THE SYNTHESIS OF GERMAN CONTRASTIVE FOCUS

Petra S. Wagner
Institut für Kommunikationsforschung und Phonetik, University of Bonn

ABSTRACT

Using a prominence–based speech synthesizer of German, different patterns of prosodic prominence have been examined as to whether they are capable of consistingly creating an impression of "contrastive focus".

The results confirm the conclusions of previous phonological and phonetic analyses, that the impression of contrast is best explained by a specific prosodic pattern. Such a pattern can be characterized by postfocal "deaccentuation" and/or a high perceptual prominence on the contrastively stressed syllable, the latter correlating with a (high) pitch accent and increase in duration. Thus, a notion of *perceptual prominence* as a relational parameter ought to be able to model contrastive focus, because it is able to capture such contextual phenomena.

1. CONCEPTS OF CONTRASTIVE FOCUS

Within the many competing definitions of focus in the semantic, syntactic and phonological literature, a relatively shared assumption is that a contrastively focussed constituent gives an alternative answer to an explicit or implicit statement provided by the previous discourse/situation. The previous statement is often illustrated as a question, the contrasting alternative as an answer.

(1) So Anna finds Benjamin attractive?

(No,) Anna finds [Catherine]$_{CF}$ attractive.

Whether the kind of prosodic marking involved in this contrastively marked utterance in (1) differs from other types of focus (e.g. "newness focus") is still an issue of debate [12].

A different but related kind of prosodic pattern which is often called a *contrastive* or *multiple focus*, occurs in coordinate sentences where main stress falls on the contrasted alternatives.

(2) An [OLD]$_{CF}$ house is more charming than a [NEW]$_{CF}$ house.

The study presented here concentrated on the type of contrast described in (1), which we also call *correction contrast* and leaves aside for now the prosodic structure given in (2). A purely semantic analysis would treat *correction contrast* (1) and *coordinate contrast* (2) alike [11]. However, they appear to differ pragmatically: In (1), the speaker introduces a contrast to the discourse *and* wants to actively override an element of what (s)he believes to be the addressee's informational state. The concentration on (1) in this paper embraces previous conceptions of contrastive stress/accent [4, 8].

Our first question was to find out, whether a contrastive focus can be adequately signalled using a prominence based approach to speech synthesis. A further question was, if it can, which prominence patterns are the most successful.

1.1 Contrast in Phonology.

Even though the existence of a unique phonological pattern characterizing contrastive focus as we defined it above, is still very controversial (e.g. see the discussion in [12]), several suggestions in favour of such a pattern do exist in the literature. Often, it has been identified with a $L + H^*$ pitch accent for both English [8] and German (e.g. [6]). Other researchers suggest a postfocal metrical deaccentuation for contrast in German [4] or an operation on the metrical tree called "contrastive relabelling" (see [2] for English and Polish) also resulting in a metrical weakening of postfocal accents and strengthening of focal ones. Previous experiments for Dutch have shown that the impression of contrast depends on the prosodic environment and cannot be perceived in isolation [12]. Therefore, a theory that deals with contrastive stress as a purely local feature of one syllable or pitch accent might not be appropriate.

1.2. Contrast in Phonetics.

Acoustic analyses of semantic/pragmatic concepts are dangerous, because they allow to formulate principles without reference to the phonological domain of meaning differentiation [5]. Still, such approaches do exist and have led to the result that in English, a correction contrast is characterized acoustically as an increase in duration on the focal word plus a postfocal flattening of the F0–contour [1].

A more useful approach would be to find quantitative/phonetic correlates of *phonological* representations

> The case for quantitative representations in the mind appears much stronger than the case for discrete representations of the speech signal. Progress needs to be made on formulating such representations and understanding their relationship to the qualitative representations of current phonological theories.[7, 391]

As the correspondences between meaning differentiating phonological representations and the acoustic level are extremely complex, a simpler representational level which is able to quantify over phonetic representations is necessary. A good candidate for such a level is the concept of *perceptual prominence* which will be dealt with in the next section.

2. PERCEPTUAL PROMINENCE IN SPEECH SYNTHESIS

The term *prominence* has been given a quantification by [3], resulting in its definition as a measure of perceptual markedness relative to the surrounding phonetic context.

