Linking bottom-up intonation stylization to discourse structure

Uwe D. Reichel

Institute of Phonetics and Speech Processing
University of Munich
reichelu@phonetik.uni-muenchen.de

Abstract

A new approach for intonation stylization is introduced, that enables the extraction of an intonation representation from prosodically unlabeled data. This approach yields global and local intonation contour classes arising from a contour-based, parametric and superpositional intonation stylization. Based on findings about the linguistic interpretation of the contour classes derived from corpus statistics and perception experiments, we created simple prediction models for the partial generation of intonation contours from discourse structure defined by discourse segment boundaries and the information status of nouns within these segments. The predicted intonation contours were evaluated by human judgments of adequacy that yielded a high accordance.

Keywords: computational intonation stylization, data-driven, contour-based, superposition, discourse structure

1. Introduction

1.1. Dichotomies of intonation models

Established intonation models can be classified along three dimensions:

1. the chosen units: *tone targets* vs. *contours*,
2. their description: *symbolic* vs. *parametric*, and
3. their arrangement: *single-layered* vs. *superpositional*.

In the following, a selection of high impact intonation models will be described with respect to these dichotomies. The tone sequence model (TSM) of [1, 2, 3] can be characterized as target-based, symbolic and single-layered, since fundamental frequency (F0) within an intonation phrase is described as a single-layered sequence of tone symbols that are assigned to pitch accented and phrase boundary syllables. The PENTA model, conceptually introduced in [4] and quantified in the form of the qTA model by [5], also considers the F0 contour as a sequence of targets that are static (horizontal) or dynamic (rising or falling) and are assigned to each syllable. While in the TSM the F0 contour generally results from connecting the targets by an interpolation function, PENTA considers the F0 contour as a result of target approximation which can be realized in different forms such as a third-order critically-damped linear system as in [5]. The PENTA model thus is target-based, parametric, and arranges intonation units in a single layer. The Fujisaki model [6, 7, 8], in every respect opposite to the TSM, is contour-based, parametric, and superpositional. It considers F0 as a superposition of a global phrase component related to

declination and a local accent component related to F0 movements on accented and phrase-final syllables. The components are parametrically represented as critically damped systems activated by phrase and accent commands, respectively. The TILT model [9] as well as the PaintE model [10] yield a contour-based, parametric and single-layered intonation representation. They provide a parameterization of the F0 contour in the scope of accented and phrase-final syllables.

1.2. Model requirements

Ideally, intonation models should allow for:

1. an appropriate abstraction from the signal,
2. the interpretability of the abstraction, and
3. its automation.

In the following sections the models presented thus far are discussed with respect to these requirements.

1.2.1. Signal abstraction

An appropriate abstraction from the signal means to capture relevant aspects of the signal, to allow for signal reproducibility, and to be reproducible itself if repeated on the same signal.

Relevant signal aspects. Tone-based approaches yield a comparably high degree of abstraction but hence have to face the criticism to neglect relevant properties of the F0 contour between tone targets due to underspecification. For example, they cannot account for the shape between the prenuclear and nuclear tone accents in Neapolitan being concave in questions and linear in statements [11]. So far two major proposals to face this issue had been made: the insertion of intervening targets [11] and the usage of interpolation rules [12]. Since tones are associated to accented and phrase boundary syllables only, additional targets require a reorganisation of the prosodic structure which needs justification. This reorganisation can for example consist in the insertion of an additional prosodic level of accentual phrases as in [11]. The usage of interpolation rules raises the question why not to work directly with contours instead of tones.

Dainora [13] argues for a contour approach due to the high predictability of tones given their predecessors indicating that the relevant intonation unit is not the tone but a tone sequence and thus a contour. Further support for this perspective is given by [14] who found for German dialog data that contours are more appropriate than tones to reveal the underlying intonation systematics for turn holding and turn yielding.

Signal reproducibility. Since parametric models are inherently more closely related to the signal, they are more appropriate for signal reproduction from the abstract intonation representation. While a parametric representation can be transformed directly into the F0 contour by parameter assignment as in an analysis-by-synthesis framework, symbolic approaches depend on additional mediation which may be hand-crafted rule-based [15] or data-driven [16].

Abstraction reproducibility. Symbolic approaches relying on manual labeling have to face the risk of low intra- and inter-labeler agreement. A common proposal to address this issue of low abstraction reproducibility is to reduce the size of the label inventory [17, 18].

Established approaches do not guarantee the reproducibility of the abstraction either, since the parameters for the chosen stylization functions are not derivable analytically but have to be estimated numerically by local optimization of the fit between representation and original contour

[8, 19]. Thus the abstraction depends strongly on the parameter value initialization. Furthermore, the relation between the parameter values and the F0 contour is usually not biunique, i.e. different parameter assignments can lead to the same contour (see [20] p. 60 for an example).

1.2.2. Interpretability

Linguistic interpretability. Empirical findings suggest that linguistic phenomena show categorical as well as gradual correspondence in the production and perception of intonation. For example, [21] found that the realizations of broad vs. narrow focus show categorical differences in terms of different pitch accent types as well as gradual differences in F0 alignment and segment durations. Therefore, both symbolic and parametric intonation representations are principally accessible for linguistic interpretation, which is discussed in depth by [22]. [23] can be referred to for the linguistic interpretation of symbolic approaches, and [7, 8] for the interpretation of parametric models.

To make use of possible advantages of more intuitive linguistic interpretations on the symbol level, approaches transforming parametrical transcriptions into symbolic ones might be helpful. The PaintE model for example offers an additional symbolic F0 representation by means of parameter vector clustering [24].

Phonetic interpretability. A superpositional arrangement of intonation units as in the Fujisaki model can account for long-term phenomena like declination. Furthermore, it offers the possibility to incorporate findings of pre-planning in intonation production [25], amongst others reflected by the relation between utterance length and declination slope.

1.2.3. Automation

Automation is highly desirable if a model is to be tested on a larger amount of data. Furthermore, it allows for fast applicability to new data of other languages. In contrast, manual adaptations of symbolic label inventories like ToBI [26] for other languages [27] are very laborious.

As said above in the context of signal reproducibility, the advantage of a parametric unit description is its direct linking to the signal. While symbolic representations need experts or additional modules to derive symbols from the signal [28] or to generate F0 contours [16], parameters can be directly inferred from and transformed into F0 values.

1.3. Discourse structure

[29] proposed three parallel discourse structures: The *linguistic structure* is given by the written or spoken text, the *attentional structure* represents the relative salience of discourse entities, and the *intentional structure* subdivides the discourse into segments of coherent speaker intentions. The linguistic concepts into which we attempt to embed our intonation stylization are discourse segmentation and the information status of words within these segments. These concepts can be linked to the intentional and the attentional structure, respectively.

1.3.1. Discourse segmentation

Discourse segmentation serves to group together coherent parts of an utterance. One way to address coherence consists in a linear discourse segmentation into subtopic units [30]. Concerning the intonational marking of discourse segmentation it has been found in production and perception studies for several languages that subtopics generally start with a higher F0 register

and end in a lower register [31, 32, 33], so that pitch reset is more pronounced at topic shifts [31]. The strength of coherence of adjacent discourse segments is prosodically encoded amongst others by boundary tones [23, 34], and the degree of final lowering [35]. High boundary tones generally mark continuation and thus a high degree of cohesion of adjacent discourse segments, while low boundary tones and final lowering indicate low coherence. A more extensive overview on intonation in discourse segmentation is given by [36].

1.3.2. Information status

The information status of a discourse entity describes whether it contains new information, which is not yet present in the discourse context, or given information, that is already available. [37] proposes a major categorical division of *givenness* into *evoked* (already mentioned or directly available as part of the dialog situation) and *inferable* (only accessible via an evoked entity). [38] distinguishes between different activation levels of discourse entities on an ordinal scale: *given* (active), *accessible* (semi-active), and *new* (inactive). A more detailed six-level scale is given by [39].

Numerous studies mainly within the TSM framework have revealed how information status is expressed by means of intonation. Different pitch accent type preferences (including deaccentuation) have been identified for given and new information for several languages as English [40, 23] and German [34, 41]. [35] and [42] relate these findings to the *attentional structure* of [29] by expressing *given* and *new* in terms of different degrees of *salience*. The intonational marking of different degrees of givenness is explored e.g. by [23, 43] who found amongst others degree-dependent pitch accent type preferences.

The aim of this study is to link our approach to simplified forms of the introduced discourse concepts. The approach was originally developed and presented in [20, 44] (in the first source it is referred to as the PKS stylization) and is based on the considerations formulated in Section 1.2. It will be described in Section 2. In Section 3 our previous work on its linguistic interpretation with respect to discourse is summarized. Based on these findings in the current study simple prediction models for global and local contour classes were handcrafted, which are introduced in Section 4 together with their perceptual evaluation.

The results are discussed in Section 5 with a focus on the relevance of our approach concerning adequate intonation representation for linguistic interpretation.

2. The CoPaSul intonation approach

2.1. General description

Our CoPaSul approach provides a contour-based (*Co*), parametric (*Pa*), and superpositional (*Sul*) F0 representation. F0 contours are treated as a superposition of global and local components. These components are anchored in a hierarchical prosodic structure defined by global and local segments which roughly correspond to intonation phrases and potential accent groups, respectively, where 'potential' means that the allowed number of accented syllables within such a group is zero or one. The stylization of the F0 contours is carried out as follows: Within each global segment a linear F0 base contour is fitted. After the subtraction of this global baseline within each local segment a third order polynomial is fitted to the F0 residual. Subsequently, a symbolic description of the intonation inventory in the form of global and local contour classes is derived by polynomial coefficient clustering. On the phonetic level, linear regression models adjust these abstract units to the respective prosodic context.

CoPaSul thus stands in the tradition of parametric (Fujisaki, PaintE, PENTA, Tilt) and superpositional (Fujisaki) models. Like the Fujisaki model it explicitly distinguishes between a global and a local intonation component and thus allows for addressing them separately. Its parametric and contour-based nature closely connects the stylization to the signal level. Furthermore, parameter clustering has been adopted from the PaintE approach yielding a symbolic intonation representation that allows for linking the stylization to the linguistic level.

2.2. Data and preprocessing

The training data originates from the SI1000P corpus [45] containing 190 minutes of German read speech by a professional standard German male newsreader. F0 contours were extracted by the Schaefer-Vincent algorithm [46] and transformed to semitones (base 50 Hz). F0 errors and voiceless segments were bridged by a shape-preserving piecewise cubic Hermite spline interpolation [47]. The contours were smoothed by a Savitzky-Golay filter [48] of order 3 and window length 5. A main advantage of this filter type over other standard smoothing methods like moving average is that it is only a little prone to temporal shifts of local peaks and valleys and therefore better preserves the F0 envelope.

