PROSODIC DISAMBIGUATION FROM DEEP SYNTACTIC STRUCTURES

Philipp von Böselager

I*f*L Phonetik, University of Cologne
philipp.boeselager@uni-koeln.de

Berthold Crysmann

DFKI GmbH & Universität des Saarlandes
crysmann@dfki.de

ABSTRACT

In this paper, we report on an experiment showing how the introduction of detailed prosodic information into synthetic speech leads to better disambiguation of structurally ambiguous sentences. Using modifier attachment (MA) ambiguities and subject/object fronting (OF) in German as test cases, we show that prosody which is automatically generated from deep syntactic information can lead to considerable disambiguation effects, and can even override a strong semantics-driven bias. The architecture used in the experiment, consisting of a large-scale generator for German, a prosody module, and the speech synthesis system MARY is shown to be a valuable platform for testing hypotheses in intonation studies.

Keywords: Prosody, Disambiguation, Synthesis, Generation, TTS, CTS

1. PROSODY FOR GENERATED SPEECH

The inclusion of prosodic information is standardly believed to play a prominent role for the improvement of CTS and TTS applications, in terms of naturalness and intelligibility, see, e.g., [4] and [5]. Another added value of prosody lies with its potential for disambiguation: it is often observed that structural ambiguities found in written texts are absent from speech, which is prosodically structured. In order to assess this potential, we carried out an experiment to establish how and to what extent prosody can contribute to improved comprehension of automatically generated speech. We conjecture that disambiguating prosody will not only lead to better intelligibility, but also enhance overall naturalness, due to an improved correspondence between intended meaning and prosodic realisation.

1.1. Background

1.1.1. Modifier attachment

Probably one of the most thoroughly studied types of structural ambiguity in human language are attachment ambiguities. While most research in this area has focused on written language, more recently, there has been a number of detailed studies of how prosody contributes to disambiguation, most notably the work of Schafer [6] and Speer et al. [8]. Using task-oriented elicited speech, [6] identified the prosodic parameters responsible for disambiguation of attachment ambiguities in English as follows: High attachments are perceived best, when there is a prosodic break before the modifier, but not between the preceding object NP and the verb. Conversely, low attachment corresponded to the absence of a prosodic break between the modifier and the NP to which it is attached; the entire object NP, including the modifier, was preceded by a prosodic break. [8] observe that, high attachment is characterised by an increased duration of the head noun and following pause, which was verified perceptually.

1.1.2. Object fronting

In German, both subjects and objects can appear in sentence-initial topic position, preceding the finite verb. Since nominative and accusative case are not always distinct, local or even global ambiguity can arise with regard to grammatical function. Subjects in topic position are generally considered unmarked. In an eye-tracking experiment using resynthesised speech, Weber et al. [9] showed that prosodic information leads to early effects with sentences involving local ambiguity. Using an L+H* contour on fronted objects followed by a steep fall [9] achieved early disambiguation, even against a strong bias for subject topics.

1.2. System architecture

The prosody component evaluated in the experiments is part of a system that implements an entire CTS pipeline from semantic input to speech output.

1.2.1. Deep syntactic generation

The tactical syntactic generator used in the experiments consists of a linguistically grounded large-scale HPSG grammar of German (GG; http://gg.dfki.de), running on the LKB system [1]. Both the grammar and the processing system are

fully reversible, i.e., they are suitable for parsing, as well as generation. The generator takes as input semantic representations, essentially encoding predicate-argument structures. Given the reversibility of both grammar and processing system, the current architecture may also be used in a TTS scenario by simply running the grammar in parsing mode. Selection of the most probable attachment for TTS can be performed using parse-selection models, which currently achieve over 80% exact match accuracy over a 25.44% baseline.

As output the generator produces surface strings, together with two isomorphic tree structures, one containing categorial information, the other encoding functional notions, such as head, subject, complement or modifier.

1.2.2. Syntax-Prosody Interface

The two tree representations provided by the generator are folded into a single XML tree representation, where functional and categorial labels are represented as attributes on the nodes. The information contained in the syntax trees is transformed into prosodic markup by means of XSLT, crucially using XPATH regular expressions. The prosodic markup generated on the basis of the syntactic representations comprises tonal and phrasing information, represented as GToBI annotations [2].