The appeal of this approach lies mainly in the possibility to have an easy description (here: a scale between 0 and 31) of the prosodic characteristics of an utterance reflecting the perceptual impressions instead of directly describing the acoustic correlates of phonological concepts. The link between perceptual prominences and acoustic realisation has been studied and integrated into the prosodic component of a rule based synthesis system [9]. This approach seems to be especially interesting in the study of prosodic focus, since contextual parameters influencing the perception can be controlled in a comparably straightforward way. Previous studies already indicate a possibility of expressing narrow focus using the prominence approach in speech synthesis [13, 9]. In [13], indications were found that subjects preferred to perceive contrastive focus in sentences synthesized with high prominence values on the syllable carrying the main stress of the focal word. Still, the correlation between (absolute) prominence values and perceived contrastive focus was low. We concluded that not only the focal syllable but also the context needed some attention.

3. A CONTRAST EXPRESSING PROMINENCE PATTERN

3.1 The Experiment

Three different declarative sentences were synthesized with contrastive stress on three different positions in each sentence, using five different prominence patterns for each constellation. The most prominent syllable within the focus exponent was chosen as focal syllable. The original prominence values were taken from the VERBMOBIL generation module which calculates prominence values using syntactic and lexical information, if no further semantic/pragmatic information is given. The result is a default prosodic pattern. Those patterns were manipulated according to the methods illustrated in Table 1. The following stimulus sentences were used (SMALL CAPITALS indicate all the possible locations of intended contrastive focus).

ENde MAI bin ICH noch im Urlaub. End of May am I still on vacation.
Es würde mich freuen, wenn WIR noch EInen TerMIN ausmachen It would me please if we another one appointment made.
ANfang MAI hätte ICH noch Zeit. Beginning of May would have I still time

Each sentence was further supplied with two question contexts. The first context matched the contrastive accentuation pattern in the answer, the second context ought to produce an odd (if not ungrammatical) impression if the accent were interpreted as a correction contrast. The following sentences are examples for a such a contrast in a matching (3) and one in a non-matching context (4).

(3) Q: Anfang Mai sind Sie also noch im Urlaub?
Q: Beginning of May are you so still on vacation

A: ENDE Mai bin ich noch im Urlaub
A: END OF May am I still on vacation

(4) Q: Ende Juni sind Sie also noch im Urlaub?
Q: End of June are you so still on vacation?

A: ??ENDE Mai bin ich noch im Urlaub.
A: END OF May am I still on vacation.

The context questions were read aloud and recorded in an anechoic chamber by a male native speaker of German and phonetic expert, who was familiar with the experiment. He was instructed to read the questions as if no context was given to prevent any bias towards a specific prosodic expectation. The resulting 180 question-answer pairs were presented to phonetic experts (n=11). Both question and answers were played via headphones, the questions were furthermore displayed on a computer screen. Subjects were allowed to listen to each answer several times and had to judge each question-answer pair on a scale between 1 and 6 (forced choice), with 1 being a very good and 6 being a very bad score (German school grades). Due to a (now fixed) mistake in the algorithm, one stimulus type (sentence 2, intended focus on *einen*) did not properly reflect the intended conditions and was thus eliminated from all further studies.

3.2 Results

Contexts matching the intended impression were rated significantly more acceptable than nonmatching ones (Kolmogorov-Smirnov, $p \leq 0.001$). Besides, there was a substantial negative correlation between the judgements and the matching vs. non-matching question-answer pairs ($\rho = -0,49$, $p \leq 0.01$) without taking into account any specific strategy of prominence manipulation. Figure 1 illustrates the distribution of judgements for matching vs. non-matching question-answer pairs.

