

A Discourse Model for Pitch-Range Control

Gregor Möhler, Jörg Mayer

Institute of Natural Language Processing
University of Stuttgart, Germany

{moehler, joemayer}@ims.uni-stuttgart.de

Abstract

The width and the position of the pitch range reveals important information about the structure of a spoken discourse. This paper studies the correlation between the pitch range and the discourse structure based on a large database. The model used to analyze the discourse is based on a two-level description of registers. Primary register features reflect the prosodic phrasing within a discourse segment. The secondary register features depend on the relations between the discourse segments, more specifically the topic structure of the discourse. The pitch-range is automatically extracted from a speech database with the help of an F_0 parametrization. This study shows that different registers exhibit pitch range values that differ clearly in position and width. These results can be used to successfully implement a global prominence model within a speech synthesis system. The ideal application is concept-to-speech, where discourse information is in principle available on the input side.

1. Introduction

Longer discourses are structured. Their building blocks, the discourse segments are tied together by a variety of specific semantic relations. Such relations are for example continuation, elaboration, contrast, etc. It has been shown that the prosodic realization of spoken discourse reflects to some degree the inherent semantic structure of the discourse. Both pause duration and the pitch range correlate strongly with the coarse topic structure of a discourse [1, 2]. Whereas most of these studies are based on relatively short stretches of speech from few speakers we are aiming at a model of discourse structure that can be evaluated on a larger speech database. Furthermore, our goal is to establish a relation between the discourse structure and its prosodic realization in terms of pitch range.

In the next section we introduce a phonological model of register. The register is assumed to be a phonological entity, which is phonetically interpreted as pitch range. It reflects semantic relations among discourse segments as described in [3] and [4]. The formal interface between register and dynamic semantics was developed in [5]. In the subsequent sections we describe the data analyzed in this work and the method used for automatically extracting the pitch range from the data. We then present how specific registers are realized in terms of pitch range. This paper extends our earlier study of registers [6].

2. Global prominence and pitch range

In the model described in [5] the pitch range of a speaker is phonologically analyzed as being divided into two categorical register levels, low $\{l\}$ and high $\{h\}$. These underlying primary register features are associated with intonation phrases and they reflect the position of the intonation phrase within a discourse

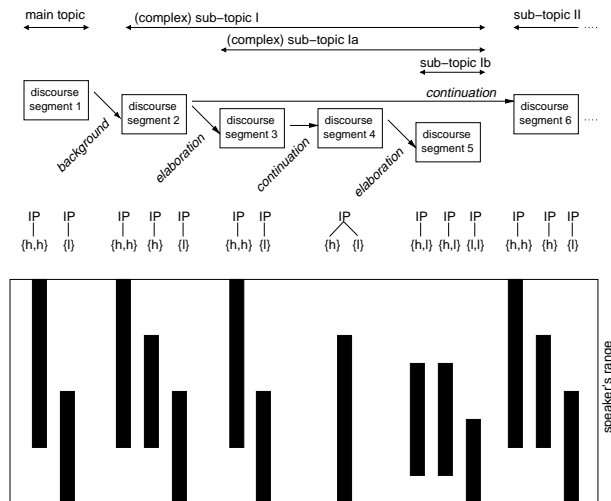


Figure 1: Overview of the global prominence model.

segment. Segment-final phrases are associated with the established finality marker $\{l\}$, whereas non-final phrases are associated with the openness marker $\{h\}$. Every intonation phrase is obligatorily associated with one primary register feature. If a short discourse segment is realized with only one intonation phrase, the phrase is double-associated with both features, $\{h\}$ and $\{l\}$. The phonetic interpretation of register features is the width and the position of the pitch range of an intonation phrase with respect to the speaker's overall range. Primary $\{h\}$ is realized in the higher portions of the speaker's range, primary $\{l\}$ is realized in the lower half. Phonological registers may and do overlap in their phonetic realization, i.e. low targets in $\{h\}$ -marked phrases are realized with lower F_0 values than high targets in $\{l\}$ -marked phrases.