Realisation of the prosody module as a separate component was a design decision, since it supports a clean separation of syntactic and phonological aspects suitable for distributed development. Furthermore, it permits the use of alternative syntactic generators, provided the structures are rich enough to make the appropriate choices as to prosodic realisation. Moreover, it enables the use of alternative speech synthesizers provided that they support prosodical XML-input.

1.2.3. Phonetic realisation

The prosodically annotated text is submitted to the MARY synthesis system [7] for phonetic realisation using diphone synthesis. MARY is a highly flexible TTS system, supporting annotation of the input data, ranging from low-level control over physical parameters to high-level phonological specification. For the experiments, we made use of GToBI-style tones and break indices, while disabling MARY default prosody rules.

2. PERCEPTION EXPERIMENT

In order to quantify the potential for prosodic disambiguation, we carried out a perception experiment, comparing how subjects interpret prosodically disambiguated stimuli as compared to their ambiguous textual counterparts. Furthermore, we used different candidate contours for each of the intended interpretations in order to measure which combination of tones and breaks will perform best.

2.1. Method

The experiment was designed as an online study; subjects were not observed. To make sure that the task was clearly explained, the main study was preceded by a pilot, involving 5 subjects.

The main study was carried out with 58 subjects (27 female, 31 male). They were aged from 17 to 54, and came from all parts of Germany.

Subjects had to assign an interpretation to each stimulus in a self-paced forced-choice test. Each stimulus could be heard as often as required.

In order to control for semantic or pragmatic preferences, subjects first had to judge stimuli presented in text form. 4 different sentences were used for modifier attachment and 2 for object fronting. From these sentences we generated 4 different speech stimuli for modifier attachment and 3 for object fronting, yielding a total of 6 textual and 22 randomised speech stimuli per subject.

2.1.1. Stimuli for Modifier Attachment Experiment

The sentences involving modifier attachment (MA) ambiguities all followed the same basic syntactic pattern subject-verb-object-modifier (S-V-O-M). For the generation of disambiguating speech stimuli, we used a combination of prosodic breaks and tones [2]. In order to determine which prosody gives the best results in terms of naturalness and disambiguation, we tested 4 different tonal patterns, 2 each for high and low attachment.

High MA: S [[V O] M]

- H1: - L* on O head noun
  - H- before M
- H2: - L+H* on O head noun
  - L- before M

Neither had any break before the direct object (O). The other two possible tone combinations, i.e., H* H- and L* L- sounded unnatural and were therefore discarded during the pilot study already.

Low MA: S [V [O M]]

- L1: - no break before O
- L2: - H- in front of O

Both versions contained an L+H* on O and no break before M.

2.1.2. Stimuli for Object Fronting Experiment

The disambiguating speech stimuli for the object vs. subject fronting subtask were based on [9]. Since timing was not an issue in our study, contrary to [9], we inserted an additional intonation phrase break after the fronted object in OVS-sentences. Also in contrast to [9], ambiguity was global, not local. The resulting utterances synthesized by use of the prosody module have the following prosody:

SVO

- no intonation phrase break after fronted S

OVS

- OVS1: L- after fronted O
- OVS2: H- after fronted O

Both, SVO and OVS, ended in an L-% boundary tone and had an L+H* accent on the fronted constituent. In the OVS-versions this accent was additionally emphasized by raising the peak, thus strengthening accent-prominence. The tonal pattern used for SVO was based on the default contour in MARY.

2.2. Results

The main experimental results are summarised in figures 2 and 3. As compared to the baseline obtained with textual stimuli (bias), the perception experiment shows a clear disambiguation effect with speech stimuli, for both modifier attachment and subject vs. object fronting. Our main claim that prosody automatically generated from deep syntactic structures can be used for the task of disambiguation in CTS and TTS was confirmed.

