contrastive focus is accompanied by further IS-configurations; in the case of sentence initial contrastive focus other constituents are given and in postfocal position deaccented.

We take the preference for IS-enriched intonation synthesis as a support for the hypothesis that systems involving complex user-system interactions (as they take place in product recommendation) should provide for discourse-driven information structure attributes, which the speech synthesis component can take as input for producing IS-sensitive intonation.

4 Acknowledgements

This paper is part of the project T2 of the SFB 632 on Information Structure at Potsdam University, financed by the DFG. We are grateful to Sebastian Varges, Andreas Peldszus and Daniel Pusch who provided valuable technical assistance.

References

[1] B. Abb, C. Günther, M. Herweg, K. Lebeth, C. Maienborn and A. Schopp, “Incremental phonological encoding – an outline of the SYPHONICS formulator,” *Proceedings of the 4th European Workshop on Natural Language Generation*, (Pisa), 1993.

[2] M. Schröder and J. Trouvain, **“**The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching,” *International Journal of Speech Technology 6*, pp. 365–377, 2003.

[3] M. White, R. A. J. Clark and J. D. Moore, “Generating tailored, comparative descriptions with contextually appropriate intonation,” *Computational Linguistics 36 (2)*, pp. 159-201. 2010.

[4] W. L. Chafe, “Givenness, contrastiveness, definiteness, subjects, topics and point of view;” in Charles N. Li*, Subject and Topic*, (New York) pp. 27-55, 1976.

[5] M. Krifka, “Basic Notions of Information Structure,” in Féry, C., Fanselow, G., Krifka, M. (eds.)*, Working Papers of the SFB632, Interdisciplinary Studies on Information Structure (ISIS) 6*. (Potsdam, Germany), pp. 13–56, 2007.

[6] F. Kügler, “The role of duration as a phonetic correlate of focus,” in Barbosa, P. A., Madureira, S., Reis, C. (eds.)*, Proceedings of the Speech Prosody 2008 Conference. Campinas*, (Brazil), pp. 591–594, 2008.

[7] C. Féry and F. Kügler, “Pitch accent scaling on given, new and focused constituents in German,” *Journal of Phonetics 36*. pp. 680–703, 2008.

[8] S. Baumann and J. Becker, M. Grice and D. Mücke, “Tonal and Articulatory Marking of Focus in German,” in *Proceedings of the 16th International Congress of Phonetic Sciences*, (Saarbrücken). pp. 1029–1032, 2007.

[9] G. Fanselow and D. Lenertová, “Left Peripheral Focus. Mismatches between Syntax and Information Structure,” *Nat Lang Linguist Theory* 29, pp. 169–209, 2011.

[10] M. Götze, T. Weskott, C. Endriss, I. Fiedler, S Hinterwimmer, S. Petrova, A. Schwarz, S. Skopeteas and R. Stoel, „Information Structure,” in Dipper, S., Götze, M. & Skopeteas, S. (Eds.) *Working Papers of the SFB632, Interdisciplinary Studies on Information Structure (ISIS) 7*, *pp.* 147-187 (Potsdam, Germany), 2007.

[11] S. Repp, “Defining 'contrast' as an information-structural notion in grammar,” *Lingua 120 (6)*, pp. 1333-1345, 2010.

[12] F. Kügler, C. Féry and R. van de Vijver, “Pitch accent realization in German,” *Proceedings of the 15th ICPhS.* (Barcelona, Spain), pp. 1261-1264, 2003.

[13] M. Stede, “Polibox: Generating desciptions, comparisons, and recommendations from a database,” in *Proceedings of the 19th Int'l Conference on Computational Linguistics (Coling)*, (Taipei), 2002.

[14] P. Prieto, “Experimental methods and paradigms for prosodic analysis,” in Cohn, A., Fougeron, C., Huffman, M. (eds.)*, Handbook of Laboratory Phonology*, Submitted.

[15] P. Boersma and D. Weenink, “Praat-doing phonetics by computer,” `http://www.praat.org` (June 26, 2012)

[16] M. Grice, S. Baumann and R. Benzmüller, “German Intonation in Autosegmental-Metrical Phonology,” in Jun, Sun-Ah (ed.) *Prosodic Typology: The Phonology of Intonation and Phrasing,* pp. 55-83, 2005.

[17] E. Grabe, “Comparative Intonational Phonology. English and German,” Ph.D. thesis, (Nijmegen), 1998.

[18] S. Baumann, “The Intonation of Givenness. Evidence from German,” (Tübingen), 2006.

[19] D. Bates and D. Sarkar, “lme4: Linear mixed-effects models using s4 classes,” R package version 0.9975-11 [Computer software], 2007.

[20] S. Pammi, "Prosody control in HMM-based speech synthesis," 2011. `http://www.dfki.de/~chandra/marticles/pammi_Report_ProsodyControl.pdf`, accessed on 26. June 2012.

[21] SSML, W3C Recommendation. `http://www.w3.org/TR/speech-synthesis/`, accessed on 26. June 2012.