Automatisation of intonation modelling and its linguistic anchoring

Uwe D. Reichel

Institute of Phonetics and Speech Processing, University of Munich

reichelu@phonetik.uni-muenchen.de

Abstract

This paper presents a fully machine-driven approach for intonation description and its linguistic interpretation. For this purpose, a new intonation model for bottom-up F0 contour analysis and synthesis is introduced, the *CoPaSul* model which is designed in the tradition of parametric, contour-based, and superpositional approaches. Intonation is represented by a superposition of global and local contour classes that are derived from F0 parameterisation. These classes were linguistically anchored with respect to information status by aligning them with a text which had been coarsely analysed for this purpose by means of NLP techniques. To test the adequacy of this data-driven interpretation a perception experiment was carried out, which confirmed 80% of the findings.

Index Terms: intonation, modelling, information status, data-driven, perception

1. Introduction

For fundamental intonation research dealing with large amounts of data as well as for its technical applications, it would be helpful to fully automatise the whole processing chain starting from developing an appropriate intonation representation and ending in its linguistic interpretation. If a linguistically interpretable label set can be automatically derived from scratch to annotate the data in a consistent way, time consuming manual labelling at the risk of low intra- and inter-labeler agreement can be avoided.

Existing intonation models can be roughly divided considering the way intonation is represented: symbolically as a sequence of tone labels [1] or parametrically in form of contours [2] which can also be superpositionally arranged [3]. While parametric approaches have the advantage to be more appropriate to automatically extract the intonation representation, symbolic representations might be easier to interprete linguistically. A solution to combine the advantages of both approaches is to adopt the procedure of [2] to transform the continuous parametric representation into a discrete symbolic one by vector clustering. A superpositional representation as in [3] serves to detach the interpretation of local F0 movements from global intonation trends.

The parametric model used in this study has been designed according to these considerations and has already successfully been linked to the concepts of semantic weight and utterance finality [4, 5]. After an introduction of this model, the procedure to relate it to the concept of information status is described as well as a perception experiment to test the adequacy of the model's linguistic anchoring.

2. The *CoPaSul* model

The *CoPaSul* model provides a contour-based (Co), parametric (Pa), and superpositional (Sul) F0 representation. As is shown in Figure 1, F0 contours are treated as a superposition of global and local contour classes which are arranged in a hierarchic prosodic structure.

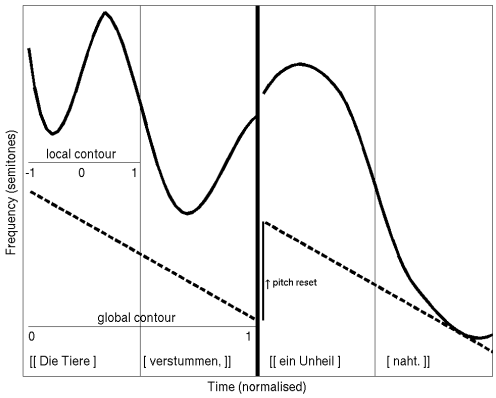

Figure 1: *CoPaSul* F0 representation as a superposition of global and local intonation contour classes for the utterance *Die Tiere verstummen, ein Unheil naht* (*The animals hush, a disaster is approaching.*)

2.1. Data and preprocessing

The data used in this study origins from the SI1000P corpus [6] containing 190 minutes of read German speech of a professional male speaker. F0 contours were extracted by the Schaefer-Vincent algorithm [7] and transformed to semitones (base 50 Hz). F0 errors and voiceless segments were bridged by shape-preserving piecewise cubic Hermite interpolation. The contours were smoothed by a Savitzky-Golay filter of order 3 and window length 5. Pauses and syllable nuclei were detected as described in [4]. On the text level, part of speech tagging was carried out by a tagger developed in [8]. Signal and text were aligned by MAUS [6].

2.2. F0 analysis

The F0 analysis consists of the following steps introduced in further detail below: Prosodic structuring, F0 stylisation, contour class extraction, and specification of their phonetic realisation.

