Prosodic Correlates of Contrastive and Non-Contrastive Themes in German

Bettina Braun *D. Robert Ladd*

Saarland University University of Edinburgh
bebr@coli.uni-sb.de bob@ling.ed.ac.uk

Abstract

Semantic theories on focus and information structure assume that there are different accent types for thematic (backward-looking, known) and rhematic (forward-looking, new) information in languages as English and German. According to Steedman [1], thematic material may only be intonationally marked (= bear a pitch accent), if it "contrasts with a different established or accommodatable" theme [p. 656]. We shall show that intonational marking of themes in German seems rather gradual. Themes in contrastive contexts have a significantly longer stressed vowel, a higher and longer rise which results in a higher and more delayed peak than non-contrastive themes. Moreover, speakers can use different strategies to signal the contrast.

Data were elicited by reading short paragraphs with a contrastive and non-contrastive pre-context. The use of many filler texts distracted subjects' attention from the contrast so that the data may be regarded as highly natural. Implementing these prosodic features in speech synthesis systems might help to avoid unnatural exaggerated prosodic realisations.

1. Introduction

Recently, there has been growing interest in integrating prosodic information into semantic formalisms (e.g. [2], [3], [1]). While this greatly improves semantic theory, the prosodic categories employed are not yet well established. Many of the theories rely more on intuitions and introspection than on empirical studies with acoustic-prosodic measurements.

Previous work on contrastive themes in German (by e.g. Wunderlich [4], Büring [3]) has identified a special pitch configuration called 'bridge accent', which is characterised by a rising accent on the contrastive theme, a sustained high pitch and a fall on the nucleus (rheme). This pattern was first described under the name 'hat pattern' by Cohen and t'Hart for Dutch [5]. Mehlhorn et al. investigated the phonetic differences between contrastive and non-contrastive topics[1] in German more closely and found that contrastive topics show a steeper rise, a higher f_0-range and longer syllable duration [6]. We shall describe an exploratory reading study that aims at reproducing Mehlhorn's results with more natural data. Utterances with contrastive and non-contrastive themes are elicited in larger contexts to distract subjects' awareness (see below). These contrast-minimal pairs were analysed phonetically to find reliable dependent variables that can best describe the differences in prosodic realisation.

Hypotheses: In addition to the findings of [6] (higher peak preceded by a steeper rise and longer syllable duration), it is assumed that the peak is reached later for contrastive themes. As Gussenhoven has pointed out, delayed peak can be a substitute for peak height [7].

[1] In this article, the terms 'topic' and 'theme' are used interchangeably. These are assumed to be sentence-initial.

2. Data Elicitation

Many studies employ a question-answer pair methodology for controlling the information structure of the test utterances (e.g. [6]). While this allows for a high degree of control, the purpose of such an experiment can hardly be hidden so that we might expect exaggerated realisations. Furthermore, question-answer pairs have a severe drawback for thematic material: Thematic material has to be present in the question (to be given) which would normally trigger elliptic answers (or reduced grammatical forms as pronouns which can not be used for comparison). The experimental setup, however, forces the subjects to use non-elliptic answers which might obscure results.

Reading studies are better suited to mask the purpose of the study and to ensure naturally produced speech.

2.1. Reading Material

For the reading experiment, short paragraphs (5 or 6 sentences) were constructed. The test utterances appear in roughly the same position in the contrastive and non-contrastive versions. The only difference thus lies in the context. For the non-contrastive context, the theme is present throughout the paragraph. The theme in the test utterance can thus be interpreted as a sort of 'topic-resumption'. For the contrastive context, criteria from Prevost are used ([8]). He argues that the use of two contrasting pairs of discourse entities (of the same type) is a sufficient condition for establishing contrast[2]. Two of our sample paragraphs, translated into English, are:

Non-contrastive theme context: Many Europeans don't know much about Malaysia. The country consists of two islands. To ease the communications between the two parts, almost every household has a computer with Internet access. However, Malaysia is not a highly technological country. *The Malaysians live from agriculture.* They are neither especially poor nor rich.

Contrastive theme context: Malaysia and Indonesia are neighbouring countries in the South China Sea. Despite their geographical adjacency, their living and working conditions differ tremendously. In Indonesia, tourism is very important and many people work in this sector. *The Malaysians live from agriculture.* They have mainly focussed on the cultivation of rice.