Pauses and syllable nuclei were detected automatically by energy (root mean squared deviation; RMS) comparison between an analysis window w_a and a longer reference window w_r with the same time midpoint. For pause assignment the energy in w_a had to be considerably lower than in w_r, more precisely $RMS(w_a) < RMS(w_r) \cdot 0.06$. The length of w_r was set to 5 s, the length of w_a to 150 ms in order to prevent the confusion of pauses and shorter stop consonant closures. For syllable nucleus assignment the energy in the relevant frequency range from 245 to 3125 Hz had to be considerably higher in w_a than in w_r, and additionally had to surpass a threshold which was defined relative to the maximum energy of the utterance, i.e. $RMS(w_a) > 0.15 \cdot \max(RMS) \wedge RMS(w_a) > RMS(w_r) \cdot 1.2$. Here, the lengths of w_a and w_r were set to 50 and 250 ms, respectively.

All length, factor, bandwidth, and threshold parameters were estimated by the non-linear Nelder-Mead Simplex optimization [49] on a SI1000P subcorpus comprising 20 hand-segmented sentences with 1011 syllables and 86 speech pauses. The error for pause detection in terms of insertions and deletions in non-final position amounts 10%, the error for syllable nucleus detection 7%. These error rates are regarded as not too severe: First, in our read speech data every punctuation mark co-occurred with a speech pause, so that a global segment boundary was set even if pause detection failed. Second, almost all nucleus detection errors occurred in low prominence function words which are expected to be of minor relevance for determining intonation events. Details of the optimization and the evaluation are described in [20].

On the text level, part of speech tagging was carried out by a tagger described in [50]. Signal and text were aligned by the Munich Automatic Segmentation System (MAUS) [45], and a grapheme-phoneme converter [51] served to locate the word-stressed syllables within this alignment.

2.3. F0 analysis

2.3.1. Prosodic structure

The segmentation into global and local segments was carried out automatically based on the preceding alignment of the signal and the tagged text and on pause detection. Global intonation segments are delimited by speech pauses and by punctuation. Local intonation segments were defined as a chunk of function words terminated by a content word or a global segment boundary.

This notion roughly corresponds to chunking approaches as those given in [52, 53] and ensures in most cases that each local segment contains one accented syllable as a maximum. As an example, the utterance illustrated in Figure 1 *Die Tiere verstummen, ein Unheil naht. (The animals hush, a disaster is approaching.)* is divided into global segments at punctuation marks, and each global segment is further divided into local segments by placing a boundary behind each content word yielding the following structure: [[*Die Tiere*] [*verstummen*]], [[*ein Unheil*] [*naht*]] *(*[[*The animals*] [*hush*]], [[*a disaster*] [*is approaching*]]*)*.

2.3.2. F0 stylization

All stylizations are based on the F0 values in 110 ms frames centered on the detected syllable nuclei. The advantage of this approach is that its demands on preprocessing are low. It requires only a syllable nucleus detection, which is robust and can be carried out automatically. There is no need for an exact syllable segmentation or for a weighting of more and less important parts of the F0 contour. Thus, ad-hoc approaches like intensity-based weighting [54] are dispensable. The choice of a frame length of 110 ms is a trade-off between the need of enough input data for reliable polynomial fitting and the need to avoid frame overlaps of consecutive syllables. If an overlap still occurs, neighboring frames are shortened by an equal amount.

Global and local contour stylizations that will be described in the next paragraphs are shown in Figures 2 and 3.

Global contours. Within each global intonation segment, a declination baseline is derived as follows: For each syllable nucleus window the F0 minimum is taken as an F0 base level. The baseline then is adjusted as the flattest bottom tangent of the sequence of these base level values that passes through two points of the sequence chosen from all linear connections of pairs of F0 minima [20]. In order to make baseline slopes comparable across different segment lengths, time is normalized to the interval from 0 to 1. This baseline is then subtracted from the F0 contours, and its slope is recorded for subsequent clustering (see Section 2.3.3).

Local contours. Within each local segment a third-order polynomial is fitted to the time-normalized residuum contour. As illustrated in Figure 3 time is normalized as follows: The time span of the local segment is set from -1 to 1. -1 is assigned to the left boundary of the first syllable nucleus window, and 1 to the right boundary of the last syllable nucleus window. 0 is placed on the nucleus of the syllable of the segment-final word (the content word), that carries the word stress. Thus, the peak of the F0 contour can be interpreted relative to the accent position. This approach requires separate normalizations of the pre- and post-accent parts of the local segment. By means of normalization it is possible to compare utterance segments of different lengths. Furthermore, it allows to ignore unstable polynomial behavior outside the chosen interval it is fitted to. On the other hand, this normalization obscures the influence of phonetic segment durations and syllable number on peak alignment (see [55] for an overview). However, due to the high number of different phonetic segmental contexts, this influence is considered as noise that overall does not affect the analysis in a systematic way.

The selected polynomial order is motivated by the trade-off between capturing relevant F0 movements and avoiding data overfitting. First and second order polynomials are not powerful enough to cover all relevant aspects of local F0 contours. Polynomials of fourth or higher order run the risk not to be well-conditioned and to complicate subsequent clustering, since they demand a larger amount of data due to the increased coefficient vector length, as well as a weighting schema for more and less relevant coefficients. In any case, the danger of oversimplification by

third-order polynomials is diminished by the specification that a local segment contains at most one pitch accent, which limits the expected complexity of the F0 contour.

2.3.3. Contour classes

Contour classes were derived by k-means clustering of the range-normalized coefficients. For global classes the baseline slope values and for local classes the polynomial coefficient vectors were clustered with respect to their squared Euclidean distances. Cluster initialization was carried out by subtractive clustering [56] that iteratively locates initial centers in the parameter space at regions with high data density. The parameters for subtractive clustering were derived by Nelder-Mead Simplex optimization. Details on the application of this procedure are given in Appendix A.

Figure 4 shows the centroids of the resulting three global and five local contour classes g_1 to g_3 and c_1 to c_5, respectively. The centroid coefficient values are listed in Appendix A. In Figure 1 intonation within the two global segments is illustrated as the superposition of the global contour class sequence g_2, g_2 and the local class sequence c_2, c_4, c_5, c_1.

The chosen squared Euclidean distance measure is not primarily perceptually motivated, since reliable perceptually grounded distance measures are not yet available (although attempts were made to develop such metrics, e.g. [54, 57]), but it is in line with the general bottom-up character of the current approach. In [57] the Euclidean distance turned out to correspond slightly better to perceptual distance judgments than other metrics.

The cluster center initialization by means of subtractive clustering guarantees stable results across disjunct data subsets and input vector randomization. The number of clusters was constant over all subsets and randomizations. All pairwise correlations of corresponding centroids are above 0.97 for the data subsets and above 0.99 for data randomization.

2.3.4. Phonetic realization

The phonetic realization of the contour classes on the basis of linear regression models is illustrated in Figure 5. The regression models serve to map the abstract contour class centroids to the intonation surface level.

Contour realizations. In order to constrain the deviance of a contour realization from the underlying class, the regression models for global and local contour realizations predict absolute values for the polynomial coefficients. The algebraic sign is taken over from the underlying centroid coefficient.

The linear regression model for the slope realizations $|b_r|$ of global contours is given by $|b_r| = w_0 + w_1 \cdot |b_c| + w_2 \cdot |b_{rp}| + w_3 \cdot l$. This means that the absolute slope realization $|b_r|$ of a global contour is derived from the underlying absolute centroid slope $|b_c|$ of the contour class, the realized slope of the preceding contour $|b_{rp}|$, and the length of the current global segment l. w_i are the predictor weights, which were calculated by minimizing the overall error between $|b_r|$ and absolute values of the observed baseline slopes obtained from the stylization. For weight calculation the realized slope of the preceding contour b_{rp} was set to the value of the respective stylization coefficient. In application contour slopes are estimated sequentially from left to right, so that b_{rp} is given by the output of the regression model for the preceding global contour.

For each local contour coefficient a separate linear regression model was trained: $|a_r| = w_0 + w_1 \cdot |a_c| + w_2 \cdot |a_{rp}| + w_3 \cdot |b_r| + w_4 \cdot p$. The absolute value of the contour coefficient realization $|a_r|$ is predicted from its underlying contour class coefficient $|a_c|$, the realized value of the preceding corresponding coefficient $|a_{rp}|$ the realized slope of the current global contour $|b_r|$,

and the relative position p of the local segment within the global segment. The predictor weights are calculated by minimizing the overall error between $|a_r|$ and the corresponding value obtained from the stylization.

Pitch reset. At junctions between global segments the pitch reset r is modeled by $r = w_0 + w_1 \cdot b_{r1} + w_2 \cdot b_{r2} + w_3 \cdot pl$. b_{r1} and b_{r2} are the realized slopes of the adjacent global contours. Their absolute values have been predicted by the global contour model introduced above. pl is the length of the interjacent pause. The predicted pitch reset is added to the final F0 value of the preceding global contour in order to derive the F0 starting level of the current global contour. The predictor weights are calculated by minimizing the overall error between r and the corresponding observed pitch reset values.

For all linear regression models, the predictors are range-normalized. The resulting regression weights are shown in Table A.4 in Appendix A along with the correlations between predictions and targets that range between 0.63 and 0.82.

Gain from the realization models. The regression models serve to set the concrete realization of a contour class in the context of extrinsic influence factors. By this it can be accounted for the negative correlation of global segment length (predictor l) and global contour slope [25]. As can be seen in Appendix A, Table A.4, this reverse relation is captured by the negative contour length weight $w_3 = -0.0221$.

An implicit F0 topline [58] can be modeled by relating the local contour coefficients to the relative position of the contour (predictor p) within the global segment. Accordingly, the corresponding regression weights w_4 turned out to be negative for all local contour coefficients. Thus, the more a local contour is located to the end of a global segment, the lower its height, and the less pronounced its steepness and shape. Moreover, the regression models serve to smooth contour sequences, since contour parameters are calculated not only from the underlying centroids but also from the correspondent coefficients of the preceding contour.

The pitch reset model accounts for the co-operation of pitch reset [59] and pause length [60] in encoding prosodic boundary strength. This positive impact of pause length on the pitch reset is reflected by the positive value of $w_3 = 0.2166$ shown in Table A.4. In principle, the pitch reset model could be extended by a predictor representing the presence or absence of a discourse segment boundary in order to incorporate the findings of [31] (cf. section 1.3.1). However, this is not yet possible on the basis of the given training data since each sentence was produced in isolation so that pitch reset could not been measured across topic shifts.

Local segment junctions. Depending on the distance between the two syllable nuclei adjacent to a segment boundary a contour gap (as in Figure 5) or overlap occurs. Both are bridged by linear interpolation. Start and endpoint of this linear bridge are smoothed by a moving median filter.

3. Previous work on automatic linguistic anchoring of the CoPaSul stylization

The text-based local contour class prediction suggested in the current study (see Section 4) integrates the findings of two previous perception experiments [44, 61] that will be reviewed in the following.