Primary registers can be modified, yielding position variation (up-shift, down-shift) or width variation (expansion, compression). In the autosegmental model, these modifications are represented as secondary register features. Other than primary features, which depend on the internal phrasing of discourse segments, secondary register features reflect semantic relations among discourse segments. According to [3] an important aspect of some discourse relations (e.g. elaboration, background) is *discourse dominance*, leading to a hierarchical representation of the topic structure of discourse. In the model in [5] this semantic hierarchy is prosodically expressed by means of secondary registers. The first phrase of a discourse segment introducing a new topic or the first phrase of a discourse segment in a sequence of dominated segments (sub-topic) is realized in a

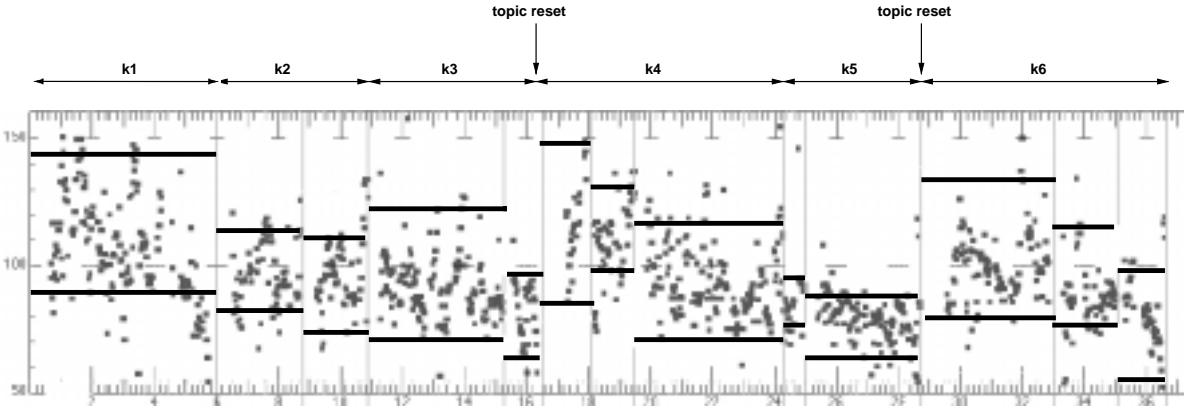


Figure 2: Pitch contour for one news message containing 6 discourse segments (k_1 - k_6). The pitch range of each intonation phrase is marked with horizontal lines. The topic reset between the segments k_3 and k_4 and between k_5 and k_6 is clearly visible.

high register expanding to the top of the speaker’s range. This is represented as $\{h,h\}$ (first symbol: primary, second symbol: secondary register feature). All phrases are modified with a secondary low register in discourse segments (1) at the end of a sequence of dominated segments (topic change) or (2) at the end of a main topic (topic change); $\{h,l\}$ is realized as down-shifted h-register, $\{l,l\}$ as compressed l-register. Primary low with secondary high register ($\{l,h\}$), realized as up-shifted l-register, is supposed to occur only in dialogues (e.g. to express turn-giving) and not considered in this study. Figure 1 gives an overview of the various aspects of the global prominence model.

3. Data description

This study is based on the IMS German Radio News Corpus recorded from satellite broadcast [7]. From the database we used 48 minutes of news messages read by a professional male news speaker, and 9 minutes read by a female speaker. The recordings were segmented into single news messages. One message consists of 13 intonation phrases on average. The speech data was augmented with phonetic and syllabic transcriptions using techniques of forced alignment. The prosody was manually annotated based on the German ToBI transcription system [8].

The discourse structure of each news story in the database was manually annotated by one labeler under the supervision of one of the authors following the discourse theoretical framework of Mann & Thompson [4] and Asher [3]. Based on the textual representation the labeler divided the news story into discourse segments and determined the discourse relations between the segments. The discourse relations are of the type $rel(k_a, k_r)$. In this form k_a is the actual discourse segment that is related to a segment k_r earlier in the discourse. k_r is said to be the *discourse reference* of k_a . Here are the characteristics of the nine discourse relations used in for labeling.

- *new topic*. The discourse segment k_a introduces a new topic to the discourse world. There is no discourse reference for a new topic.
- *background*. A link between the fact k_a and a fact k_r that has already existed before (*link relation*). The segment k_a increases the ability of the reader to comprehend k_r . k_a is dominated by k_r .

- *explanation*. A link relation between the two discourse segments k_a and k_r . More specific than *background*, because the relationship is causal. k_a is dominated by k_r .
- *elaboration*. The discourse universe (which contains k_r) is expanded by k_a (*expansion relation*). k_a provides additional detail and is dominated by k_r .
- *comment*. k_a is dominated by k_r , but adds no additional information (e.g. “. . . , the politician said.”)
- *contrast*. k_a and k_r are comparable facts, but have at least one difference.
- *continuation*. k_a follows on k_r . Both segments have the same topic and the same relation to all other segments dominating them.
- *condition*. k_r is only valid if k_a is valid.
- *consequence*. k_a is the result of k_r

The labels also contain a parameter to mark the labeler’s confidence in the label (using the binary decision “good”/“bad”). The labeling was carried out with the help of a graphical tool. The labels were stored in XML-format and checked for consistency afterwards.