Prosodic structuring A hierarchic prosodic structure as shown by the square brackets in the word tier in Figure 1 is imposed on the data in form of global and local segments which roughly correspond to intonation phrases and accent groups respectively. Global intonation segments are delimited by speech pauses and punctuation. Local intonation segments were de-

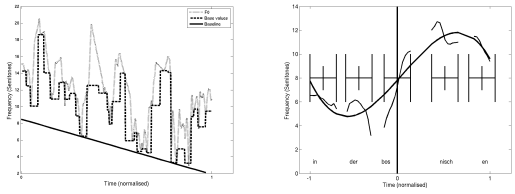

Figure 2: **Left:** Linear global contour stylisation in form of a baseline within a global intonation segment. **Right:** Local contour stylisation within a local intonation segment by a third order polynomial; The stylisation is based only on the F0 values around the syllable nuclei.

fined as a chunk of function words terminated by a content word or a global segment boundary. This notion roughly corresponds to chunking approaches as in [9] and ensures in most cases that each local segment maximally contains one accented syllable.

F0 stylisation All stylisations are based on the F0 values in windows of 110 ms length centered on the detected syllable nuclei. This approach does not require neither an exact syllable segmentation nor a weighting of more and less important parts of the F0 contour.

As can be seen in the left half of Figure 2, within each global intonation segment, a declination baseline is derived by calculating a F0 base value for each syllable defined here as the median of all values below the 10th percentile. The baseline is then adjusted as the flattest possible bottom tangent of the sequence of these base values. The baseline is subtracted from the F0 contours, and its slope is recorded for subsequent clustering.

As shown in the right half of Figure 2, within each local segment a third-order polynomial is fitted to the residuum contour, whereat time is normalised as follows: the time span of the local segment is set from -1 to 1, 0 placed on the nucleus of the stressed syllable of the segment-final word (the content word), so that the peak of the F0 contour can be interpreted relative to the accent position.

Contour classes Contour classes were derived by Kmeans clustering of the range-normalised coefficients. For global classes the baseline slope values were clustered, for local classes the polynomial coefficient vectors with respect to their squared Euclidean distances. Cluster initialisation was carried out by subtractive clustering which itself was optimised by a simplex method on a data subset for mean cluster silhouette maximisation [4]. Figure 3 shows the centroids of the resulting global and local contour classes.

Phonetic realisation models The mapping of the level of abstract contour class centroids to the phonetic level of the intonation surface is carried out by linear regression models. See [4] for further details. Another linear regression model is trained in order to predict pitch reset values [4].

3. Linking the model to information status

3.1. Information status

One way to determine the information structure of an utterance is to divide it into the background containing given information and the focus generally containing new information (see e.g. [10, 11]). Given information is related to the mutual beliefs between speaker and listener about the common state of knowledge. An information is assumed as given if (1) it has already been transferred in the course of the discourse, or (2) it is part

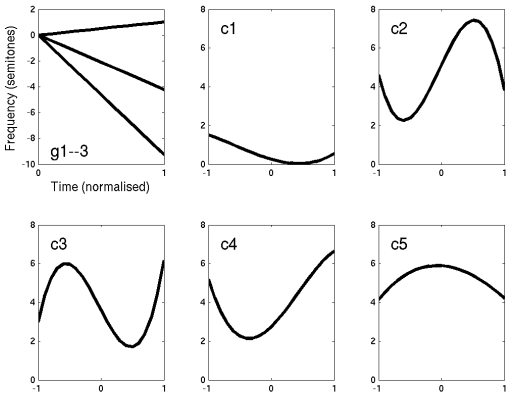

Figure 3: global (g_{1-3}) and local (c_{1-5}) contour classes.

of the common knowledge about the world, or (3) it can be inferred from the situational context. Numerous studies, e.g. [12] and [13], have revealed aspects of prosodic marking of information status, i.e. how given and new information is encoded by means of deaccentuation and pitch accent patterns.

Since modelling world knowledge is far beyond the scope of this study, and the situational context for our data is constantly "reading political newspaper texts", the identification of given as opposed to new information is restricted to information already transferred in the discourse, thus givenness (1).

3.2. Text analysis

To automatically identify given information within the corpus, it was first segmented into thematic units. Subsequently coreference resolution was carried out within each of these units.