3.1. Linguistic concepts

We had tested the linguistic adequacy of the local contour classes with respect to semantic weight [44] (defined in terms of word predictability [62]), utterance finality [44], and information status [61]. Due to strong correlations between the contour classes' relations to semantic weight and information status and the inherent correlations of these two concepts [20], only the two discourse structure concepts *discourse segmentation* and *information status* are further considered in this study. For these initial attempts to relate the data-driven approach to discourse structure, terminology was strongly simplified: First, we reduced information status to the dichotomy *given vs. new information* not further distinguishing between different levels of givenness. Second, we considered discourse segments to be subtopics in a linear sequential order as in [30] leaving aside hierarchical discourse structures or more sophisticated discourse coherence analyses reflecting its intentional structure. Third, we set equal utterance and discourse segment boundaries based on the assumption that the intonation of declarative utterance ends is accepted as a marker of potential discourse segment ends, due to their common coincidence.

3.2. General Method

The general procedure to link the intonation stylization to linguistic units can be described as follows: First, linguistic concepts were extracted by means of simple sentence segmentation and natural language processing (NLP) methods. Then, based on co-occurrence statistics between the linguistic events and the contour classes, hypotheses were formulated about the linguistic function of the classes. For global classes these hypotheses concern discourse segmentation, for local classes discourse segmentation and information status are covered. For the local classes these hypotheses were subsequently tested by perception experiments. 24 German mother tongue subjects between 22 and 47 years took part in these experiments. All subjects were students or scientists of phonetics in Munich.

3.3. Discourse Segmentation

From the studies referred to in section 1.3.1 it can be concluded that in intonation segment finality is encoded on a local level by means of boundary tones [23, 34] and on a global level by means of F0 register [31, 32, 33]. Segment starts are predominantly marked on the global level in terms of F0 register [31, 32]. Therefore, we examined the function of local contour classes in segment final position only, with the aim to classify them by the dichotomy *final vs. non-final (continuation)*. The function of global contour classes was examined in segment final and additionally in initial position using the two dichotomies *final vs. non-final* and *initial vs. non-initial*, respectively. *Non-final* and *non-initial* both subsume medial positions.

Since in our data the reader produced each sentence in isolation, the last global, respectively local segment of each sentence was classified as discourse segment final, and the others as non-final. Accordingly, the first global segment of a sentence was labeled as discourse segment initial, all others as non-initial.

3.3.1. Global contour classes

For each class it was tested separately (a) whether there is a significant relation to initial position (i.e. initial vs. non-initial the latter subsuming the medial position) and (b) whether there is a significant relation to final position (i.e. final vs. non-final the latter subsuming medial). In case a significant relation between a contour class and a position was revealed, the direction of this relation was determined by comparing the observed occurrence count of the class in the

class	segmentation
g_1	non-initial, final
g_2	initial, final
g_3	non-initial, non-final

Table 1: Occurrence preferences of global contour classes within discourse segments.

respective position with the count to be expected in the case of independence. By this comparison a positional preference (e.g. initial) or obstruction (non-initial) was concluded. For both positions each class showed significant preferences or obstruction (χ^2 values between 28.43 and 257.18 were obtained; $\alpha = 0.001$). The relations are summarized in Table 1. One can see from this table, that in our data the global contour marking of discourse segment initial and final parts is disjunct only for class g_1, that preferably occurs in final position but is only reluctantly used in initial position. Class g_2 in contrast has a preference for both positions from which can be inferred that it is rarely to be observed in medial position. Both classes are characterized by a negative slope to establish low F0 register values at the end of discourse segments. g_3 in contrast has a preference to occur neither in initial nor in final position from which can be inferred its preference for the medial position to mark continuation by its positive slope.

In line with these findings χ^2 tests relating contour classes and medial position showed a significant preference for this position only for class g_3 while g_1 and g_2 significantly more often occur in non-medial (i.e. initial or final) position (χ^2 values from 24.91 to 124.38, $\alpha = 0.001$).

Unfortunately, the impact of pitch reset on discourse segmentation allowing for high topic initial F0 register [31] could not be addressed directly, since in the training data each sentence was produced in isolation, so that pitch reset values were not available for discourse segment boundaries.

3.3.2. Local contour classes

Testing the significance of co-occurrences of contour classes and utterance finality by χ^2 tests ($\alpha = 0.05$; [63]) resulted in the hypotheses: Finality is encoded by c_1 and c_5, continuation by c_2, c_3, and c_4, which is summarized in Table 2.

class	segmentation	information status
c_1	final	given
c_2	non-final	new
c_3	non-final	new
c_4	non-final	given
c_5	final	new

Table 2: Relations between contour classes as discourse segmentation (finality vs. non-finality) and information status markers (new vs. given information). All relations are significant (χ^2 tests, $\alpha = 0.05$).

In a perceptual validation of these hypotheses single local segment utterances of the form *'Eine X (an X).'* were presented. The stimuli had been synthesized by means of MBROLA [64] using a male German voice database (*de4*). Segment durations had been calculated by a regression tree which will be described in Section 4.2. For all F0 contours the F0 baseline was constantly set to 80 Hz (declination slope 0), so that the contour variation was determined solely by the local contour classes. The class-related contours were laid on the time-normalized

utterances as illustrated in Figure 3. Time 0 was associated with the nucleus midpoint of the stressed syllable in the target word *X*.

The 60 target words *X* were amongst others controlled for uniform syllable number and structure, word frequency, and voicing, in order to rule out that word-intrinsic properties interfere with its prosodic realization. The task was to allocate the stimuli on a 5-point bipolar scale with the end points *'Eine X und eine Y (an X and a Y)'* and *'Eine X'*. Allocating a stimulus to *'Eine X'* implies that it is considered as a completed utterance and is thus characterized by segment-final intonation, whereas allocating it to *'Eine X und eine Y'* implies that subjects expect more to come triggered by an intonation contour signaling continuation.

The subjects' judgments for each class are shown in Figure 6 in the right boxplots together with the corpus-derived hypotheses marked by filled circles placed at 1 (non-final) or 5 (final). Except for class c_5 all hypotheses were perceptually verified (Kruskal-Wallis test, $\chi^2_4 = 316.9, p < 0.001$ [65]; Dunnett post-hoc test, $\alpha = 0.01$; [66]). Furthermore, all classes were perceptually identified with respect to finality, since all judgments differed significantly from the mean level 3 of undecidedness (sign tests, $\alpha = 0.05$, Bonferroni-corrected, $|z| > 3.4, p < 0.001$; [67, 68]).

3.4. Information status

As described in [61] we automatically labeled nouns co-referring to preceding ones with the information status *given*, and the others with *new*. To achieve this, first, the text was segmented into thematic units using an adapted version of the TextTiling algorithm [30]. Subsequently, coreference resolution was carried out within each of these units by means of an iterative pattern matching procedure proposed by [69], extended by a transitive closure based on compound analyses.

Considering the stylization parameters, only the offset polynomial coefficient differed significantly across the conditions and was as expected higher in the *new* condition compared to *given*. For local contour classes corpus statistics yielded that *givenness* was encoded by c_1 and c_4, whereas the classes c_2, c_3, and c_5 encoded new information (χ^2 tests, $\alpha = 0.05$). These results are summarized in Table 2.

In order to perceptually verify these hypotheses the same subjects as introduced in Section 3.3 were asked to participate. They had to judge a stimulus *Yes, an X (e.g. Yes, a flower)* on a five level bipolar scale whether it is rather an answer to the question *'Is this an X? (Is this a flower?')* or to *'Is this a hypernym(X)? (Is this a plant?)'*. If the stimulus *'Yes, a flower'* is perceived as an answer to *'Is this a flower?'*, it is considered as a confirmation not containing any new information. However, as an answer to *'Is this a plant?'* it contains new information that specifies the hypernym *'plant'*. Thus, the judgment of the stimulus on the bipolar scale reflects the subjects' opinion whether its intonation encodes rather given or new information.

The generation of the stimulus part '*... an X*' was carried out as described in the previous section. The initial particle '*Yes, ...*' received a linear F0 contour falling from 90 to 80 Hz and was followed by a 300 ms pause.

As is shown by the left boxplots of Figure 6, except for class c_4 the results were in line with the hypotheses: c_2, c_3, and c_5 were clearly perceptually bound to new information, and c_1 to givenness (Kruskal-Wallis test, $\chi^2_4 = 217.1, p < 0.001$, Dunnett post-hoc test, $\alpha = 0.05$). As with finality all classes were perceptually attributed to information status, again expressed by the significant differences of all judgments from the mean level 3 (sign tests, Bonferroni-corrected, $|z| > 4.3, p < 0.001$).

4. Text-based local contour class prediction

4.1. Prediction models

For the global classes the predictions simply emerge from the Table 1. g_1 can be chosen for any non-initial (thus medial or final) position within a discourse segment. g_2 is suggested for both initial and final but not medial position, whereas g_3 is complementary restricted to medial (thus non-initial and non-final) positions.

For the local classes in accordance with our hypotheses in Table 2 derived from corpus statistics a tree was handcrafted for choosing the appropriate class based on the simplified discourse structure of an utterance. The linguistic concepts are represented by the non-terminal nodes of the tree, their values by the outgoing branches, and local contour classes by the leafs. Thus, each path through the tree represents one discourse structure setting and ends in a leaf suggesting a corresponding local contour class. This tree is presented in Figure 7.

4.2. Perceptual evaluation

4.2.1. Subjects and method

Subjects. 14 subjects took part in the perceptual evaluation of global classes, and 15 subjects in the evaluation of local classes. All subjects are German native speakers between 21 and 53 years, and all of them are phonetic experts working at the Institute of Phonetics and Speech Processing in Munich or post-graduate students of Phonetics. The subject groups for both evaluations are not identical but overlap. The author did not take part as a subject.

Method for global class evaluation. The subjects were presented with pairs of sentences that were either coherent or not. They had to judge the intonation with respect to whether or not it can correctly express the connectedness of the sentences. Coherence was established by pronouns and topic relatedness, separation by topic incongruity.

Examples for a coherent and an incoherent sentence pair used in this validation are:

> **coherent:** Dort gibt es Bienen. Ihr Honig ist lecker. (There are bees. Their honey is tasty.)
> **incoherent:** Dort gibt es Bienen. Die Fähre kam pünktlich. (There are bees. The ferry arrived in time.)

All eight sentence pairs are listed in Appendix B.1. Each sentence is considered as a global segment, within which a global contour is fitted. For each sentence pair four global contour variant types V_* were generated:

- V: the global contour classes suggested by the prediction model in section 4.1 in order to establish sentence coherence or incoherence,
- V_1: only the contour class of the first sentence matches the prediction,
- V_2: only the contour class of the second sentence matches the prediction, and
- V_0: both contour classes contradict the prediction.