2.2.1. Modifier attachment

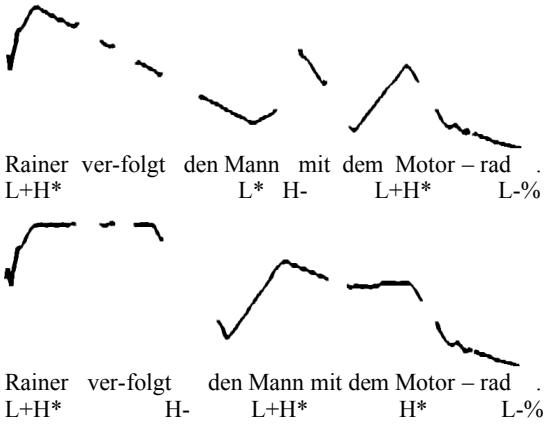

Figure 1: The prosodic contour of high (H1, top) vs. low MA (L2, bottom) with GToBI-Annotation.

Best disambiguation results were obtained with contours H1 and L2, given in Figure 1. The results for these contours are given in Figure 2, where a value of 1 corresponds to perceived high attachment, and 0 to low. Interpretations assigned are provided for each of the 4 test sentences, averaged over all 58 subjects. Test sentences differed as to their inherent semantic attachment preferences (bias calculated from textually presented stimuli): while (a) does not display any clear preference, (b) and (c) have a strong preference for low attachment, while (d) is mainly attached high.

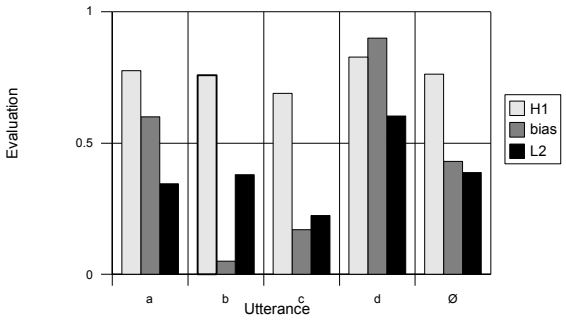

Figure 2: Interpretation of disambiguating contours for modifier attachment: high (H1), textual bias, low (L2), for each 4 test sentences.

The most important result is that a clear disambiguation effect could be found not only for ambiguous sentences without any clear semantic attachment preference, such as (a), but that automatically generated prosody could effectively override even strong preferences for low (b,c) or high attachment (d).

With sentences showing a strong bias for low attachment, we observed that the speech stimuli designed to suggest low attachment do not quite reach the level of the bias. We tentatively attribute this difference to a mismatch between expected and actual prosodic contours in synthetic speech, which can hopefully be overcome on the basis of better prosody planning to be obtained from future experimental studies.

As a measure of the disambiguation effect we take the span between perceived high and low attachment. For H1 and L2 it ranges from 0.23 (=0.83-0.60; utterance d) to 0.47 (=0.69-0.22; utterance c), with an average around 0.37 (=0.76-0.39). The value for H2 (average: 0.63) shows a far lower disambiguation potential than H1, while the value reached by L1 (average: 0.50) proves this contour unsuitable for the task. This latter result confirms for German the findings made in [6] for English, namely that insertion of a pre-object boundary enhances perception of low attachment.

2.2.2. Object Fronting

Results confirm previous findings on prosody-induced early effects, as well as our own claim concerning the disambiguation potential of prosody in speech synthesis. Details are provided in Figure 3, where a value of 1 corresponds to perceived object fronting, and 0 to subject fronting. In contrast to modifier attachment, however, the bias for SVO was extremely high (=0.02 for OVS). Still, by means of carefully designed disambiguating prosody, it was possible, with an average value of 0.43, to make available, for interpretation, a reading that was practically inaccessible with textual stimuli.