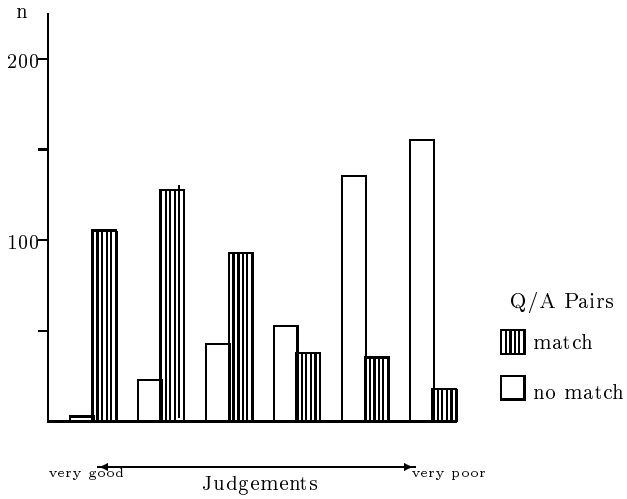

Figure 1: Judgements for matching and non-matching question-answer pairs

In the subsequent analysis, only the "matching" question-anwer pairs were examined. Judgements were significantly better for conditions 1 and 2, where a combination of strategies (high prominence on contrastive syllable *plus* deaccentuation *plus* additional duration manipulation) was employed (Kolmogorov-Smirnov, $p \leq 0.01$). Figure 2 shows, how the good judgements (1 and 2) spread over the different conditions.

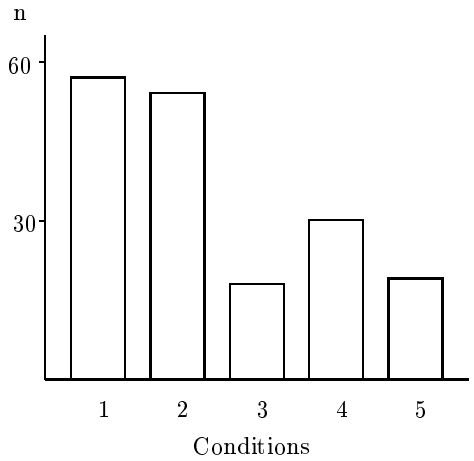

Figure 2: Number of good and very good judgements (1 or 2) for the different conditions

It was impossible, however, to isolate a single factor determining the impression of contrast; the judgements for individual strategies (either deaccentuation, durational manipulation or high prominence) were all worse than for the combinational strategies but were not rated significantly different among each other. Another finding was that a pre-*plus* postfocal deaccentuation (condition 1) did not yield significantly better judgements than a postfocal deaccentuation alone (Kolmogorov-Smirnov, $p = 0.44$).

Further tests were run in order to isolate other factors influencing the subjects' judgements. No significant impact of the sentences were found; however, placement of the focal syllable proved to be important. Subjects rated sentences where a sentence-final focus was intended significantly worse than foci in other environments (Kolmogorov-Smirnov, $p \leq 0.01$). Such positional effects have ocurred before [13, 10] and need further attention in the definition of realization rules.

3.3 Discussion

Our results indicate that a combination of contextual deaccentuation *plus* a further increase in duration of the focal syllable yields the best results. The isolation of a single influential factor was impossible. The need for a further increase of duration in order to achieve the impression of a contrast is probably explained by the prominence realization rules within the synthesis system. These rules were originally developed within a TTS-applications. Consequently, they were tuned to model "default prosody". Correction contrast, however, may be regarded as a prosodic deviation that is currently not adequately modelled. An adaption of the prominence realization rules appears to be necessary. Another solution to this problem may be an extension of the prominence scale beyond the (so far) maximal value of 31. The need for an improvement of the durational model receives further support by the result that the position of the contrastive focus had an impact on listeners' judgements. This indicates that the relatively simple manipulation used for an additional durational enhancement was not completely appropriate and needs further improvement, even though it resulted in a more successful auditory impression.

An important result was, that the scope of deaccentuation appears to be postfocal in accordance with previous analyses [4, 2, 1] and prominence–based synthesis is in priciple able to model results of those phonological and acoustic analyses. It cannot be concluded, however, that the most successful method would exclusively lead to an impression of contrastive focus, because subjects were not asked to identify the type of focus but the appropriateness of a specific interpretation.

4. CONCLUSION

Simple question-anwer pairs triggering contrastive foci were rated by subjects who had to determine to what extent they match or not. The stimuli were created using different prominence patterns using synthetic speech which has been particularly designed to model those.