4. Register rules

A rule set was established that transforms the discourse structure of the labeled data into the appropriate register following the model described in section 2.

The rules use the notion of *sub-topic*, *complex sub-topic* and *topic reset*. A *sub-topic* spans discourse segments that are not dominated by segments that lie outside of the sub-topic. A *complex sub-topic* is a sub-topic that embeds other sub-topics. This is the case when segments are dominated by other segments of the same sub-topic. A *topic reset* takes place when the subsequent discourse segment does not belong to the same sub-topic.

In the example from Figure 1 we can see that the discourse segments 2 and 3 introduce a complex sub-topic. Segment 5 is a sub-topic on its own (but not complex). Between segment 5 and 6 a topic reset takes place, because segment 6 is again dominated by segment 1. This is because segment 2 is dominated

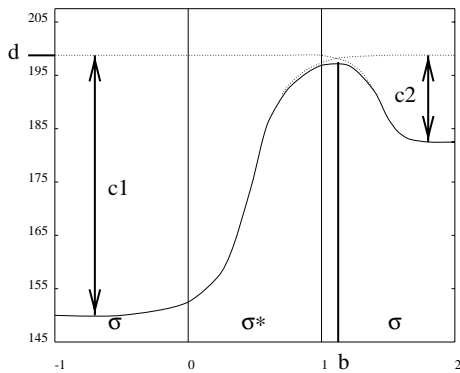


Figure 3: The parameters of the PaIntE parametrization.

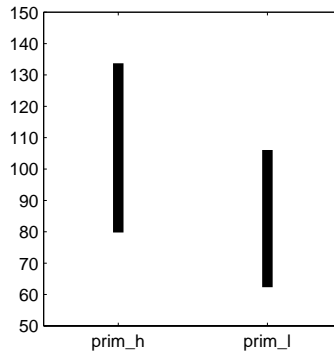


Figure 4: Pitch range of the primary registers $\{h\}$ and $\{l\}$ for the male speaker.

by segment 1 and the continuation relation links segment 2 and 6 on the same level of embedding.

With these terms we define the rules that determine a primary and a secondary register for each intonation phrase (IP).

1. Supply a *primary* $\{h\}$ register if this is the only IP within the discourse segment.
2. For discourse segments containing more than one IP, supply a *primary* $\{l\}$ register if this IP is the final IP within a discourse segment, and a *primary* $\{h\}$ register otherwise.
3. Supply a *secondary* $\{l\}$ register if a topic reset takes place after the current discourse segment (for all IPs of the discourse segment).
4. Supply a *secondary* $\{l\}$ register if the IP belongs to the final discourse segment of the news story.
5. Supply a *secondary* $\{h\}$ register if the IP is the initial IP in the discourse segment *and* the appropriate discourse segment introduces a complex sub-topic.
6. Supply a *secondary* $\{h\}$ register if the IP is the initial IP in the discourse segment *and* a topic reset took place just before the current discourse element.
7. Supply a *secondary* $\{h\}$ register if the IP is the initial IP of the news story.

It follows from these rules that the beginning of a sub-topic is only marked if it is a complex sub-topic, if a topic reset has occurred before, or if it is actually the main topic (first discourse segment in the news message). The beginning of simple sub-topics that are not complex are not treated especially.

5. Pitch-Range Determination

The pitch-range of every intonation phrase in the database is automatically extracted using the PaIntE (*Parametric Intonation Event*) parametrization described in [9].

Six parameters describe the basic movement of the fundamental frequency around a particular pitch accent (cf. Figure 3). The parameters allow a phonetic interpretation of the pitch event. For the pitch range determination we use the parameter d that describes the frequency of the peak, as well as c_1 and c_2 expressing the amplitude of the rising and falling part of the curve. The alignment parameter b and the two steepness parameters a_1 and a_2 for the rising and falling slope, respectively, are of no interest to this study.

We define the pitch range following the principles of the tone-sequence model of intonation [10]. In this model an intonation contour is described as a sequence of high (H) and low (L) targets. Our definition of the pitch range is the range between the F_0 value of the lowest L-target and the F_0 value of the highest H-target within an intonation phrase. Thus, to determine the upper edge of the pitch range we take the highest d parameter found within the intonation phrase. The same method leads to the lower margin of the pitch range taking into account the lowest $d - \max(c_1; c_2)$ of the intonation phrase. We excluded outliers from the analysis to achieve a more robust result. Figure 2 shows the intonation phrase of an utterance from the news corpus with the pitch range marked by horizontal lines.