3.2.1. Text segmentation

First, words were stemmed by removing all word final inflectional and derivational affixes which were identified by the morphological analysis of [14]. Then, two types of text segmentation were carried out in parallel, a simple sentence segmentation and an adaptation of the TextTiling algorithm developed by [15]. His basic algorithm consists of three components: the *cohesion scorer*, that measures the degree of topic cohesion as the similarity of neighbored text segments, the *depth scorer*, that assigns a depth score to each local minimum within a sequence of cohesion values, and the *boundary selector* deciding whether or not a local minimum indicates a topic shift.

In this study the cohesion scorer operates on neighbored text windows of length 35 (words) separated by a sentence boundary. The text windows were represented as binary term windows weighted by the terms' information contents. This weighting ensures that terms with lower occurrence probability and therefore higher capability to be topic distinctive contribute more to the text similarity measure. The degree of cohesion between the neighbored text segments was then measured as the cosine similarity between their weighted term vectors.

After the assignment of the depth scores to the local minima of the cohesion score sequence, topic shifts were located by the boundary selector at boundaries exceeding a threshold defined with reference to the depth scores' mean and standard deviation.

The TextTiling algorithm was complemented by some

heuristics as "sentence initial conjunctions and pronouns preceeding the first noun in a sentence indicate topic continuity". In an informal evaluation, this approach correctly classified 90% of 103 sentence pairs with respect to topic shift or continuity.

3.2.2. Coreference resolution

Within each text segment coreference relations between nouns were identified. For this purpose coreferentiality was defined as a transitive and anti-symmetric relation on the vocabulary derived from hyperonym-hyponym pairs, hyperonyms being considered as coreferents of their hyponyms. Hyperonym-hyponym pairs were extracted by an iterative pattern matching procedure proposed by [16] and by compound analysis based on [14] treating the less specific compound parts as hyperonyms of the more specific ones. The number of hyperonym-hyponym pairs was further increased by means of the reflexive-transitive closure.

Finally, within each text segment nouns pointing backwards to a coreferent noun were classified as given, otherwise as new. Since only nouns were marked with respect to given and new information for the subsequent examinations only local segments containing a noun were taken into consideration.

3.3. Corpus statistics

3.3.1. Interpretation of the parameterisation

Among the polynomial coefficients of the local intonation contours, only for the F0 offset coefficient a significant difference between new and given information was found (Welch test, $\alpha = 0.5$, $t_{245} = 7.10$, $p < 0.005$). But information status is strongly linked to the F0 maximum and span of the stylised local contours, both values being significantly higher in connection to new information compared to given information (F0 maximum: Welch test, $t_{243} = 5.13$, $p < 0.005$; F0 span: Welch test, $t_{248} = 2.79$, $p < 0.01$).

3.3.2. Interpretation of the local contour classes

The relations of local contour classes and information status derived from χ^2 tests is presented in table 1. Each class is classified with respect to whether it encodes given or new information whenever significant relations have been found for the sentence-level or TextTiling segmentation.

Table 1: *Relations between contour classes and information status depending on the segmentation (t: TextTiling, s: sentence-by-sentence). * marks significance ($\alpha = 0.05$). $P(c|x)$: probability of intonation class c given information status x. $P(c)$: a priori class probability.*

c	status	seg.	χ^2	$P(c\|\text{given})$	$P(c\|\text{new})$	$P(c)$
c_1	given	t	5.09*	0.22	0.18	0.21
		s	52.20*	0.38	0.19	
c_2	new	t	0.92	0.18	0.19	0.18
		s	3.87*	0.14	0.19	
c_3	new	t	20.12*	0.15	0.20	0.19
		s	2.72	0.15	0.19	
c_4	given	t	13.48*	0.21	0.25	0.22
		s	0.13	0.21	0.22	
c_5	new	t	1.07	0.20	0.21	0.20
		s	11.68*	0.12	0.21	

Class c_1 consistently turned out to encode given information for both segmentations. For the other classes complementary and therefore non-contradictive significance patterns have been found. From these corpus statistics five hypothesis can be inferred:

- classes c_1 and c_4 encode given information
- classes c_2, c_3, and c_5 encode new information

These hypotheses were subsequently tested by a perception experiment described in the next section.

3.4. Perceptual validation

3.4.1. Subjects and Method

Subjects 24 subjects (age between 22 and 47, German mother tongue, 19 females, students or researchers of Phonetics) took part in this experiment.

