THE COPASUL INTONATION MODEL

Uwe D. Reichel

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

Abstract: A new data-driven and linguistically interpretable intonation model for the automatic analysis and synthesis of fundamental frequency contours is introduced: the CoPaSul model, which provides a contour-based (Co), parametric (Pa), and superpositional (Sul) intonation representation. Its application in F0 analysis and generation is described as well as its linguistic anchoring with respect to semantic weight and utterance finality.

1 Introduction

Established intonation models can roughly be divided with respect to three dichotomies:

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

Some examples: the tone sequence approach of Pierrehumbert [11] can be characterised as tone-based, symbolic and single-layered, since fundamental frequency (F0) within an intonation phrase is described as a single-layered sequence of tone symbols. The Fujisaki model [7] is to be characterised as contour-based, parametric and superpositional. F0 contours are considered 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 PaintE model [9] is to be described as contour-based, parametric and single-layered. It provides a parameterisation of the F0 contour in the scope of accented and phrase-final syllables. By means of parameter vector clustering the PaintE model additionally offers a symbolic F0 representation [10].

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 [16] or to generate F0 contours [2], parameters can be directly inferred from and transformed into F0 values.

Concerning the character of the chosen units, contours are inherently more related to parametric descriptions than tones. On the one hand, one might argue for a tone-based approach supported by the perceptual *tonal movement coding* hypothesis [8]. On the other hand, [5] demonstrated tone labels to be almost entirely determined by the preceeding tones concluding that not single tones but rather tone sequences and thus contours are the relevant intonation units.

The advantage of a superpositional arrangement lies in its capability to adequately model global phenomena like declination. Furthermore, it offers the possibility to account for findings of preplanning in intonation production [4], e.g. reflected by the relation between utterance length and declination slope.

The design of the CoPaSul model originally developed in [13] (where it is referred to as the PKS model) is based on these considerations and will be described in the following section.

2 The CoPaSul model

2.1 General description

The CoPaSul model provides a parametric, contour-based, and superpositional F0 representation. F0 contours are treated as a superposition of global and local components. These components are anchored in a hierarchic prosodic structure defined by global and local segments which roughly correspond to intonation phrases and accent groups respectively. The stylisation 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 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.

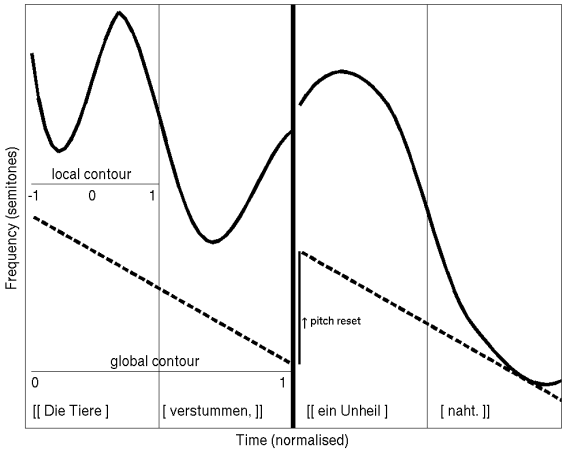

Figure 1 - CoPaSul F0 representation as a superposition of global and local intonation contour classes. See section 2.3 for further explanations.

2.2 Data and preprocessing

The data origins from the SI1000P corpus [15] containing 190 minutes of read German speech of a professional male speaker. F0 contours were extracted by the Schaefer-Vincent algorithm [14] 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 [13]. On the text level, part of speech tagging was carried out by a tagger developed in [12]. Signal and text were aligned by MAUS [15].

2.3 F0 analysis

2.3.1 Prosodic structure

Global intonation segments are delimited by speech pauses and punctuation. This segmentation was carried out automatically based on the preceeding signal-text alignment and the pause detection. 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 in [1] and ensures in most cases that each local segment maximally contains one accented syllable. 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*]].