6. Results

In this section we investigate the pitch-range values found for the registers of the male speaker in the database. Since in general the samples are not normally distributed we use median values to represent the pitch range of a particular register.

6.1. Pitch range of primary registers alone

As described above primary registers represent the position within a discourse segment: Final intonation phrases receive a primary $\{l\}$ register and all other phrases are marked with a primary $\{h\}$ register. Due to the phonetic phenomenon of *final lowering* we would expect a lower pitch range for phrases with a primary $\{l\}$ register. This is what we found in our database analysis (cf. Figure 4). The primary $\{l\}$ register extends from 62 Hz to 106 Hz, the primary $\{h\}$ register from 80 Hz to 134 Hz. The phonetic realization of the two registers overlap by approximately half of the registers' range.

6.2. Pitch range of registers with primary $\{h\}$

Many of the intonation phrases only carry a primary $\{h\}$ register, but are not further modified by a secondary register. They usually occur in positions that are not directly influenced by discourse relations. For the male speaker in the database the pitch range of these *high registers* extends from 80 Hz up to 132 Hz (cf. Figure 5).

The $\{h-l\}$ register appears in non-final position within the last discourse segment before a topic reset. The pitch range of these registers is slightly (but not significantly) lower than the one of the basic $\{h-0\}$ registers extending from 77 Hz to 127 Hz. We call it a *lowered high register*.

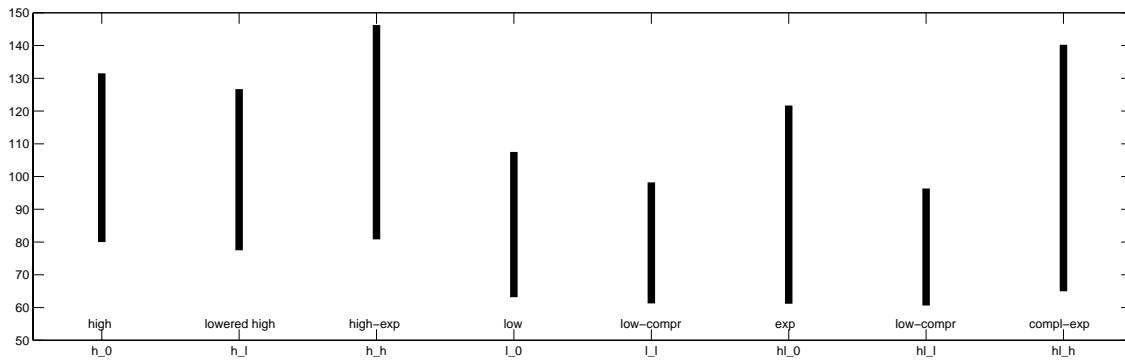


Figure 5: Pitch range of the registers for the male speaker.

The beginning of a new complex sub-topic is marked by an {h-h} register and so is the first IP after a topic reset. The pitch range is clearly expanded to the top (from 81 Hz to 146 Hz) showing the properties of a *high expanded register*. This finding is in accordance with our earlier study on registers [6]. The expansion of the pitch range can be interpreted as a stylistic means to guide the attention to a new sub-topic.

6.3. Pitch range of registers with primary {1}

Registers with a primary {1} but with no further modification occur at the end of discourse segments not preceding a topic reset. We call them *low registers*. For our male speaker they range from 63 Hz to 108 Hz.

In IPs that precede a topic-reset the primary {1} register is further modified with a secondary {1} register. Compared to the low registers the pitch range is compressed. According to these findings the {1-1} registers are called *low-compressed registers*. They range from 61 Hz to 98 Hz. These results confirm our earlier findings described in [6].

Whereas there is a strong influence of the topic reset on the final IP (low compressed register) we could not find a significant effect on earlier IPs (see the pitch range of the lowered high register above). We conclude that the topic reset mainly affect the two IPs that surround it. This may call for a refinement of the register rules. The rules No. 3 and 4 should possibly be restricted to segment-final IPs only.

6.4. Pitch range of registers with primary {hl}

Discourse segments containing only one IP are marked with both primary {1} and primary {h} registers noted as primary {hl}. We will see how the two distinct phonological features affect the pitch range.

With no secondary modification both primary registers influence the pitch range, resulting in an *expanded register* that ranges from 61 Hz to 122 Hz for our male speaker.

The realization of the {hl-1} register is similar to the {1-1}, which is why we call it *low compressed register* as well. The influence of the primary {h} seems to be lost for this register. Its pitch range spans from 61 Hz to 96 Hz.

For the {hl-h} register both the primary {1} and primary {h} registers affect the pitch range. With a pitch range ranging from 65 Hz up to 140 Hz the register occupies almost the complete overall range of the speaker. We may consider this *complete expanded register* as a fusion between the {h-h} and the {1-0} registers, which are the two registers that would occur if the discourse segment consisted of more than one IP.