In addition to the contrastive and non-contrastive paragraph-pair, distractor paragraphs about the same topic were constructed to distract subjects from the presence of minimal-pair utterances. These distractor paragraphs and other filler paragraphs were intermingled with the test paragraphs.

To get a wide variety of data and to explore possible influencing factors, the test utterances exhibit 3 different word orders

[2] It might be argued that even the "non-contrastive" context involves some degree of contrastiveness in which case the two conditions can be described as "more contrastive" and "less contrastive". This affects the interpretation but not the validity of the acoustic distinction.

(4 x subj-NP initial, 4 x PP-initial, and 4 x existential sentence). The test words have three main stress patterns (7 x initial stress, 7 x stress on 2nd syllable, 7 x stress on 3rd syllable).

2.2. Recording Procedure

Eleven native naïve German subjects (mostly postgraduate students at the University of Edinburgh) voluntarily participated in the recording. Due to the restricted choice of subjects, the dialect origin of the speakers could not be fully controlled, so there is a bias towards northern German speakers (eight northern Germans vs. three southern Germans).
Subjects were seated in a sound-proof room in the Linguistics Laboratory of Edinburgh University. They were given a pile of 52 A5 cards, each containing one paragraph and written instruction to read the paragraphs at normal speed as fluently as possible. They were told that they could have breaks between paragraphs and that they could scan the paragraphs before reading them aloud. Permission to stay with them in the recording room was obtained in order to ask for repetitions of paragraphs in cases with too many misreadings and slips of the tongue. The overall recording procedure lasted between 20 and 30 minutes. The presentation of the paragraphs was block-wise randomised, maintaining the order fillers, stimuli, distractors, stimuli. There were four different randomisations. To disguise the contrast-minimal pairs, the respective paragraph pairs are separated by at least eight other paragraphs. Data were digitised with a sampling rate of 44kHz.

2.3. Evaluation

Not all subjects were equally good readers. Two poor readers with many mispronounciations and hesitations were excluded from further analysis.
Besides these two overall exclusions, some individual minimal pairs had to be discarded. One was due to severe hesitation in a non-contrastive token where sentence planning could not be regarded as completed. Another sample was excluded because of bored, impatient attitude that strongly influenced the prosody. Two test utterances were badly designed in that they led to stress clashes (with de-accented themes). Furthermore, test words with initial stress were excluded because they did not allow to investigate certain parts of the pitch contour. This data selection leaves 83 minimal pairs for phonetic analysis.

3. Analysis

Data are analysed using xwaves; f0-tracking was conducted with the in-built pitch-tracker (get_f0). Artifacts introduced by the pitch-tracking algorithm (pitch doubling or halving) were manually corrected. Missing f_0-values were linearly interpolated. Then, the pitch-contour was smoothed using a 7-frame window (7.5ms each) with mean smoothing.

3.1. Labelling Procedure

Data annotation was done on the segmental and suprasegmental level, concentrating around the area of the f_0-rise. To illustrate the annotation procedure, which is crucial for all further analyses, the label points are summarised in figure 1, including suprasegmental, segmental, and lexical labels.

3.1.1. Segmental Landmarks

On the segmental level, four landmarks were labelled. Since the test words consisted almost entirely of sonorant sounds (to ensure a smooth f_0-contour), segmentation was sometimes difficult. Label points were always set at the positive f0-crossing, using information from a wide-band spectrogram:
C0: Start of the stressed syllable
V0: Start of the stressed vowel
C1: Start of the first post-stressed syllable
V1: Start of the unstressed vowel following the stress

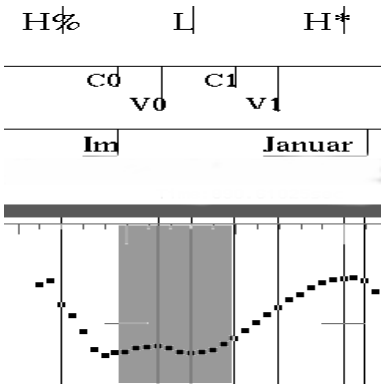

Figure 1: *Suprasegmental and segmental labels, together with lexical information ("im* JA*nuar", with shaded stress). Lexical labels mark the end of words. Segmental labels mark the start, suprasegmental labels the target point.*