2.3.2 F0 stylisation

All stylisations are based on the F0 values in frames of 110 ms centered on the detected syllable nuclei. This approach has the advantages, that (1) it relies on robust syllable nucleus detection, (2) there is no need for an exact syllable segmentation, and (3) no weighting of more and less important parts of the F0 is required. Global and local contour stylisation as described in the subsequent paragraphs is shown in Figure 2.

Global contours Within each global intonation segment, a declination baseline is derived as follows: for each syllable nucleus window the median of the F0 values below the 10th percentile is taken as an F0 base value. The baseline then is adjusted as a bottom tangent of the sequence of these base values [13]. This baseline is the 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-normalised residuum contour. 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. This approach requires separate normalisations of the pre- and post-accent parts of the local segment.

2.3.3 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 [13]. Figure 3 shows the centroids of the resulting global and local contour classes. In Figure 1 the intonation aspects related to the two global seg-

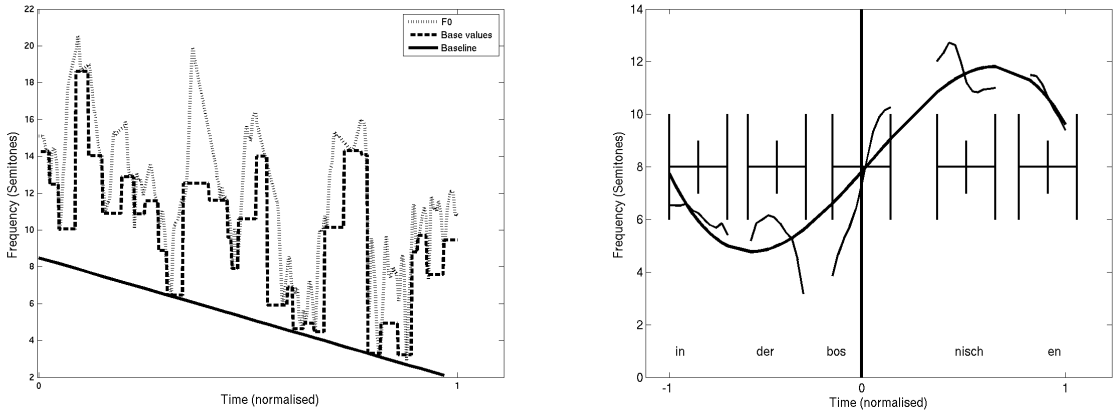

Figure 2 - **Left:** Linear global contour stylisation in form of a baseline in a global intonation segment. **Right:** Local contour stylisation in a local intonation segment by a third order polynomial; The stylisation is based only on the F0 values around the syllable nuclei as indicated by the plotted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

ments are represented by the global contour class sequence g_2, g_2, and the intonation aspects related to the four local segments by the local class sequence c_2, c_4, c_5, c_1.

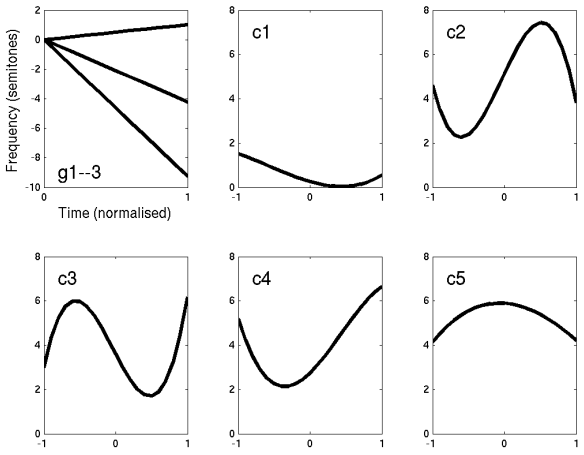

Figure 3 - Global (g_{1-3}) and local (c_{1-5}) contour classes.