Method The stimuli were resynthesised using MBROLA [17] and a German diphone database (*de4*) available on the MBROLA project web page. They consisted of a carrier sentence *"Ja, eine X" ("Yes, an X")* terminated by a target word X.

Target words The 60 target words were controlled for the following criteria: (1) to be entirely voiced to avoid discontinuous F0 contours, (2) to have uniform syllable characteristics, that is to contain two syllables the first one heavy, open, and stressed, (3) morphologically, to be a simplex form, and (4) a concrete noun to exclude potential semantic and part of speech influences. (5) Their frequencies derived from a newspaper text corpus had to be higher than a lower threshold set to 10. Finally, (6) to allow for a uniform carrier sentence, all target words were of female gender. For all target words, a hyperonym, as *plant* for *flower* was determined.

Duration model The segment durations needed for the resynthesis were derived from the model: $\hat{d}_x = \bar{d}_x \cdot f$. $\hat{d}_x$ is the predicted duration of phoneme x, $\bar{d}_x$ its intrinsic duration set to the mean duration found in a hand-segmented sub-part of the SI1000P corpus. f is a factor adjusting the intrinsic duration to the concrete context defined by accentuation, phrase finality and phoneme class. This factor is predicted by a regression tree [18] trained on the same hand-segmented SI1000P part.

F0 generation For each target word ten intonation variants, one for each of the five local intonation class and five distractors, have been generated the following way: the F0 of the initial part of the carrier sentence (*"Ja, ..."*) was for all stimuli set to a baseline declining from 90 to 80 Hz, followed by a pause of 300 ms. The rest of the carrier sentence and the target word *"eine X"* together form a local segment for which the F0 baseline was set constant to 80 Hz with a slope of 0 to eliminate any global influence on the intonation contour. The contour was thus solely defined by the class-related polynomial centroids. The distractor contours were derived from mean values of varying contour classes triplets in order to construct ambiguous stimuli and to hinder subjects from developing judgment strategies.

Presentation The stimuli were presented in constrained random order via closed head phones. The number of allowed repetitions was not restricted, and the subjects were allowed to pause whenever needed.

The subjects had to judge a stimulus *"Yes, an X"* (e.g. *"Yes, a flower"*) according to its speech melody on a five level bipolar Likert scale whether it is rather an answer to the question *"Is this an X? (Is this a flower?)"* or to *"Is this an hyperonym(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

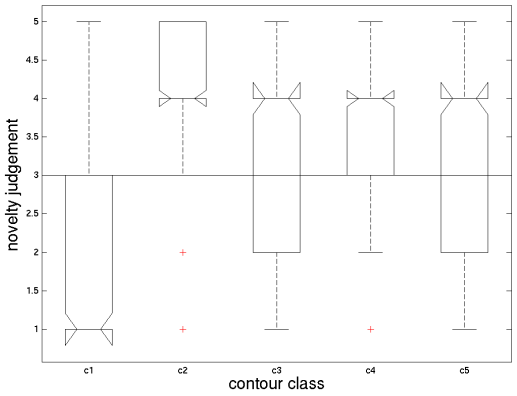

Figure 4: Novelty ratings for local contour classes c_{1-5} on a bipolar 5 level scale, 5 standing for surely new information, and 1 for surely given information.

answer to *"Is this a plant?"* it contains new information that specifies the hyperonym *plant*. Thus, the placing of the stimulus on the bipolar scale reflects the subjects' opinion whether its intonation encodes rather given or new information.

3.4.2. Results

We found significant class-dependent novelty rating differences shown as boxplots in figure 4 (Kruskal-Wallis test, $\chi^2_4 = 217.12$, $p < 0.001$). The mean novelty judgment related to class c_1 was significantly lower compared to the other classes, and it was significantly higher for class c_2 (Dunnett post hoc test, $\alpha = 0.05$). Classes c_3, c_4, and c_5 did not differ significantly.

All judgment mean values differed significantly from the *undecided* level 3 (one-sided single-sample sign test for median comparison, $|z| > 4.27$, $p < 0.001$), being lower for c_1 and higher for all other classes. Reflected in low interquartile ranges the judgments for all contour classes were significantly more consistent than random level obtained when assuming a uniform judgment distribution (one-sided single sample sign tests for median comparison, $z < -2.65$).