This can be related to the coherent sentence pair example above by the following schema:

	sentence 1	sentence 2
Discourse	non-final	non-initial
Variants		
V	g_3	g_1
V_1	g_3	g_2
V_2	g_1	g_1
	g_2	g_1
V_0	g_1	g_2
	g_2	g_2

Since in this case sentences 1 and 2 are coherent, a non-final contour should be used for the first sentence, and a non-initial contour for the second. The variant type V in line with these recommendations thus assigns the non-final g_3 class to the first sentence and the non-initial g_1 class to the second. Variant type V_1 agrees with the recommendations only concerning the finality-status of the first contour class g_3 but contradicts the recommendations in assigning the initial class g_2. Variant type V_2 agrees with V with respect to the contour class of the second sentence (non-initial g_1) but not for the first sentence, to which it assigns final g_1 or g_2. V_0 does not agree with the recommendations at all assigning a final class to sentence 1 (g_1 or g_2) and the initial class g_2 to sentence 2. All four variation types are shown in Appendix B.1. Since the subject's task was to judge the intonation marking of the sentence coherence and not the marking of the end of the utterance or the sentence mode, for the final sentence only the falling slope classes g_1 and g_2 were applied.

The F0 contour of a stimulus was generated in the following way: Each sentence was assigned a global contour according to the schema described above. No phonetic realization model (cf. Section 2.3.4) was applied for the global contours in order to directly examine the impact of the contour class on the listener judgments. The initial F0 value of the first sentence was set in dependence of the first contour class to 75 Hz for g_1, 85 Hz for g_2 and 70 Hz for g_3 to ensure a realistic pitch range throughout the whole utterance. For the second sentence the first F0 value results from the pitch reset regression model described in Section 2.3.4.

Within each local segment a local contour was added to the declination lines. For all global contour variants the F0 values of these local contours were determined by the tree predictor in Figure 7. In the example given above the contour specification of variant V is thus given by

[[*Dort gibt es Bienen*]$_{c2}$]$_{g3}$. [[*Ihr Honig*]$_{c2}$ [*ist lecker*]$_{c5}$]$_{g}$.
(There are bees. Their honey is tasty.)

Bienen (bees) and *Honig (honey)* provide new information and are both in non-final discourse segment position. Thus class c_2 is assigned to both local segments. Note that we have not yet elaborated an intonation class prediction for other word classes than nouns, since the concept of information status is not easily transferable. Nevertheless, in an informal pretest the class c_5 turned out to be appropriate for the adjective *lecker (tasty)* with respect to its discourse characteristics (new, final) as well as to the assigned F0 contour. This could be an indication that the automatic contour predictions may be extended to local segments containing predicative adjectives.

For all incoherent cases the local contour class c_5 was assigned to sentence 1 since it is suggested by the tree for discourse segment final position. A variant V for the incoherent cases is thus:

[[*Dort gibt es Bienen*]$_{c5}$]$_{g1}$]. [*Die Fähre*]$_{c2}$ [*kam pünktlich*]$_{c5}$]$_{g2}$.
(There are bees. The ferry arrived in time.

Thus, the local contour classes were always set to be in line with the tree predictions, whereas the global classes were systematically varied.

To anchor the local contour classes within each local segment the time point 0 of the time-normalized contours was aligned to the midpoint of the nucleus in stressed syllable of the head word, i.e. the final content word. The local contours were subsequently adjusted to the context by means of the corresponding regression models (cf. Section 2.3.4). As described Section 2.3.4 the joints of neighboring local contours were bridged by linear interpolation and smoothed by a moving median filter.

The segment durations needed for the synthesis were derived from the model $\hat{d}_x = \bar{d}_x \cdot f$, where $\hat{d}_x$ is the predicted duration of phoneme x, $\bar{d}_x$ its intrinsic duration set to the mean duration found in a manually segmented sub-part of the SI1000P corpus. f is a factor adjusting the intrinsic duration to the given context defined by accentuation, phrase finality and phoneme class. This factor is predicted by a regression tree [70] trained on a SI1000P sub-part. With one exception (one schwa occurrence was considered to be too long by one subject) the subjects did not report any duration abnormalities.

Between the coherent sentences a pause of length 400 ms was inserted, while for the non-coherent sentences the pause duration was set to 250 ms. This duration difference reduces the impact of pause duration on coherence marking, thus the realization of incoherent sentences is not only judged to be adequate due to a long pause, and vice versa for the realization of coherent sentences. By reducing the reliability of the pause as a cue for coherence, the subjects' attention is expected to be drawn to the intonation contours.

As in the preceding studies the stimuli were synthesized using MBROLA [64] and a German database (*de4*) available on the MBROLA project web page.

Eight test items were presented to the subjects: one for each of the eight sentence pairs shown in Appendix B.1. In each of these trials the appropriateness of each of the four intonation contour variant types had to be judged with respect to how well it marks the presence or absence of coherence. Since always two variant types consist of two class pairs (final, non-initial: g_1, g_1 and g_2, g_1; final, initial: g_1, g_2 and g_2, g_2) actually six contour variants were presented. The subjects were asked to assign one value for each variant on a five-level bipolar Likert scale between the endpoints *non-adequate 1* and *adequate 5*. Thus, in each trial six judgments had to be made. The stimuli were presented via closed headphones, and no upper limit was set to stimulus repetition.

Method for local class evaluation. The subjects had to judge the intonation adequacy of local target segments within different discourse contexts as in:

Dort steht eine Buche. [*Die Buche*]$_{s1}$ *verliert* [*ihre Blätter*]$_{s2}$.
(There is a beech tree. [*The beech tree*]$_{s1}$ *loses* [*its leaves*]$_{s2}$*.)*

The discourse context is provided by the preceding sentence, in this example *There is a beech tree.* It determines the information status of the referents in the local segments s_1 and s_2. For each of the segments four local contour variant types V_* were generated:

- V: the local contour class suggested by the tree,

- V_i: a contour class only matching the information status encoding requirements,
- V_s: a contour class only matching the discourse segmentation encoding requirements, and
- V_0: a contour class neither matching the information status nor the segmentation encoding requirements.

This can be related to the example above by the following schema:

Context	*There is a beech tree.*	
Carrier	[*The beech tree*]$_{s1}$ *loses* [*its leaves*]$_{s2}$.	
s_1	Discourse	given, non-final
	Variants	V: c_4, V_i: c_1, V_s: c_2, c_3, V_0: c_5
s_2	Discourse	new, final
	Variants	V: c_5, V_i: c_2, c_3, V_s: c_1, V_0: c_4

The local segment s_1 is located in the *non-final* position and carries *given information*. As can be seen in Figure 7 the tree suggestion is a c_4 type contour. Three types of contrastive variants to this prediction were generated: V_0 (class c_5) differs from c_4 regarding information status as well as discourse segmentation encoding. V_s differs with respect to information status encoding (classes c_2 and c_3), and V_i regarding segmentation (class c_1). All four context-carrier configurations are shown in Appendix B.2.

Again the stimuli were synthesized using MBROLA [64] and the German database *de4*. F0 contours were generated in the following way: Each context and carrier sentence was assigned a global contour component in accordance with the global class prediction model, thus the first sentence received an "initial" g_2 contour and the second a "final" g_1 contour. Both contours were adjusted amongst others to sentence length by the regression model introduced in Section 2.3.4. The first onset was set to 85 Hz, and the second onset was derived dynamically from the pitch reset regression model of Section 2.3.4.

On these declination lines one local contour per local segment was added. The F0 values of these contours were determined by the underlying local contour class. Opposed to the declination lines no phonetic realization models had been applied in order to directly examine the impact of the contour classes in the target segments.

As documented in Appendix B.2 in detail, constant classes were chosen for the context sentences and the verbs in the carrier sentences, and varying classes for the target segments. Within each local segment the time point 0 of the time-normalized contours was aligned to the midpoint of the nucleus in stressed syllable of the head word, i.e. the final content word. The joints of neighboring local contours were bridged as described above. The segment durations were derived from the duration model introduced in the preceding section. The pause between the context and the carrier sentence was constantly set to 200 ms.

Eight test items were presented to the subjects: For each of the four sentence pairs shown in Appendix B.2 one item for the target segment s_1, and a one item for s_2. In each of these trials the appropriateness of each of the four intonation contour variant types on the marked target segment had to be judged with respect to information status and discourse segmentation. Since always one variant type consists of two classes, c_2 and c_3, actually five contour variants were presented. The subjects were asked to assign one value for each variant on a five-level bipolar Likert scale between the endpoints *non-adequate 1* and *adequate 5*. Thus, in each trial five judgments had to be made. The stimuli were presented via closed headphones, and no upper limit was set to stimulus repetition.

4.2.2. Results

Global class prediction. The judgments for all variants are presented as boxplots in Figure 8. Results were the following:

- The predicted global contours were generally accepted. The judgment median is 4 and thus significantly higher than the mean judgment level 3 (one-sided one-sample sign test for median comparison, $z = 4.12$, $p < 0.001$; [71]).
- The predictions V were significantly higher rated than alternatives V_2 and V_0 (Kruskal-Wallis test, $\chi^2_3 = 9.36$, $p < 0.05$, Scheffé post-hoc test, $\alpha = 0.05$; [72]).
- There was no significant difference between V and V_1 indicating that listeners rather relate the declination pattern of the first sentence to topic coherence or shift.

Local class prediction. The judgments for all variants are presented as boxplots in Figure 9. Results were the following:

- The contours suggested by the tree were generally accepted. The judgment median is 4 and thus significantly higher than the mean judgment level 3 (one-sided one-sample sign test for median comparison, $z = 7.28$, $p < 0.001$).
- The tree predictions V were rated significantly higher than variant types V_i and V_o in isolation (Kruskal-Wallis test, $\chi^2_3 = 117.12$, $p < 0.001$, Scheffé post hoc test, $\alpha = 0.05$). V showed also a tendency to be rated higher than V_s which is reflected in the arithmetic mean rating difference of these variants (3.88 vs. 3.52) and a significant difference yielded by the less conservative Dunnett post-hoc test, $\alpha = 0.05$.
- All other variant ratings differed significantly (Kruskal-Wallis test, $\chi^2_3 = 117.12$, $p < 0.001$, Scheffé post-hoc test, $\alpha = 0.05$). Variant V_s was rated higher than variant V_i indicating that the subjects focused on the discourse segmentation function of intonation rather than on the information status. As to be expected V_0 received the lowest scores.

5. Discussion and Conclusions

5.1. Point of departure for bottom-up modeling

In order to get started with a knowledge-free bottom-up approach of intonation modeling, several restrictions and simplifications concerning the data and its analyses were made. So far only a single speaker was examined. However, due to the choice of a professional newsreader producing clear and easily intelligible speech, it can safely be stated that the data contains intonation patterns that are linked to linguistic concepts with high proficiency. In total, 10 out of 12 corpus-based predictions of these links were confirmed in subsequent perception experiments by [44, 61]. As in the current study in these experiments not the original voice but an MBROLA voice had been presented, which indicates that the found relations are not just speaker-idiosyncratic but more general.