3.1.2. Suprasegmental Events

On the suprasegmental level, the following events in or before the test words were marked. In uncertain cases, the same criteria were always chosen for both items of the contrast-minimal pairs. Note that these labels are not meant to correspond to (standard) ToBI labels:
H%: High point before the fall. In most cases this value was found in the middle of the vowel of the first unstressed syllable of the prosodic word. If this value was not reliable (e.g. glottalisation, devoicing), the value in the following sonorant was taken (often the case in PPs beginning with 'in' or 'im'). Otherwise, if there was no reliable value in the first unstressed syllable (often the case with the definite article "die"), the value of the unstressed syllable preceding the stressed one was used.
L: Local minimum near the rise. This label was extremely difficult to assign, because the valleys may be quite broad (as in figure 1) or the local minimum found only in consonantal areas or well before the stressed syllable. There, local pitch perturbations may influence the labelling procedure. In some rare cases the contour is monotonously increasing, i.e. there is no fall to an L-point.[3]
H*: First local maximum after the stressed syllable. In rare cases where the rise proceeded into the following word, the H* was nevertheless assigned within the test word[4].

[3] In order not to lose possibly important information, two elbow points were labelled that mark a change in slope:
E1: This point marks a considerable change in the slope of the fall. As already pointed out for H%, the fall was often hard to detect. Therefore, E1 was an optional label that was only assigned if a) there was a broad valley, where E1 marks the start of the broad valley or b) if there were two dips in the f_0-contour, a case that was also interpreted as a broad valley (see figure 1).
E2: Point where the rise starts, i.e. a considerable change in slope. This label may coincide with L.

[4] This criterion is rather strict and needs further investigation.

3.2. Phonetic Parameters (= Dependent Variables)

The aim of the production study was to explore which variables might be meaningfully used to distinguish contrastive from non-contrastive themes. There are four groups of variables[5]:

F_0 Variables: The most obvious feature in intonation research is f_0. Five f_0 variables were analysed, three static (f_0(H%), f_0(L) and f_0(H*)) and two dynamic ones (Δf_0(fall), Δf_0(rise)).

Temporal Variables: Besides f_0 variables, the temporal organisation may differ. Three variables were extracted from the data, duration of the stressed vowel Δt(V0), and duration of the fall and the rise (Δt(fall) and Δt(rise)).

Alignment: The above variables are restricted to information from only one tier, segmental or suprasegmental. The alignment variables [9] represent a link between these auto-segmental tiers, insofar as they encode the temporal alignment of suprasegmental events with respect to the segmental structure.

Since the anchor points of suprasegmental events are not yet well understood, different alignment variables were explored. For the peak, alignment was calculated to the start and the end of the stressed vowel and to the start of the first post-stressed vowel: al(H*,V0), al(H*,V0end) and al(H*,V1). It is hypothesised that al(H*,V1) is the best predictor because the peak in German is rather late. The alignment of the valley was calculated to the beginning of the stressed vowel and to the start of the post-stressed syllable: al(L,V0), al(L,C1).

Apart from the alignment relative to the start and end of the stressed vowel, the comparison of alignment data is obscured if the syllables following the stress have different structures (e.g. CV vs. V). In almost all test words, the post-stressed syllable was of the type CV, i.e. segmental labelling had the order C0–V0–C1–V1. The last two labels were, however, reversed in the case of *Malayen* and *Bayern*. These data were therefore not taken into account for the alignment variables that do not calculate the temporal difference to V0. Similarly, test words where the *schwa* in the post-stressed syllable was deleted would have caused artifacts (e.g. *Mormonen*). This reduced the amount of data for these variables to 67 contrast-minimal pairs.

Derived Variables: Besides the basic phonetic variables, the slope of the rise was calculated by dividing the f_0-range by the duration (slope(rise)).

4. Results and Discussion

In this section we first assess the predictive power of the dependent variables. Secondly, we investigate whether there are interactions between the dependent variables which may be important for the interpretation of suprasegmental events. In the last part we discuss speaker idiosyncrasies.