2.3.4 Phonetic realisation models

Contour realisations The mapping of the "phonologic" level of abstract contour class centroids to the phonetic level of the intonation surface is carried out by linear regression models. By this means, the slope realisation of global contours is derived from the underlying centroid slope, the realised slope of the preceeding contour, and the length of the current global segment (given as the number of syllables). Each local contour coefficient realisation is predicted from its underlying centroid coefficient, the realised slope of the global contour, and the relative position of the local segment within the global segment.

Pitch reset At junctions between global segments as shown in Figure 1 pitch reset is modelled by a linear regression model using as predictors the realised slopes of the two global contours, the length of the interjacent pause and the F0 baseline value of the final syllable in front of the juncture. For all linear regressions, the predictors are range-normalised and orthogonalised by means of a principal component analysis.

2.4 F0 generation

For F0 generation an utterance is first segmented into global and local segments according to pause, punctuation and part of speech information. After having chosen appropriate contour classes they are mapped to their phonetic realisations by the linear regression models described in the preceeding section. For every global segment the pitch reset model determines the global contour F0 starting point. Finally global and local contour realisations are time-denormalised to the concrete utterance, superimposed by addition and transformed to Hertz values. Within a global segment shape-preserving piecewise cubic Hermite interpolation is carried out to smooth F0 discontinuities between adjacent local contours.

2.5 Results

Objective evaluation In a ten-fold cross validation the CoPaSul model was evaluated on held-out test data with respect to the root mean square error (RMSE) and the correlation between original and predicted contours. The RMSE represents the distance of model prediction from the original values, the correlation expresses the form similarity between the contours. A mean RMSE of 21.14 Hz, and mean correlation of 0.47 was achieved on the held out data. There was no significant performance difference between training and test data (Mann-Whitney, $p > 0.4$), altogether speaking for a good generalisation capability on an average performance level.

Perceptual evaluation Using TD-PSOLA as implemented in PRAAT resynthesised stimuli with original and modelled F0 contours were rated by 24 subjects (age between 22 and 47, German mother tongue, 19 females, students or researchers of Phonetics) with respect to their naturalness on a five-level scale. The mean naturalness rating amounted 3.12 for the modelled contours on the held out data, which is significantly below the mean judgment of the original contours 4.07 (Mann-Whitney, $p < 0.001$) but also significantly above the average 3 (one-sided sign test, $p < 0.05$).

3 Linguistic interpretation

The linguistic anchoring of the CoPaSul model was examined the following way: by automatic linguistic corpus analyses hypotheses about possible relations between local contour classes and linguistic concepts were generated. These hypotheses were subsequently tested in perception experiments.

3.1 Semantic weight

3.1.1 Modelling

The concept of semantic weight is based on a definition given by [3] as the *predictability* of a word, which can be expressed in quantitative terms. In this study it was measured in form of

linear interpolated trigram probabilities of the final content words of the local segments. Unigram probabilities provide an approximation of their context-independent predictability while bi- and trigram probabilities reflect their context-dependent predictability.

3.1.2 Corpus statistics

The class-dependent predictabilities are presented as boxplots in the left panel of Figure 4. A one-factor analysis of variance revealed significant differences in the word predictabilities with respect to the factor *local intonation class* ($F[4,9214] = 31.7, p < 0.001$). By post hoc comparison class c_2 was identified to contain content words of significantly lower probability and thus higher semantic weight than the other classes (Turkey-Kramer, $\alpha = 0.001$). At the opposite end, words realised by class c_1 are of significantly lower semantic weight than words of the other classes ($\alpha = 0.005$). From these results two hypotheses are formulated:

H1 class c_1 encodes low semantic weight

H2 class c_2 encodes high semantic weight

3.1.3 Perceptual validation

Subjects and Method H1 and H2 were subsequently tested by a perception experiment, the same 24 subjects as described in section 2.5 took part in. Single local segment utterances were resynthesised with the five local intonation contour classes as well as with 5 distractor contours by MBROLA [6] based on a German diphone database and presented via headphones. Segment durations were calculated by multiplying their intrinsic (mean) durations by a factor predicted by a regression tree model (see [13] for details). All utterances were derived from the template *Das ist eine X (This is an X)*. The 60 target words X were amongst others controlled for uniform syllable number and structure, word frequency, and voicing. The subjects had to rate the importance of the utterance as signaled by the speaker on a 5-point Likert scale.