Inter-speaker variability has not yet been addressed, but will be investigated in future studies. In principle, our approach can capture variability by distinguishing between abstract contour classes and concrete F0 realizations. Within this framework it has to be tested whether

it is possible to derive speaker-independent classes which can be mapped to the F0 surface by speaker-dependent regression models as described in Section 2.3.4.

Concerning prosodic structure, the global segmentation guided only by speech pauses and punctuation needs further improvement at the current state, since pauses are neither necessary nor sufficient boundary markers. In a current study [73] we explore how automatically derived F0 discontinuity features can be used to supplement the identification of prosodic boundaries.

In subsequent studies prosodic structure could be extended by the insertion of an intermediate layer between local and global segments, that is defined directly in information structure terms. This layer would contain a sequence of *theme (topic)* and *rheme (comment)* segments [74], that has been proven to have an impact on prosodic phrasing. To give an example, [75] found for English that the prosodic attachment of a verb depends on its information status. It is attached to the subject (theme) if its status is *given*, and to the object (rheme), if its status is *new*. Furthermore, the intermediate layer could serve to examine how information structure is encoded by the syntagmatic combination of local contour classes.

The discourse structure concepts examined here were highly simplified. Again this simplification is considered to be an appropriate starting point given the bottom-up nature of the present approach. Next steps should include interpretations in the context of more fine-grained discourse concepts like different degrees of givenness [23, 43]. Since rather accent type preferences instead of unique mappings of intonation and linguistic events are reported, it can be concluded that principally already a small number of contour classes can be sufficient to encode more complex information status concepts.

Also the NLP methods to generate the hypotheses for the linguistic anchoring were rather crude. They are definitely not sufficient to replace sophisticated methods used for example in speech synthesis for text-based intonation prediction, but nevertheless, they *are* of use for discovering relations between intonation events and their linguistic functions, since in two former studies of the author, [44] and [61], in total 83% of the predictions had been verified by the subsequent perceptual validations.

Taken altogether, we argue that these simplifications are justified to find an appropriate point of departure for our bottom-up modeling approach.

5.2. Intonation units for linguistic interpretation

5.2.1. From parameters to categories

To anchor the polynomial contour stylization that contains a high degree of variation to linguistic concepts the CoPaSul framework provides an intermediate abstraction level by inference of discrete classes from the continuous parameter space. The quality of this mapping depends amongst others on the choice of the stylization function. Due to its analytical nature, the polynomial contour fit is more suited for parameter clustering than the numerical fit of more complex functions [24], since by the latter parameter vector distances are more loosely linked to contour distances.

Other parametric approaches like the Fujisaki model allow for a direct parameter interpretation [7, 8]. But again this semantically rich parameterization does not guarantee the reproducibility of the intonation abstraction due to the underlying numerical optimization. If different parameter values can be derived from the same underlying original contour, the linking of the parameters to linguistics may not be possible in a straight-forward linear manner.

Generally, there is no straightforward solution for an appropriate derivation of categories amongst others due to the various kinds of objects that could be clustered. We followed [24]

in clustering stylization coefficient vectors, but alternatively, the stylized F0 vectors themselves could have been used. First, there is a methodological justification for using coefficient and not F0 vectors: In machine learning dimension reduction and orthogonalization of the feature vectors is beneficial [76]. While the F0 vectors consist of a high number of elements that are highly correlated among each other, polynomial coefficients are exactly the result of a dimension reduction and orthogonalization of these vectors, and are thus expected to be more appropriate.

Second, to test the linguistic adequacy of this alternative classification, we clustered the stylized F0 vectors in exactly the same way as the coefficients. As introduced in section 3.3 we tested by means of χ^2 tests, whether or not the four derived contour co-occur significantly with the examined discourse events. The relation to utterance finality was significant for all four classes ($\alpha = 0.05$). However for none of the classes a significant relation to information status has been found ($\chi < 1.2$). It can thus be concluded that for our data coefficient vector classes are more appropriate to reflect discourse structure than F0 vector classes.

5.2.2. Contours instead of tones

Analogously to the approach of [17] for ToBI label sequences, we trained a contour class trigram probability model on the local contour class sequences of our data. Given that the vocabulary consists of only five classes all trigram probabilities turned out to be relatively low (all maximum likelihood estimates are below 0.34). This means that the contour class predictability in a given intonation context is much lower than was found in [13] for pitch accent and boundary tone sequences within the tone sequence framework. One can draw two conclusions from this observation: First, the intonotactics of CoPaSul intonation events is not as restricted as found for tone sequence approaches and for other models like the IPO model [77, 78]. Second, following the argument of [13] low overall n-gram predictability can be seen as an indication for having extracted proper basic building blocks and not block fragments co-occurring with probabilities near one. Thus, these findings support the position that elementary intonation units are better expressed as contours than tones.

5.3. Relations between contour classes and linguistic concepts

So far the local contour have been linked to semantic weight and elementary discourse structure concepts [44, 61]. As already pointed out the test of many other links like sentence mode, contrast constructions and paralinguistics is pending. It is not yet clear how far one can get with linguistic interpretations of this knowledge-free bottom-up approach of intonation modeling, but nevertheless, the initial attempts are promising.

Since the local contour classes cover all combinations of information status (given, new) and discourse segmentation (non-final, final), the derived set is sufficient at least for a crude encoding of discourse structure. It was further possible to perceptually attribute a linguistic function to each contour class since all mean perceptual judgments differed significantly from the undecidedness level 3 on the 5 point scales. The relations turned out to be stable, since in the present study the predicted contours received high scores by the subjects across different intonation contexts.

There is no one-to-one mapping between local intonation classes and linguistic concepts: More than one class can be used to encode the same concept (class correlations) and the same class can be used to encode several concepts (concept correlations). Concept correlations are in line with the results in [79] and [23] within the tone sequence framework revealing that the same sequence of tones can have different pragmatic functions context-dependently and can encode

different sentence modes. The class correlations that were found here do not allow for a compositional approach as in [23] distinguishing between pitch accents connected to the information status of a discourse referent and boundary tones encoding finality or continuation. All local contour classes derived in this study serve at the same time to encode information status and discourse segmentation.

It is to be pointed out that our contour prediction is not yet fully automated, since local contour classes can so far only automatically be assigned to local segments containing nouns. Thus, further effort is needed to extend the predictions to local segments of any type. As stated in section 4.2 the prediction was already applicable also to adjectives in predicative position.

5.4. Perceptual evaluation

This study aimed to integrate previous findings about the linguistic anchoring of the CoPaSul approach which had been derived by perception experiments with a higher number of subjects. These findings were incorporated into a tree for text-based intonation prediction whose output was evaluated by phonetic experts. The choice of expert subjects allowed for a compact experimental setup. The relatively complex demands on the subject decisions would have made it necessary to run several experiments with naive subjects. Only simple sentences have been presented in order to allow for a straightforward linking of the subjects' judgments to the responsible intonation characteristics which eases the interpretation of the results. In other perception studies [80, 81] the discourse context is specified in more detail by longer texts. The advantage over the single-sentence contexts of the current study is, that the subjects are constrained not to add other contextual factors than the ones being specified. On the other hand, processing more complex discourse contexts is more demanding for the subjects, and thus may lead to misunderstandings or even interferences across different contexts. In any case, in the experiments introduced in this study no subject reported any difficulties arising from vague linguistic contexts.

Despite the small sample sizes and the required conservative statistical tests, significant results emerged concerning the acceptability of the generated contours. The intonation that conformed to the predictions was judged to be adequate significantly above the midpoint of the rating scale for global and local contour classes. Furthermore, for local contour classes it was judged to be significantly more adequate than intonation patterns contradicting the predictions. It is very likely that these differences remain significant with increasing sample size, which would additionally allow for less conservative tests. Thus, the claim that within our purely data-driven framework acceptable and adequate intonation contours can be predicted from text is supported by these results. For global contour classes it turned out that the intonation prediction before and after potential topic shifts (variant V) does not offer any gain compared to the prediction of the preceding intonation only (variant V_1). The global contour class prediction model may thus be reduced to cover only the sentence final global segments.

5.5. Relevance of the CoPaSul approach

The CoPaSul approach is partly eclectic in the sense that it aggregates characteristics of established intonation models presented in Section 1. Regarding the requirements defined in Section 1.2, namely an appropriate abstraction from the signal, the interpretability of the abstraction, and its automation, in our opinion the current approach provides several additional benefits: Opposed to the other parametric models PaintE, PENTA, TILT, and the Fujisaki model, CoPaSul offers a biunique mapping of the F0 contour to the parameter values which is expected to ease linguistic interpretation. Like PaintE, CoPaSul additionally offers a symbolic representation in

form of intonation contour categories, that are derived by clustering. However, while for PaintE the cluster number was set manually in an ad hoc manner, in the current approach it arises directly from the data, using a subtractive clustering technique. This approach is suitable to obtain repeatable results for disjunct data subsets concerning contour class number and shapes.

For PaintE, TILT, and most applications of the Fujisaki model, the domain of local F0 behavior is defined in terms of realized pitch accents and boundary tones, while CoPaSul operates on potential accent groups containing zero or one accent. The advantages of this approach are that no preceding labeling of these prosodic events is required, and not only pitch accents but also deaccentuation can be captured as an intonation event. In contrast to the PENTA model which can also capture deaccentuation by describing local F0 behavior for each syllable, CoPaSul makes contour characteristics spanning over several syllables more explicit, which is considered to be favorable at least for languages that do not contain lexical tones.

5.6. Conclusion

It was possible to generate a perceptually acceptable intonation representation in a data-driven partly automatic way. This representation can be interpreted linguistically with respect to discourse structure and can thus be derived from the signal as well as from text. Therefore, this approach can be of relevance for intonation analysis and synthesis and can be useful for speech technology applications as well as for phonetic fundamental research for the automatic analysis of speech data.

References

[1] J. Pierrehumbert, The phonology and phonetics of English intonation, Ph.D. thesis, MIT, Cambridge, Massachusets, 1980.

[2] C. Gussenhoven, The Phonology of Tone and Intonation, Research Surveys in Linguistics, Cambridge University Press, 2004.

[3] D. Ladd, Intonational Phonology, Cambridge University Press, Cambridge, Massachusets, 2 edition, 2008.

[4] Y. Xu, Speech melody as articulatorily implemented communicative functions, Speech Communication 46 (2005) 220–251.

[5] S. Prom-on, Y. Xu, B. Thipakorn, Modeling tone and intonation in Mandarin and English as a process of target approximation, Journal of the Acoustical Society of America 125 (2009) 405–424.

[6] H. Fujisaki, A note on physiological and physical basis for the phrase and the accent components in the voice fundamental frequency contour, in: O. Fujimura (Ed.), Vocal physiology: voice production, mechanisms, and functions, Raven, New York, 1987, pp. 165–175.