4.1. Descriptive Power of Dependent Variables

Since the phonetic variables were highly correlated and each of the speakers produced both contrastive and non-contrastive themes, a paired t-test was preferred over a discriminant analysis. Those phonetic variables that showed significant difference between contrastive and non-contrastive themes were interpreted as reliable variables to encode the distinction. Due to multiple t-tests (for the 21 variables, including the elbow variables), the standard significance level of p=0.05 was adjusted to p=0.0024 (Bonferroni correction). The overall results of the paired t-test for the four groups of variables are shown in table 1[6], together with the averaged mean values over all speakers. The results are discussed below:

Table 1: *Overall means of variable values in contrastive and non-contrastive context, including number of samples (alignment variables that relate to C1 and V1 are analysed excluding* Malayen *and* Bayern*) and significance value of the paired t-test ('ns' meaning non-significant on p=0.0024).*

variable	#	non-contr.	contr.	p
f_0(H%)	83	164.4 H z	163.7 Hz	ns
f_0(L)	83	153.2 Hz	150.4 Hz	ns
f_0(H*)	83	214.1 Hz	223.2 Hz	0.002
Δf_0(fall)	83	11.3 Hz	13.2 Hz	ns
Δf_0(rise)	83	60.9 Hz	72.7 Hz	0.000
Δt(V0)	83	96.8 ms	104.4 ms	0.002
Δt(fall)	83	107.8 ms	119.7 ms	ns
Δt(rise)	83	181.9 ms	200.6 ms	0.001
al(H*,V0)	83	170 ms	219 ms	ns
al(H*,V1)	67	71.6 ms	93.8 ms	0.000
al(H*,V0end)	83	-6.1 ms	36.5 ms	ns
al(L,C1)	67	109.2 ms	109.1 ms	ns
al(L,V0)	83	-11.5 ms	19.1 ms	ns
slope(rise)	83	0.34	0.37	ns

F_0 Variables: Descriptively, all f_0 variables except for f_0(H%), behave according to the hypotheses. A higher f_0(H%) was expected for contrastive themes because this would have emphasised the fall. This deviation may be caused by the difficulty of reliably assigning H%. Although f_0(H%) behaves contrary to expectation, the f_0-range of the fall is nevertheless larger for contrast. As expected, contrastive themes have a lower valley and a higher peak which results in a more expanded rise. The peak height and the range of the f_0-rise differ significantly for contrast and non-contrast.

Temporal Variables: We expected that the duration of the stressed syllable, as well as the duration of the fall and the rise would be longer for contrastive themes. Descriptively, this is reflected in the means. But only the duration of the stressed vowel and the duration of the rise (Δt(rise)) are significantly different.

Alignment Variables: It was hypothesised that the peak is aligned later for contrastive themes. This tendency emerges from all three H*-alignment variables, but only the variable that calculates the alignment to V1 is significantly different. This is because the peak in German is only found in the post-stressed syllable. Calculating the alignment of the peak with respect to the start or end of the stressed vowel (as in the variables al(H*,V0, al(H*,V0end)) therefore introduces more variation since there is more segmental material between H* and the segmental anchor point.

Neither of the explored L-alignment variables reach significance, i.e. the L-alignment is not significantly different in contrastive and non-contrastive themes. It might be assumed that L has a rather stable anchor point in the syllabic structure which was also found in the study of [10]. This assumption is evaluated in more detail in section 4.2.

Derived Variables: While descriptively there is a tendency to a steeper rise for contrast, which is according to the hypothesis, this is not consistent enough to be reflected in the statistics.

[5] All variables that were calculated with respect to L were also calculated with respect to E2, see footnote 3.

[6] The variables related to the elbow points E1 and E2 do not appear in the table because they showed no significant difference at all.

4.2. Interactions between Dependent Variables

From the descriptive statistics and the paired t-test alone, we do not learn whether the dependent variables are correlated or independent. Some correlations that are very straightforward (e.g. peak height and Δf_0(rise)) are not investigated here. Some less obvious relations, however, are interesting for the phonetic interpretation of tonal events: e.g. has the delay of $f_0(H^*)$ something to do with the height of the peak or are these two variables adjusted independently? In cases where more variables were explored, the variables with the best significant values in the paired t-test are used (e.g. al(H^*,V1) for peak alignment).

Because the raw data are subject to much (unwanted) variation, such as different f_0-level or speech rate, the correlation analyses are based on the *ratios* between the contrastive and non-contrastive values of the variables.

Peak height significantly correlates with the slope of the rise (r=.55, p=0.000) and with the height of L (r=0.46, p=0.000) but not with peak alignment. That is, the height of the peak is controlled independently from the alignment and is accompanied by a steeper slope and a higher L.

Peak alignment, on the other hand, correlates with the height of L (r=-.31, p=.009), but *not* with the alignment of the L, nor with the slope. That is, a later peak has approximately the same slope as an early peak. This later peak is not achieved by changing the position of L, but by lowering it.