Taken altogether, the predicted encoding of givenness was confirmed for class c_1 but disproved for class c_4. The predicted encoding of novelty was confirmed for all affected classes, so that in four out of five cases the automatic assignment of linguistic function to the intonation classes was successful.

4. Discussion

It could been shown in this and in previous studies [4, 5], that it is possible to link an automatically derived intonation representation at least to crude linguistic concepts as semantic weight, dichotomous information status, and utterance finality.

To be principally accessible for linguistic interpretation an intonation representation should be reproducible, meaning that the same F0 contour always is mapped on the same parameter values. In contrast to other parametric models based on numeric optimisation [3] the *CoPaSul* model facilitates such a biunique relation, simply by using polynomials for a stylisation in an analytic and not numeric way.

By the restriction of the underlying data to only one professional speaker the issue of inter-speaker prosodic variability has not yet been addressed, but at least it is assured that the examined speaker's intonation is commonly acceptable.

Given the model's linguistic anchoring, it offers an intonation representation which can be inferred from the signal as well as from text and therefore can be of interest for intonation analysis and synthesis in fundamental research and speech technology.

5. References

[1] J. Pierrehumbert, "The phonology and phonetics of Englisch intonation," Ph.D. dissertation, MIT, Cambridge, MA, 1980.

[2] G. Möhler and A. Conkie, "Parametric modeling of intonation using vector quantization," in *Proc. 3rd ESCA Workshop on Speech Synthesis*, 1998, pp. 311–316.

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

[4] U. Reichel, "Datenbasierte und linguistisch interpretierbare Intonationsmodellierung," Ph.D. dissertation, Institut fr Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität, München, 2010.

[5] ——, "The CoPaSul intonation model," in *Elektronische Sprachverarbeitung 2011*, ser. Studientexte zur Sprachkommunikation, B. Kroeger and P. Birkholz, Eds. TUDpress, 2011, pp. 341–348.

[6] F. Schiel, "Automatic Phonetic Transcription of Non-Prompted Speech," in *Proc. ICPhS*, San Francisco, 1999, pp. 607–610.

[7] K. Schaefer-Vincent, "Pitch period detection and chaining: Method and evaluation," *Phonetica*, vol. 40, pp. 177–202, 1983.

[8] U. Reichel, "Improving Data Driven Part-of-Speech Tagging by Morphologic Knowledge Induction," in *Proc. AST Workshop*, Maribor, 2005, pp. 65–73.

[9] S. Abney, "Parsing By Chunks," in *Principle-Based Parsing*, R. Berwick, S. Abney, and C. Tenny, Eds. Dordrecht: Kluwer Academic Publishers, 1991, pp. 257–278.

[10] W. L. Chafe, "Givenness, contrastiveness, definiteness, subjects, topics, and point of view," in *Subject and topic*, C. Li, Ed. New York: Academic Press, 1976, pp. 25–55.

[11] E. Vallduví, "Information packaging: A survey," Tech. Rep. HCRC/RP-44, 1993.

[12] J. Hirschberg and J. Pierrehumbert, "The intonational structuring of discourse," in *Proc. 24th Annual Meeting, Association for Computational Linguistics*, New York, 1986, pp. 136–144.

[13] J. Hirschberg, D. Litman, J. Pierrehumbert, and G. Ward, "Intonation and the intentional structure of discourse," in *Proc. 10th international joint conference on Artificial intelligence*, Mailand, 1987, pp. 636–639.

[14] U. Reichel and K. Weilhammer, "Automated Morphological Segmentation and Evaluation," in *Proc. LREC*, Lisbon, Portugal, 2004, pp. 503–506.

[15] M. Hearst, "TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[16] ——, "Automatic acquisition of hyponyms from large text corpora," in *Proc. International Conference on Computational Linguistics*, vol. 2, Nantes, 1992, pp. 539–545.

[17] T. Dutoit, F. Bataille, V. Pagel, N. Pierret, and 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*, Philadelphia, 1996, pp. 1393–1396.

[18] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and Regression Trees*. Pacific Grove, CA.: Wadsworth & Brooks, 1984.