[7] B. Möbius, Ein quantitatives Modell der deutschen Intonation: Analyse und Synthese von Grundfrequenzverläufen, Niemeyer-Verlag, Tübingen, 1993.

[8] H. Mixdorff, Intonation Patterns of German – Model-based Quantitative Analysis and Synthesis of F0-Contours, Ph.D. thesis, TU Dresden, 1998.

[9] P. Taylor, Analysis and Synthesis of Intonation using the Tilt Model, Journal of the Acoustical Society of America 107 (2000) 1697–1714.

[10] G. Möhler, Describing intonation with a parametric model, in: Proc. ICSLP 1998, Sydney, Australia, pp. 2851–2854.

[11] C. Petrone, M. D'Imperio, Tonal structure and constituency in Neapolitan Italian: Evidence for the Accentual Phrase in statements and questions, in: Proc. Speech Prosody 2008, Campinas, Brazil, pp. 301–304.

[12] J. Pierrehumbert, Synthesizing intonation, Journal of the Acoustical Society of America 70 (1981) 985–995.

[13] A. Dainora, Does intonational meaning come from tones or tunes? evidence against a compositional approach, in: Proc. Speech Prosody 2002, Aix-en-Provence, France, pp. 235–238.

[14] E. Dombrowski, O. Niebuhr, Acoustic patterns and communicative functions of phrase-final rises in German: activating and restricting contours, Phonetica 62 (2005) 176–195.

[15] M. Jilka, G. Möhler, G. Dogil, Rules for the Generation of ToBI-based American English Intonation, Speech Communication 28 (1999) 83–108.

[16] A. Black, A. Hunt, Generating F0 contours from ToBI labels using linear regression, in: Proc. ICSLP 1996, volume 3, Philadelphia, Pennsylvania, pp. 1385–1388.
[17] A. Dainora, Eliminating downstep in prosodic labeling of American English, in: Proc. ISCA Workshop on Prosody, Speech Recognition and Understanding 2001, Red Bank, New Jersey, pp. 41–46.
[18] C. Wightman, ToBI Or Not ToBI?, in: Proc. Speech Prosody 2002, Aix-en-Provence, France, pp. 25–29.
[19] K. Dusterhoff, A. Black, P. Taylor, Using Decision Trees within the Tilt Intonation Model to Predict F0 Contours, in: Proc. European Conf. on Speech Communication and Technology 1999, Budapest, Hungary, pp. 1627 – 1630.
[20] U. Reichel, Datenbasierte und linguistisch interpretierbare Intonationsmodellierung, Ph.D. thesis, Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität, München, 2010.
[21] S. Baumann, M. Grice, S. Steindamm, Prosodic Marking of Focus Domains – Categorical or Gradient, in: Proc. Speech Prosody 2006, Dresden, Germany, pp. 301–304.
[22] P. Taylor, The rise/fall/connection model of intonation, Speech Communication 15 (1995) 169–186.
[23] J. Pierrehumbert, J. Hirschberg, The Meaning of Intonational Contours in the Interpretation of Discourse, in: P. Cohen, J. Morgan, M. Pollack (Eds.), Intentions in Communication, MIT Press, Cambridge, 1990, pp. 271–311.
[24] G. Möhler, A. Conkie, Parametric modeling of intonation using vector quantization, in: Proc. 3rd ESCA Workshop on Speech Synthesis 1998, Jenolan Caves, Australia, pp. 311–316.
[25] W. Cooper, J. Sorensen, Fundamental frequency in sentence production, Springer, New York, 1981.
[26] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg, TOBI: A standard for labeling English prosody, in: Proc. ICSLP 1992, Banff, Alberta, Canada, pp. 867–870.
[27] M. Grice, R. Benzmüller, Transcription of German Intonation using ToBI tones; The Saarbruecken System, in: Phonus, volume 1, University of the Saarland, 1995, pp. 33–51.
[28] A. Schweitzer, B. Möbius, Experiments in Automatic Prosodic Labeling, in: Proc. Interspeech 2009, Brighton, England, pp. 2515–2518.
[29] B. Grosz, C. Sidner, Attention, intentions, and the structure of discourse, Computational Linguistics 12 (1986) 175–204.
[30] M. Hearst, TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, Computational Linguistics 23 (1997) 33–64.
[31] S. Nakajima, J. Allen, A study on prosody and discourse structure in cooperative dialogues, Phonetica 50 (1993) 197–210.
[32] M. Swerts, D. Bouwhuis, R. Collier, Melodic cues to the perceived "finality" of utterances, Journal of the Acoustical Society of America 96 (1994) 2064–2075.
[33] A. Botinis, B. Granström, B. Möbius, Developments and paradigms in intonation research, Speech Communication 33 (2001) 263–296.
[34] C. Féry, German intonational patterns, Niemeyer, Tübingen, 1993.
[35] J. Hirschberg, J. Pierrehumbert, The intonational structuring of discourse, in: Proc. 24th Annual Meeting, Association for Computational Linguistics 1986, New York, pp. 136–144.
[36] J. Venditti, J. Hirschberg, Intonation and discourse processing, in: Proc. ICPhS 2003, Barcelona, Spain, pp. 107–114.
[37] E. Prince, Toward a taxonomy of given-new information, in: Radical Pragmatics, Academic Press, New York, 1981, pp. 223–255.
[38] W. Chafe, Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing, The University of Chicago Press, Chicago, 1994.
[39] J. Gundel, N. Hedberg, R. Zacharski, Cognitive status and the form of referring expressions in discourse, Language 69 (1993) 274–307.
[40] G. Brown, Prosodic structure and the given/new distinction, in: D. Ladd, A. Cutler (Eds.), Prosody: Models and Measurements, Springer-Verlag, 1983, pp. 67–78.
[41] M. Grice, S. Baumann, R. Benzmüller, German Intonation in Autosegmental-metrical Phonology, in: Jun, Sun-Ah (Eds.), Prosodic Typology: The Phonology of Intonation and Phrasing, OUP, Oxford, 2005, pp. 55–83.
[42] J. Hirschberg, D. Litman, J. Pierrehumbert, G. Ward, Intonation and the intentional structure of discourse, in: Proc. 10th international joint conference on Artificial intelligence 1987, volume 2, Milan, Italy, pp. 636–639.
[43] S. Baumann, The Intonation of Givenness – Evidence from German, Ph.D. thesis, Saarland University, 2006.
[44] U. Reichel, The CoPaSul intonation model, in: B. Kroeger, P. Birkholz (Eds.), Elektronische Sprachverarbeitung 2011, Studientexte zur Sprachkommunikation, TUDpress, 2011, pp. 341–348.
[45] F. Schiel, Automatic Phonetic Transcription of Non-Prompted Speech, in: Proc. ICPhS 1999, San Francisco, California, pp. 607–610.
[46] K. Schaefer-Vincent, Pitch period detection and chaining: Method and evaluation, Phonetica 40 (1983) 177–202.
[47] C. de Boor, A Practical Guide to Splines, number 27 in Applied Mathematical Sciences, Springer, 1978.
[48] P.-O. Persson, G. Strang, Smoothing by Savitzky-Golay and Legendre Filters, in: D. Gilliam (Ed.), Mathematical systems theory in biology, communications, computation, and finance, Springer, 2003, pp. 301–315.

[49] J. Nelder, R. Mead, A simplex method for function minimization, Computer Journal 7 (1965) 308–313.
[50] U. Reichel, Improving Data Driven Part-of-Speech Tagging by Morphologic Knowledge Induction, in: Proc. AST Workshop 2007, Maribor, Slovenia, pp. 65–73.
[51] U. Reichel, Perma and Balloon: Tools for string alignment and text processing, in: Proc. Interspeech 2012, Portland, Oregon, p. paper no. 346.
[52] J. Gee, F. Grosjean, Performance structures: a psycholinguistic and a linguistic appraisal, Cognitive Psychology 15 (1983) 411–458.
[53] S. Abney, Parsing By Chunks, in: R. Berwick, S. Abney, C. Tenny (Eds.), Principle-Based Parsing, Kluwer Academic Publishers, Dordrecht, 1991, pp. 257–278.
[54] D. Hermes, Measuring the perceptual similarity of pitch contours, Journal for Speech, Language, and Hearing Research 41 (1998) 73–82.
[55] O. Niebuhr, Categorical perception in intonation: a matter of signal dynamics?, in: Proc. Interspeech 2007, Antwerpen, Belgium, pp. 109–112.
[56] S. Chiu, Fuzzy Model Identification Based on Cluster Estimation, Journal of Intelligence & Fuzzy Systems 2 (1994) 267–278.
[57] U. Reichel, F. Kleber, R. Winkelmann, Modelling similarity perception of intonation, in: Proc. Eurospeech 2009, Brighton, England, pp. 1711–1714.
[58] B. Connell, D. Ladd, Aspects of pitch realisation in Yoruba, Phonology 7 (1990) 1–30.
[59] J. de Pijper, A. Sandermann, On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues, Journal of the Acoustical Society of America 96 (1994) 2037–2047.
[60] M. Swerts, R. Geluykens, Prosody as a marker of information flow in spoken discourse, Language and Speech 37 (1994) 21–43.
[61] U. Reichel, Automatisation of intonation modelling and its linguistic anchoring, in: Proc. Speech Prosody 2012, Shanghai, China, pp. 63–66.
[62] D. Bolinger, Intonation: Levels Versus Configurations, Word 7 (1951) 199–210.
[63] K. Pearson, On the criterion that a given system of derivations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50 (1900) 157–175.
[64] T. Dutoit, F. Bataille, V. Pagel, N. Pierret, O. van der Vreken, The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes, in: Proc. ICSLP 1996, Philadelphia, Pennsylvania, pp. 1393–1396.
[65] W. Kruskal, W. Wallis, Use of ranks in one-criterion variance analysis, Journal of the American Statistical Association (1952) 583–621.
[66] D. C.W., A multiple comparison procedure for comparing several treatments with a control, Journal of the American Statistical Association 50 (1955) 1096–1121.
[67] J. Karas, I. Savage, Publications of Frank Wilcoxon (1892–1965), Biometrics 23 (1967) 1–10.
[68] O. Dunn, Multiple Comparisons Among Means, Journal of the American Statistical Association 56 (1961) 52–64.
[69] M. Hearst, Automatic acquisition of hyponyms from large text corpora, in: Proc. International Conference on Computational Linguistics 1992, volume 2, Nantes, France, pp. 539–545.
[70] L. Breiman, J. Friedman, C. Stone, R. Olshen, Classification and Regression Trees, Wadsworth & Brooks, Pacific Grove, California, 1984.
[71] H. Mann, D. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Annals of mathematical Statistics 18 (1947) 50–60.
[72] H. Scheffé, The Analysis of Variance, Wiley, New York, 1959.
[73] U. Reichel, K. Mády, Parameterization of F0 register and discontinuity to predict prosodic boundary strength in Hungarian spontaneous speech, in: P. Wagner (Ed.), Elektronische Sprachverarbeitung 2013, Studientexte zur Sprachkommunikation, TUDpress, 2013, pp. 223–230.
[74] M. Halliday, Intonation and Grammar in British English, Mouton, Den Haag, 1967.
[75] M. Steedman, The syntactic process, MIT Press, Cambridge, Massachusets, 2000.
[76] P. Cunningham, Dimension Reduction, Technical Report UCD-CSI-2007-7, University College Dublin, 2007.
[77] J. t'Hart, R. Collier, A. Cohen, A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody, Cambridge University Press, Cambridge, 1990.
[78] S. Noteboom, The Prosody of Speech: Melody and Rhythm, in: W. Hardcastle, J. Laver (Eds.), The Handbook of Phonetic Sciences, Blackwell, Oxford, 1997, pp. 653–668.
[79] G. Ward, J. Hirschberg, Implicating uncertainty: The pragmatics of fall-rise intonation, Language (1985) 747–776.
[80] O. Niebuhr, Perzeption und kognitive Verarbeitung der Sprechmelodie: Theoretische Grundlagen und empirische Untersuchungen, Ph.D. thesis, IPDS, Christian-Albrechts-Universität zu Kiel, 2007.
[81] P. Welby, Effects of Pitch Accent Position, Type, and Status of Focus Projection, Language and Speech 46 (2003) 53–81.