Results As can be seen in the middle panel in Figure 4 class c_2 was perceptually attributed to significantly higher and c_1 to significantly lower semantic weight than the other classes (Kruskal-Wallis, $\chi^2_4 = 515.36, p < 0.001$; Dunnett post hoc, $\alpha = 0.01$). Thus, hypotheses $H1$and $H2$ were perceptually confirmed.

3.2 Utterance finality

3.2.1 Corpus statistics

Since the reader produced each sentence in isolation, the last local segment of each sentence was classified as utterance final, and the others as non-final. Testing the significance of co-occurrences of contour classes and utterance finality resulted in two hypotheses ($\chi^2 > 17.5, \alpha = 0.001$):

H3 c_1 and c_5 encode finality

H4 c_2, c_3, and c_4 encode non-finality

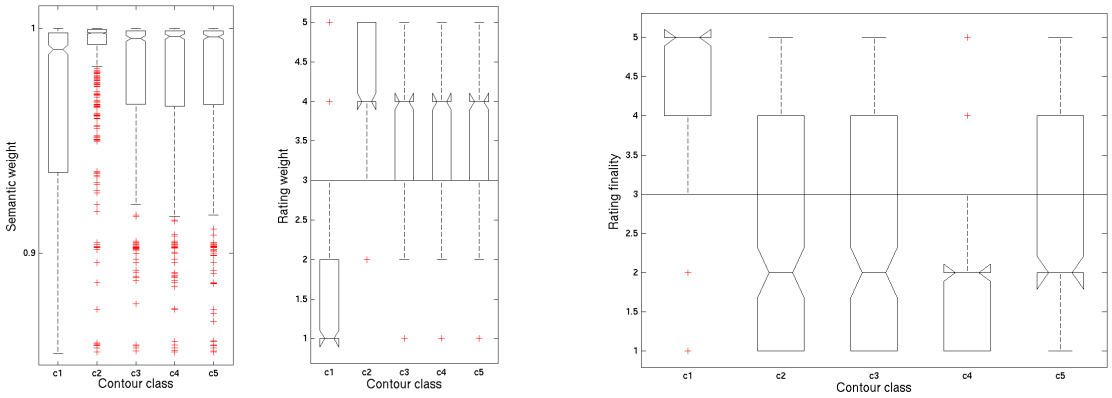

Figure 4 - **Left:** Corpus statistic results: content word trigram probabilities for the 5 local contour classes. **Mid:** Perceptual validation: Class-dependent weight ratings. **Right:** Perceptual finality judgments depending on local contour classes.

3.2.2 Perceptual validation

Subjects and Method Again the same 24 subjects as described in section 2.5 took part in this experiment. The stimuli were resynthesised as specified in section 3.1.3. This time the template sentence was: *Eine X (an X). X* was taken from the same target word pool as above. 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 thus is characterised by an utterance-final intonation. Allocating it to *'Eine X und eine Y'* implies, that subjects expect more to come triggered by a progredient intonation contour.

Results As can be seen on the right hand side in Figure 4, finality judgments differed significantly for classes c_1 (final) and c_4 (non-final; Kruskal-Wallis, $\chi^2_4 = 316.92, p < 0.001$; Dunnett post hoc, $\alpha = 0.01$). All classes were perceptually classified with respect to finality since all judgments differed significantly from the mean 3 representing "undecided" (one-sided sign tests, $\alpha = 0.05$, Bonferroni-corrected, $|z| > 3.40, p < 0.001$).