The experimental results clearly indicate a preference for those environments where a deaccentuation of the postfocal domain and a further durational increase of the focal syllable was employed. Thus, further evidence has been retrieved that the domain of deaccentuation is postfocal. Prefocal deaccentuation does not play a role in the perception of contrast. Our study sheds light on previous attempts to isolate a specific "contrastive accent" as it appears to be the case that the impression of contrast is neither a local nor a purely intonational phenomenon but involves several

Condition 1	Prominence of the focal syllable was assigned the highest possible value (31). Additionally, the prominence was enhanced by manipulating the duration of the focal syllable, and the remaining accented syllables in the prosodic phrase were deaccented by reducing their prominence values.
Condition 2	Prominence of the focal syllable was assigned the highest possible value (31). Additionally, prominence was enhanced by manipulating the duration of the focal syllable, and the accented syllables occurring *after* the focal syllable in the prosodic phrase were deaccented by reducing their prominence values.
Condition 3	Prominence of the focal syllable was assigned the highest possible value (31). Additionally, prominence was enhanced by manipulating the duration of the focal syllable. No deaccentuation was involved.
Condition 4	Prominence of the focal syllable was assigned the highest possible value. No further manipulations were used.
Condition 5	Prominence of the focal syllable was assigned the highest possible value and the accented syllables following the focal syllable were reduced in prominence.

Table 1: The different conditions used for synthesizing contrastive focus

factors. These factors can be modelled using the concept of perceptual prominence as it has been defined in [3].

However, the answer of this study is only a partial one, because it was not examined whether the identified pattern exclusively triggers the impression of contrast.

ACKNOWLEDGEMENTS

This research was funded by the German Federal Ministy of Education and Research under grant 01 IV 101 G within the VERBMOBIL project. In first place, I would like to thank all my patient subjects. Helpful comments on this paper and/or the research involved were provided by Thomas Portele, Gunnar Fant, Anita Kruckenberg, Berhard Schröder and many other people.

REFERENCES

[1] Cooper, W.E., S.J. Eady and P.R. Mueller (1985). *Acoustical Aspects of contrastive stress in question—answer contexts* JASA *77*,(6), 2142—2156.

[2] Dogil, G. (1979). *Autosegmental Account of Phonological Emphasis.* Current Inquiriy into Language and Linguistics *25*. Alberta: Edmonton.

[3] Fant, G. and A. Kruckenberg (1989). *Preliminaries to the study of swedish prose reading and reading style.* STL-QPSR2/1989, 1—68.

[4] Caroline Féry (1988). Rhythmische und tonale Struktur der Intonationsphrase. H. Altmann (ed.). *Intonationsforschung.* Tübingen: Niemeyer, 41—64.

[5] Ladd, D.R, (1993). An introduction to intonational phonology. G. D. Docherty and D.R. Ladd (eds.). *Papers in Laboratory Phonology II. Gesture, Segment, Prosody.* 321—334.

[6] Reyelt, M., M. Grice, R. Benzmüller, J. Mayer and A. Batliner (1996). Prosodische Etikettierung des Deutschen mit ToBI. D. Gibbon (ed.). *Natural Language Processing and Speech Technology. KONVENS 96, Bielefeld.* 144—155.

[7] Pierrehumbert, J. (1990). Phonological and phonetic representation. *Journal of Phonetics 17*, 231–278.

[8] Pierrehumbert, J. and J. Hirschberg (1990). The Meaning of Intonational Contours in the Interpretation of Discourse. *Intonations in Communication* P. Cohen, J. Morgan & M. Pollack (eds.) Cambridge MA: MIT Press. 342—369.

[9] Portele, T. and B. Heuft (1997). Towards a prominence–based speech synthesis system. *Speech Communication*, *21*, 61—72.

[10] Portele, T. (1999). A perceptually motivated intonation model for German. this volume.

[11] Rooth, M. (1995). Focus. S. Lappin (ed.) *Handbook of Contemporary Semantic Theory.* Blackwell:Oxford, 271—297.

[12] Krahmer, E. and M. Swerts (1998). Reconciling two competing views on contrastiveness. *Proceedings of ICSLP '98, Sydney.*

[13] Wolters, M. and P. Wagner (1998). Focus Perception and Prominence. B. Schröder, W. Lenders, W. Hess and T. Portele. *Computers, Linguistics, and Phonetics between Language and Speech. Proceedings of KONVENS '98, Bonn.* Lang: Frankfurt. 227—239.