Appendix A. Clustering, contour class, and regression model characteristics

For cluster initialization subtractive clustering [56] was applied, that iteratively locates initial centers in the parameter space at regions with high data density. It was carried out using the SUBCLUST Matlab function. The following four parameters need to be initialized: (1) the cluster radius, that determines the size of the clusters, (2) the squash factor, which is to be multiplied with the radius and defines the neighborhood of cluster centers within which new centers are discouraged, (3) the accept ratio to set the minimum neighborhood density as a fraction of the density of the first cluster center, above which another data point will be accepted as a center, and (4) the reject ratio, that gives the neighborhood density as a fraction of the density of the first cluster center, below which a data point will be rejected as a center. Generally, a low number of non-overlapping clusters is derived by reasonably high values for all four parameters. To find a local optimum for these parameter values for a sample of 20% of the data, the Nelder-Mead Simplex optimization (Matlab function FMINSEARCH) was utilized. The error to be minimized was derived from the mean clustering silhouette as in [20]. The silhouette for the data point i is defined as $S(i) = \frac{d_B(i)-d_A(i)}{\max(d_A(i),d_B(i))}$. $d_A(i)$ is the mean squared Euclidean distance of point i to all points of the same cluster. $d_B(i)$ is mean distance of point i to all points of the most i-similar cluster $B \neq A$. Its values fall within the range from -1 to 1. The closer the silhouette approaches 1 the more clearly and thus better is the assignment of i to its cluster. Values close to -1 indicate and erroneous assignment. The mean silhouette $\bar{S}$ over all data points was transformed to an error measure e ranging from 0 to 1 by the following equation: $e = \frac{1-\bar{S}}{2}$. The optimized parameter values and the errors are listed separately for global and local contour clustering in Table A.1.

Subsequently, k-means clustering was carried out starting with these initial cluster centers. In Tables A.2 and A.3 the polynomial coefficients of the global and local contour class centroids are presented.

Table A.4 shows the coefficients of the linear regression models mapping the contour class centroids to context dependent realizations.

	global classes	local classes
radius	0.3075	0.3075
squash factor	1.1875	1.2812
accept ratio	0.5125	0.5125
reject ratio	0.1537	0.1425
error	0.1693	0.3078

Table A.1: Subtractive clustering parameter values optimized by the Nelder-Means method. Details on these parameters and the error definition are presented in Appendix Appendix A.

class	slope	p	l
g_1	−4.2176	0.46	17
g_2	−9.2595	0.29	15
g_3	1.0208	0.25	14

Table A.2: Global contour class characteristics. p: relative frequency; l: mean length in syllables.

class	a_0	a_1	a_2	a_3	p	l
1	0.2537	−0.9297	0.7758	0.4436	0.22	5
2	5.1403	6.8721	−0.9293	−7.2646	0.18	4
3	3.5853	−6.2229	0.9980	7.8085	0.17	6
4	2.7439	3.1747	3.1763	−2.4418	0.20	5
5	5.8955	−0.2384	−1.7163	0.2772	0.23	5

Table A.3: Local contour class characteristics. Polynomial coefficients a_k from $\sum_{k=0}^{3} a_k \cdot t^k$; p: relative frequency; l: mean length in syllables.

	Local contours				Global contours	Pitch reset
	a_o	a_1	a_2	a_3	b	r
w_o	-0.2187	-0.3150	-0.4874	-0.2890	-0.2954	0.1168
w_1	0.3000	0.4092	0.0338	0.4086	0.5077	-0.5490
w_2	0.2912	0.0102	0.0227	0.0031	0.0067	-0.4956
w_3	0.0658	-0.0357	-0.0371	-0.0388	-0.0221	0.2166
w_4	-0.0381	-0.0100	-0.0159	-0.0409	–	–
ρ	0.69	0.67	0.63	0.71	0.80	0.82

Table A.4: Phonetic realization models introduced in section 2.3.4: regression weight values w_x to adjust the contour class polynomial coefficients and the pitch reset r. a_x: local class coefficients, b: global contour class slope. In the bottom row the Pearson's correlation coefficient ρ of target and predicted values is presented.

Appendix B. Stimuli for adequacy ratings

Appendix B.1. Global contour classes

Coherent sentence pairs.

- *'Dort steht eine Buche. Ihre Blätter sind grün. (There is a beech tree. Its leafs are green.)'*
- *'Da liegt eine Geige. Ihre Saiten sind neu. (There is a violin. Its strings are new.)'*
- *'Dort gibt es Bienen. Ihr Honig ist lecker. (There are bees. Their honey is tasty.)'*
- *'Hier gibt es Birnen. Sie schmecken sehr saftig. (Here we have pears. They are very juicy.)'*

	Discourse	Local contour classes
Sentence 1	non-final	[*Dort steht eine Buche*]$_{c2}$
Sentence 2	non-initial	[*Ihre Blätter*]$_{c2}$ [*sind grün*]$_{c5}$
Global contour pair variants	V: $g_3 + g_1$; V_1: $g_3 + g_2$; V_2: $g_1 + g_1$, $g_2 + g_1$; V_0: $g_1 + g_2$; $g_2 + g_2$	

Incoherent sentence pairs.

- *'Dort steht eine Buche. Meine Schwester hat Fieber. (There is a beech tree. My sister suffers from fever.'*
- *'Da liegt eine Geige. Es gibt Nudeln aus der Dose. (There is a violin. We'll have noodles from the can.)'*

- *'Dort gibt es Bienen. Die Fähre kam pünktlich. (There are bees. The ferry arrived in time.)'*
- *'Hier gibt es Birnen. Der Kellner trägt keine Socken. (Here we have pears. The waiter does not wear socks.)'*

	Discourse	Local contour classes
Sentence 1	final	[*Dort steht eine Buche*]$_{c5}$
Sentence 2	initial	[*Meine Schwester*]$_{c2}$ [*hat Fieber*]$_{c5}$
Global contour pair variants	V: $g_1 + g_2$, $g_2 + g_2$; V_1: $g_1 + g_1$, $g_2 + g_1$; V_2: $g_3 + g_2$; V_0: $g_3 + g_1$	

Appendix B.2. Local contour classes

- **Sentence pair 1:** *'Dort steht eine Buche. Die Buche verliert ihre Blätter. (There is a beech tree. The beech tree is losing its leaves.)'*

Context	*Dort steht eine Buche.*	
Carrier	[*Die Buche*]$_{s1}$ *verliert* [*ihre Blätter*]$_{s2}$.	
s_1	Discourse	given, non-final
	Variants	V: c_4; V_i: c_1; V_s: c_2, c_3; V_0: c_5
s_2	Discourse	new, final
	Variants	V: c_5; V_i: c_2, c_3; V_s: c_1; V_0: c_4

- **Sentence pair 2:** *'Dort steht eine Buche. Auch ein Traktor und ein Ochse. (There is a beech tree. Also a tractor and an ox.)'*

Context	*Dort steht eine Buche.*	
Target	[*Auch ein Traktor*]$_{s1}$ [*und ein Ochse*]$_{s2}$.	
s_1	Discourse	new, non-final
	Variants	V: c_2, c_3; V_i: c_5; V_s: c_4; V_0: c_1
s_2	Discourse	new, final
	Variants	V: c_5; V_i: c_2, c_3; V_s: c_1; V_0: c_4

- **Sentence pair 3:** *'Dort steht eine Buche. Die Kinder bewundern die Buche. (There is a beech tree. The children admire the beech tree.)'*

Context	*Dort steht eine Buche.*	
Target	[*Die Kinder*]$_{s1}$ *bewundern* [*die Buche*]$_{s2}$.	
s_1	Discourse	new, non-final
	Variants	V: c_2, c_3; V_i: c_5; V_s: c_4; V_0: c_1
s_2	Discourse	given, final
	Variants	V: c_1, V_i: c_4, V_s: c_5, V_0: c_2, c_3

- **Sentence pair 4:** *'Dort steht eine Buche und eine Scheune. Die Buche verdunkelt die Scheune. (There is a beech tree and a barn. The beech tree darkens the barn.)'*

Context	*Dort stehen eine Buche und eine Scheune.*	
Target	[*Die Buche*]$_{s1}$ *verdunkelt* [*die Scheune*]$_{s2}$.	
s_1	Status	given, non-final
	Variants	V: c_4; V_i: c_1; V_s: c_2, c_3; V_0: c_5
s_2	Status	given, final
	Variants	V: c_1; V_i: c_4; V_s: c_5; V_0: c_2, c_3

For the background segments, i.e. the local segments other than the target segments s_1 and s_2 following contour classes have been used:

Background Segment	Contour class
[*Dort steht eine Buche*]	c_5
[*Dort stehen eine Buche*] ...	c_4
... [*und eine Scheune*]	c_5
... [*verliert*] ...	c_1
... [*bewundern*] ...	c_5
... [*verdunkelt*] ...	c_5

Remarks: The verb *steht (is)* was treated as an auxiliary as in English, so that it does not demand its own local segment. In order to avoid that the background segments affect the adequacy judgments of the target segments (1) the first sentence was separated from the second one, that contains the targets, by means of the finality encoding contour class c_5. (2) For the other background segments the 3 least prominent contour classes c_1, c_4, and c_5 had been chosen, whereas class prominence had been derived from a former perception experiment [20].