With the obvious exception of the {hl-1} registers, we may conclude that the realization of single-IP discourse segments can be seen as a merge between the properties of primary {1} and primary {h} registers. The upper margins of these registers are slightly lower than we would expect from the appropriate multi-IP registers. A simple overlay of the pitch ranges would possibly lead to registers that are too expressive, which explains the slight compression.

7. Conclusion

The interface between phonological registers and their phonetic realization, the pitch range, has been established by the study described in this paper. It can be implemented in a speech synthesis system in a straightforward manner if discourse information is available on the input side.

In a concept-to-speech system such information is in principle available from a language generation engine. It is, therefore, the ideal application to successfully implement a global prominence model. The method would be to first apply the rule set described in section 4 to the input discourse. Primary register features reflect the prosodic phrasing within a discourse segment. The secondary register features are a function of the relations between the discourse segments, more specifically the topic structure of the discourse. In a second step every register is mapped to the according pitch range derived from Figure 5.

The rule set described in section 4 can be seen as a direct

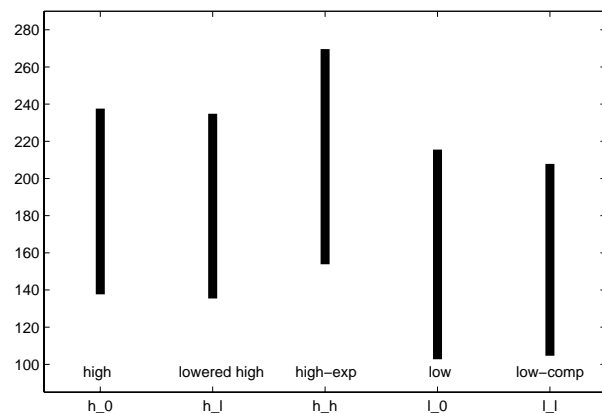


Figure 6: Pitch range of the registers for the female speaker.

implementation of the global prominence model in [5]. Further prosodic investigations of the database but also taking into account other discourse structural information than dominance will result in a refinements of the rules — and may finally lead to register classes with an even stronger phonetic coherence.

A question not addressed so far is how much the results demonstrated here depend on the one male speaker only. We have carried out a preliminary analysis of the pitch ranges of a female speaker in the database. This result is shown in Figure 6. Only the pitch range for the primary {l} and primary {h} is displayed. The analysis of primary {hl} was not reliable due to a low frequency of occurrence. The registers included in this analysis show a very similar behavior as for the male speaker, suggesting that our findings are speaker independent.

We should like to add that pitch range is not the only phonetic correlate of discourse structure. The duration of pauses after intonation phrases is also an important cue to mark the relation between discourse segments [5] and should, therefore, taken into account when developing a discourse model for speech synthesis.

8. References

- [1] B. Grosz and C. L. Sidner, “Some intonational characteristics of discourse structure,” in *Proceedings of International Conference on Spoken Language Processing*, Banff, Canada, 1992, pp. 429–432.
- [2] G. M. Ayers, “Discourse functions of pitch range in spontaneous and read speech,” *Ohio State University Working Papers in Linguistics*, vol. 44, pp. 1–49, 1994.
- [3] N. Asher, *Reference to Abstract Objects in Discourse*, Kluwer Academic Publishers, Dordrecht, 1993.
- [4] William C. Mann and Sandra A. Thompson, “Rhetorical structure theory: a theory of text organization,” ISI Reprint Series, 87-190, June 1997.
- [5] Jörg Mayer, “Prosodische Merkmale von Diskursrelationen,” *Linguistische Berichte*, vol. 177, pp. 65–86, 1999.
- [6] Gregor Möhler and Jörg Mayer, “A method for the analysis of prosodic registers,” in *Proceedings of Eurospeech*, Budapest, 1999.
- [7] Stefan Rapp, *Automatisierte Erstellung von Korpora für die Prosodieforschung*, Ph.D. thesis, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung, 1998.
- [8] Jörg Mayer, *Intonation und Bedeutung. Aspekte der Prosodie-Semantik-Schnittstelle im Deutschen*, Ph.D. thesis, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung, 1997.
- [9] Gregor Möhler and Alistair Conkie, “Parametric modeling of intonation using vector quantization,” in *Proceedings of 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.
- [10] Janet B. Pierrehumbert, *The Phonology and Phonetics of English Intonation*, Ph.D. thesis, PhD Thesis, MIT, Cambridge, MA, 1980.