$H4$ was confirmed: c_2, c_3, and c_4 pere perceived as progredient and thus encode non-finality. $H3$ was just partly confirmed: while c_1 was perceived as utterance final, c_5 was rather perceived as non-final, however to a lesser degree than the other classes.

4 Discussion

4.1 The CoPaSul model

An essential demand for designing the CoPaSul model has been to keep the required training corpus preparations as low as possible. All preprocessing steps have been automatised in this study. This allows for a fast adaptation of the model to other speech data and avoids inter-labeler inconsistencies.

Polynomial stylisation even of low order is capable to encode prominence and progredience, and guarantees as an analytic approximation method a biunique mapping between signal and

stylisation which is considered to be crucial for a subsequent interpretable parameter clustering. Up to now, the model was developed on the basis of read speech of one professional speaker. By this choice widely acceptable intonation contours are guaranteed, but nevertheless, the next step will be to confront this model with spontaneous speech data.

4.2 Linguistic anchoring

Based on corpus statistics hypotheses about the linguistic functions of local contour classes were derived and subsequently tested by perception experiments. With one exception the hypotheses have been confirmed. By this linguistic anchoring it has been shown that the CoPaSul model is appropriate to generate an intonation representation which can be derived from the signal as well as from text, which could make this model interesting for intonation analysis and synthesis in fundamental research and speech technology.

References

[1] ABNEY, S.: *Parsing By Chunks*. In BERWICK, R., S. ABNEY and C. TENNY (eds.): *Principle-Based Parsing*, pp. 257–278. Kluwer Academic Publishers, Dordrecht, 1991.

[2] BLACK, A. and A. HUNT: *Generating F0 contours from ToBI labels using linear regression*. In *Proc. ICSLP*, vol. 3, pp. 1385–1388, Philadelphia, 1996.

[3] BOLINGER, D.: *Accent is predictable (if you're a mind reader)*. Language, 48:633–644, 1972.

[4] COOPER, W. and J. SORENSEN: *Fundamental frequency in sentence production*. Springer, New York, 1981.

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

[6] DUTOIT, T., 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*, pp. 1393–1396, Philadelphia, 1996.

[7] FUJISAKI, H.: *A note on physiological and physical basis for the phrase and the accent components in the voice fundamental frequency contour*. In FUJIMURA, O. (ed.): *Vocal physiology: voice production, mechanisms, and functions*, pp. 165–175. Raven, New York, 1987.

[8] HOUSE, D.: *Tonal Perception in Speech*. Lund University Press, Lund, 1990.

[9] MÖHLER, G.: *Describing intonation with a parametric model*. In *Proc. ICSLP*, pp. 2851–2854, Sydney, 1998.

[10] MÖHLER, G. and A. CONKIE: *Parametric modeling of intonation using vector quantization*. In *Proc. 3rd ESCA Workshop on Speech Synthesis*, pp. 311–316, 1998.

[11] PIERREHUMBERT, J.: *The phonology and phonetics of Englisch intonation*. PhD thesis, MIT, Cambridge, MA, 1980.

[12] REICHEL, U.: *Improving Data Driven Part-of-Speech Tagging by Morphologic Knowledge Induction*. In *Proc. AST Workshop*, pp. 65–73, Maribor, 2005.

[13] REICHEL, U.: *Datenbasierte und linguistisch interpretierbare Intonationsmodellierung*. PhD thesis, Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität, München, 2010.

[14] SCHAEFER-VINCENT, K.: *Pitch period detection and chaining: Method and evaluation*. Phonetica, 40:177–202, 1983.

[15] SCHIEL, F.: *Automatic Phonetic Transcription of Non-Prompted Speech*. In *Proc. ICPhS*, pp. 607–610, San Francisco, 1999.

[16] SCHWEITZER, A. and B. MÖBIUS: *Experiments in Automatic Prosodic Labeling*. In *Proc. Eurospeech*, pp. 2515–2518, Brighton, 2009.