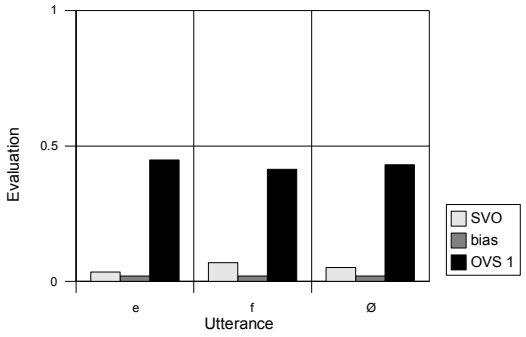

Figure 3: Interpretation of disambiguating contours for object fronting: SVO, textual bias, OVS1

Although the values for unmarked SVO (e: 0.03; f: 0.07; average: 0.05) do not fully reach the bias determined with textual stimuli, we believe that the difference is negligible. The strength of the disambiguation effect, that is, how well prosody distinguishes SVO and OVS interpretation averages at 0.38 for OVS1 and at 0.36 for OVS2. The contours for subject fronting and the best prosodic contour for object fronting (OVS1) are given in figure 4.

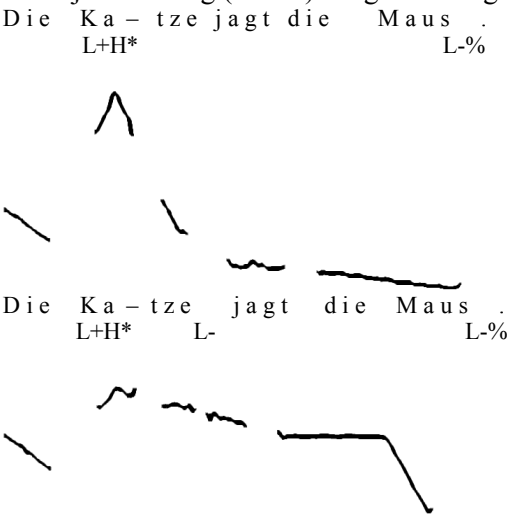

Figure 4: The prosodic contour of SVO (top) vs. OVS (1, bottom) with GToBI-Annotation.

3. CONCLUSION

In this paper we have presented experimental evidence showing how prosody automatically generated from deep syntactic trees can be used successfully to disambiguate structural ambiguities in German. The results we obtain using prosodically enhanced diphone synthesis actually outperform disambiguation rates previously achieved for English modifier attachment ambiguities using human speech stimuli: under forced-choice conditions, [6] reports a value of 0.651 in response to high attachment stimuli similar to our H1, and a value of 0.472 with low attachment stimuli similar to our L2, yielding an overall disambiguation effect of 0.18, compared to our 0.37. The result we obtained with object fronting further suggest that the disambiguating effect (0.38) of our automatically generated prosody is very robust, even against a very strong bias for subject fronting.

The disambiguation effects we obtain with synthesised speech also underline the potential of prosody derived from deep syntactic structures for the improvement of intelligibility in CTS and TTS applications. Finally, the fact that automatically generated stimuli can achieve disambiguation rates comparable to human speech makes our system a valuable test bed for studies at the syntax-prosody interface.

4. REFERENCES

[1] Copestake, A. 2001. *Implementing Typed Feature Structures.* Stanford: CSLI Publications.

[2] Grice, M., Baumann, S. 2002. Deutsche Intonation und GToBI. *Linguistische Berichte* 191, 267-298.

[3] Herwijnen, O. van, Terken, J., Bosch, A. van den, Marsi, E. 2003. Learning PP attachments for filtering prosodic phrasing. *EACL*, 139-146.

[4] McKeown, K.R., Pan, S. 2000. Prosody modelling in concept-to-speech generation: methodological issues. *Philosophical Transactions of the Royal Society* 358(1769), 1419-1431.

[5] Olaszy G., Németh G. 1997. Prosody generation for German CTS/TTS systems (from theoretical intonation patterns to practical realisation). *Speech Communication* 21(1), 37-60.

[6] Schafer, A. 1997. *Prosodic parsing: the role of prosody in sentence comprehension.* PhD dissertation, U Mass.

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

[8] Speer, S., Warren, P., Schafer, A. 2003. Intonation and sentence processing. *Proc. 15th ICPhS* Barcelona, 95-106.

[9] Weber, A., Grice, M., Crocker, M. 2006. The role of prosody in the interpretation of structural ambiguities: a study of anticipatory eye movements. *Cognition* 99(2), B63-B72.