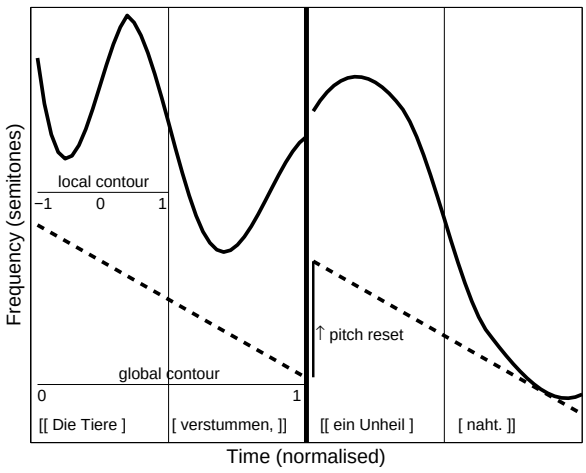

Figure 1: CoPaSul F0 representation as a superposition of global and local intonation contour classes for the utterance *Die Tiere verstummen, ein Unheil naht* (*The animals hush, a disaster is approaching.*)

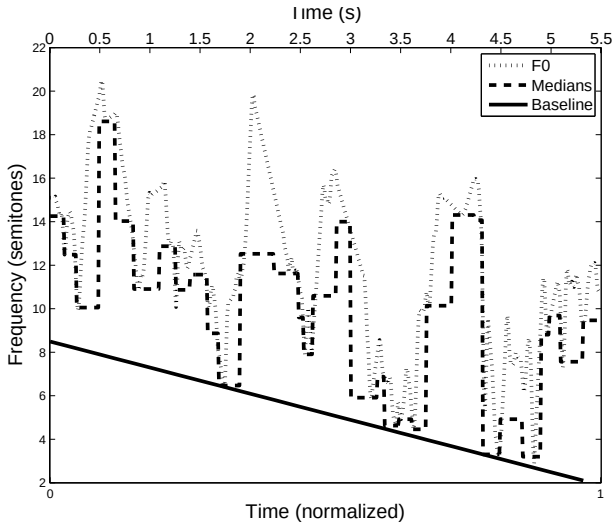

Figure 2: Linear global contour stylization in form of a baseline within a global intonation segment of our data. The contour is given by the flattest bottom tangent through syllable related F0 minima. Absolute and normalized time values are given in the top and bottom abscissa, respectively.

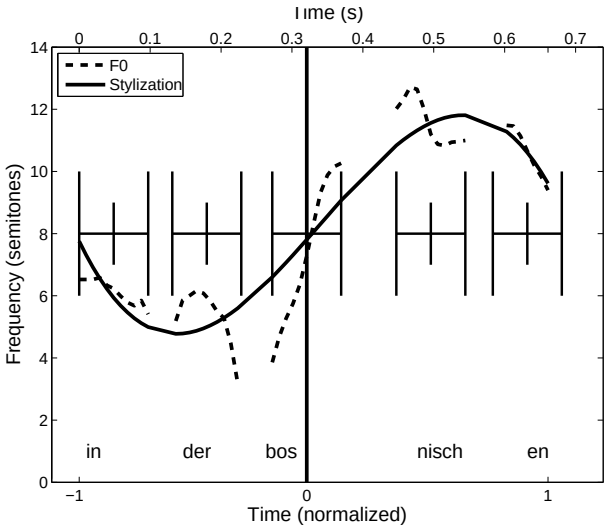

Figure 3: Local contour stylization within a time-normalized local intonation segment by a third order polynomial. The stylization is based on the F0 values in 110 ms time windows centered on the syllable nuclei. The original F0 values are given by the thin discontinuous line, the stylization is shown by the solid continuous line. The five blocks represent the time windows around the syllable nuclei marked by the thin midline. Absolute and normalized time values are given in the top and bottom abscissa, respectively. Normalized time 0 is assigned to the nucleus of the stressed syllable of the segment-final content word, so that the peak of the F0 contour can be interpreted relative to the accent position. Underlying utterance part: *in der bosnischen (in the Bosnian)*.

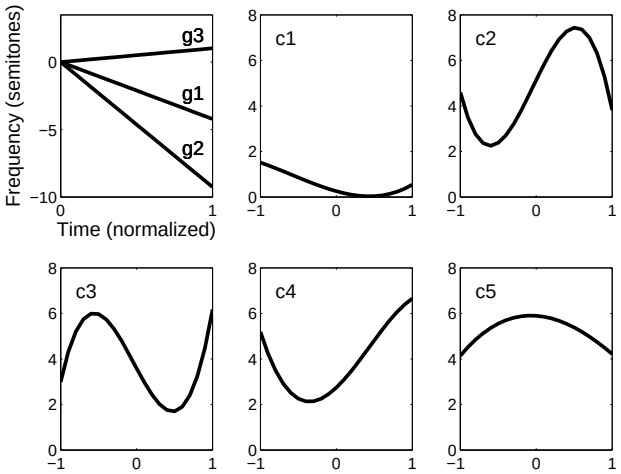

Figure 4: Global (g_{1-3}) and local (c_{1-5}) contour classes.

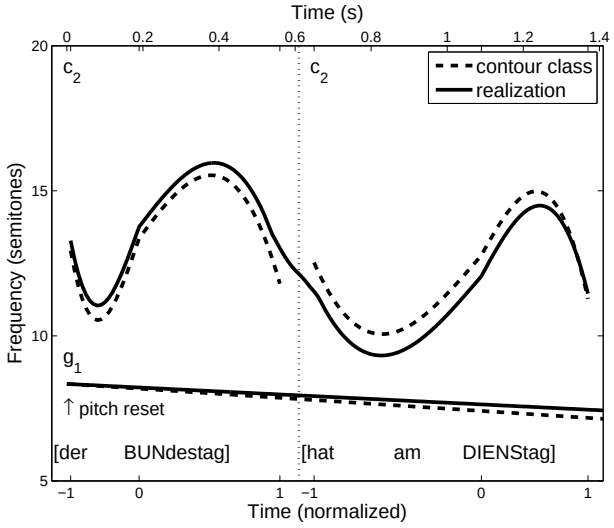

Figure 5: Phonetic realization of the superposition of the global contour class g_1 and a sequence of two local contours of class c_2. Linear regression models mediate between contour classes (dashed lines) and their context-dependent realizations (solid lines). The global contour's initial level is derived from a pitch reset model. The local contours are warped from normalized time (bottom abscissa) to the segmental durations of the utterance (top abscissa) and are added to the global contour. The gap between the local contours is bridged by linear interpolation and subsequently smoothed by a moving median filter. Underlying utterance part: *der Bundestag hat am Dienstag (the Bundestag* [*has*] *on Tuesday).*

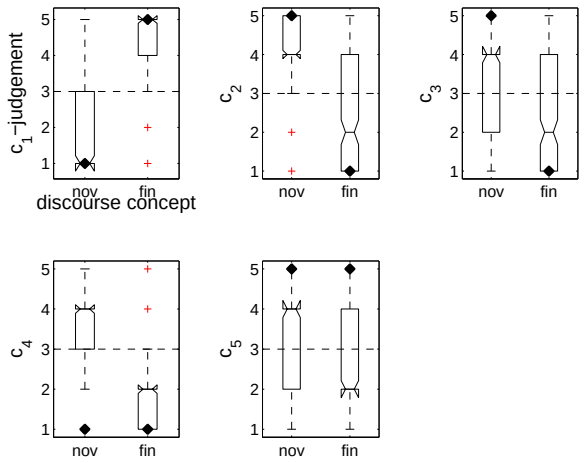

Figure 6: Boxplots for the perceptual judgment distributions of the local contour classes $c_1 - c_5$ with respect to novelty (left boxplots; 1=given, 5=new) and finality (right boxplots; 1=non-final, 5=final). The corpus-statistics-based hypotheses about the linguistic function of the contour classes are marked by filled diamonds placed at level 1 or 5, respectively. The hypotheses where confirmed in 8 out of 10 cases (exceptions: c_4 and novelty, c_5 and finality).

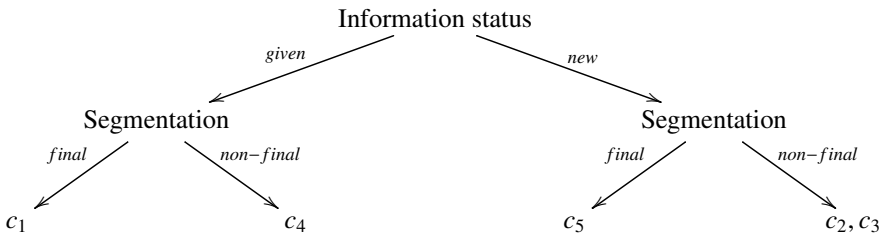

Figure 7: Tree based on the results of the corpus analyses to choose the appropriate local contour class with respect to discourse structure.

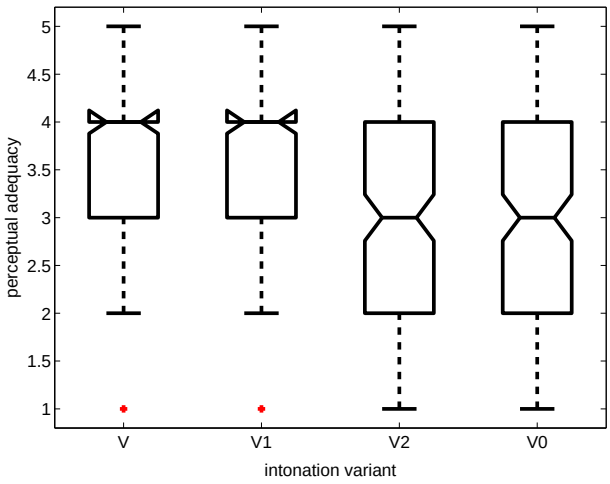

Figure 8: Boxplots for the distributions of adequacy judgments for the predicted global intonation pattern V and its variants V_*. V_1 is in line with the prediction only for the first sentence, V_2 only for the second sentence, and V_0 not at all.

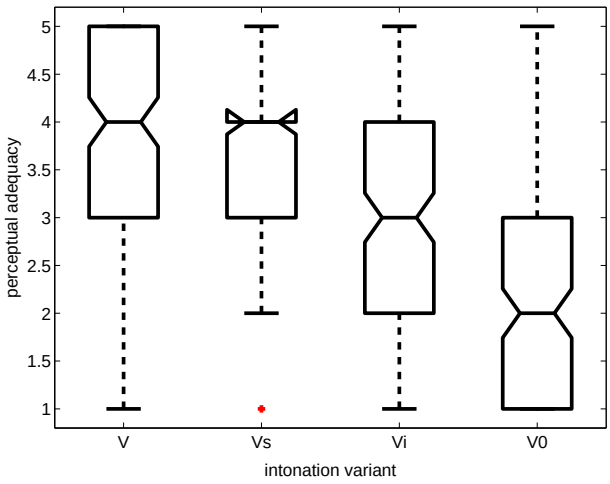

Figure 9: Boxplots for the distributions of adequacy judgments for the predicted local intonation pattern V and its variants V_*. V_i is in line with the prediction only for information status, V_s only for discourse segment finality, and V_0 not at all.