4.3. Speaker Idiosyncrasies

The labelling phase revealed that speakers use different strategies to encode the contrast: some make heavier use of f_0-range, others of alignment (which is in line with Gussenhoven [7]). As noted above, the correlation analysis showed that height and alignment of the peak are uncorrelated. In order to further investigate the speaker strategies, the ratios between contrastive and non-contrastive values of $f_0(H^*)$ and al(H^*,V1) for each speaker were plotted against each other. In the case of a trading relation between these two derived variables we would expect a negative line, indicating that subjects who make heavy use of range do not vary the alignment and vice versa. And this is indeed the case as shown in figure 2.

There is one outlier where both the ratio of peak heights and the ratio of peak alignments were close to one, i.e. she did not prosodically distinguish contrastive from non-contrastive themes. Without this speaker, we get a significant correlation (r=-0.77, p=0.027) which statistically corroborates the trading relation between peak height and peak alignment.

The different speaker strategies may be the reason why some of the variables do not differ significantly for contrastive and non-contrastive themes (e.g. slope(rise) and f_0(L)).

5. Conclusions

The present study shows that sentence-initial marked themes in contrastive contexts are prosodically distinguished from those in non-contrastive contexts, most importantly by peak height and alignment, range and duration of the rise, and duration of the stressed vowel. This finding is especially significant, given that the subjects were not aware of the contrast-minimal pairs. The relation between contrast and prosodic features (and between the prosodic features themselves) is rather complex. Furthermore, the observed trading relation between peak height and peak alignment may undermine the common assumption of a 1:1 mapping of ToBI-style pitch accents to semantic function (as in [1]). It rather suggests that prosodic features have to

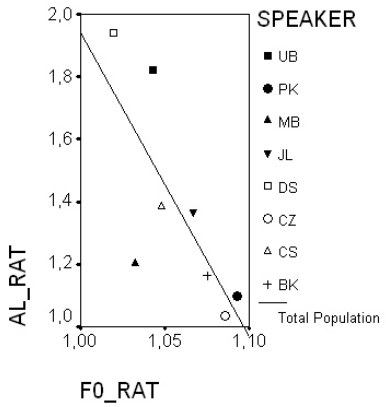

Figure 2: *Correlation between the ratio of $f_0(H^*)$ (contrastive:non-contrastive) and the ratio of al(H^*,V1)' (contrastive:non-contrastive). One speaker was excluded because she did not realise a difference between the two contrast conditions.*

be carefully selected for speech synthesis to avoid exaggerated contours. The predicted bridge accent is also not observed in all contrastive cases and could be seen as an extreme configuration which can be gradually weakened depending on the context.

It might be argued that the pre-verbal position is inherently contrastive if it contains a lexical word and not a pronoun. This would partly explain the gradual marking of contrast which might be related to continuous discourse functions (e.g. topic-resumption, topic-change, contrastive topic). Contrast could, however, be *perceived* categorically. Perception studies that assess this are in progress.

6. References

[1] Steedman, M. "Information structure and the syntax-phonology interface", Linguistic Inquiry 31/4, 2000, p 649–689.

[2] Rooth, M. "A theory of focus interpretation", Natural Language Semantics 1, 1992, p 75–116.

[3] Büring, D. "The Meaning of Topic and Focus – the 59^{th} Street Bridge Accent", 1997, Routledge, London.

[4] Wunderlich, D. "Intonation and Contrast", Journal of Semantics 8, 1991, p 239–251.

[5] Cohen, A., J. 't Hart. "On the anatomy of intonation", Lingua 19, 1967, p 177–192.

[6] Mehlhorn, G. "Produktion und Perzeption von Hutkonturen im Deutschen", Linguistische Arbeitsberichte 77, 2001, p 31–57.

[7] Gussenhoven, C. "Intonation and Interpretation: Phonetics and Phonology", Proc. Prosody 2002. p 47–57.

[8] Prevost, S. "A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation", PhD thesis, 1995.

[9] Arvaniti, A., D. R. Ladd, I. Mennen. "Stability of tonal alignment: The case of Greek prenuclear accents", Journal of Phonetics 26, 1998, p 3–25.

[10] Caspers, J., V. J. van Heuven. "Effects of time pressure on the phonetic realisation of the Dutch accent-lending pitch rise and fall", Phonetica, 50, 1993, p 161–171.