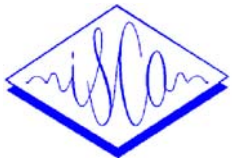

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

A SYSTEM OF STYLIZED INTONATION CONTOURS IN GERMAN

Hannes PIRKER, Kai ALTER+, Erhard RANK, John MATIASEK, Harald TROST and Gernot KUBIN++

Austrian Research Institute for Artificial Intelligence (ÖFAI), Schottengasse 3, A–1010 Vienna, Austria.
Email: {hannes|erhard|john|harald}@ai.univie.ac.at
+Max-Planck-Institute of Cognitive Neuroscience, Inselstraße 22-26, D–04103 Leipzig, Germany.
Email: alter@cns.mpg.de
++Institute of Communications and High-Frequency Engineering, Vienna University of Technology,
Gußhausstraße 25/E389, A–1040 Vienna, Austria. Email: g.kubin@ieee.org

ABSTRACT

Modeling intonation, i.e., specifying adequate fundamental frequency (F0) contours, remains a challenging task for speech synthesis systems. This paper discusses the development of a system for phonetically specifying intonation contours for German. It deals with the problem of translating an abstract phonological representation of intonation - namely the tone-sequence model - into a concrete phonetic model. Design options and evaluation methods are discussed.

10.21437/Eurospeech.1997-125

1. PHONOLOGICAL SPECIFICATION

The study of phonological aspects of intonation has strongly been influenced by the tone-sequence model of Pierrehumbert [12]. As a consequence, the ToBI (Tone and Break Indices, [3]) notation system has become a generally accepted means to describe intonation structure *phonologically*.
In this autosegmental model intonation contours are modelled by specifying a sequence of high (H) and low (L) target tones.
In [13] a ToBI offspring for the description of German intonation is presented. Neglecting details like phrase accents and downstepping for now, this system comprises the following inventory of tones:

H*	Normal peak accent
L+H*	Steep rise within the accented syllable
L*+H	Peak is moved behind the accented syllable
L*	Valley accent
H+L*	Falling accent
L%	Low boundary tone
H%	High boundary tone

2. PHONETIC MODELS

However, in order to use these phonological, i.e., rather abstract descriptions of intonational contours in speech synthesis they still need to be interpreted *phonetically*. The realization of the abstract tonal markers has to be defined with respect to the time and frequency domain. For an accent labelled L+H* (steep rise) for instance, it has to be specified where the F0 excursion starts, where the peak is located in alignment to the segmental string, and how high the pitch peak is. In other words: abstract tones have to be transformed into the acoustical parameters *Hz* and *ms*.

The study of intonation is hampered by the strong variability observed in naturally occurring pitch contours. Speaker dependent variations as well as purely accidental fluctuations have to be considered. Specifying a mapping from abstract phonological features to acoustic parameters not only is a necessity for speech synthesis but may also be used in order to evaluate the soundness of the phonological labelling. This could be accomplished by comparing resynthesized samples using both the original contours and those produced on the basis of the phonological labels - which should be perceived as functionally equivalent.
For German we are aware of phonetically influenced intonation systems that employ theories such as Fujisaki's model (e.g., Möbius' work [10]) as well as data driven approaches like [15].

3. STYLIZED CONTOURS

Our system is particularily influenced by approaches that use so called *stylized contours*. In this context stylization is defined as the replacement of naturally occurring F0 contours with standardized stretches of pitch movements that remain functionally equivalent. Stylization schemes have been proposed for German before which have to be taken into account. Of special interest are the IPO model presented in by Adriaens [1] and the *Kieler Intonation Model* (KIM) devoloped by Kohler [9].
Obviously, a stylized stretch of a pitch contour can be specified as shown in fig.(1) - i.e. by supplying its length, pitch range and temporal alignment with respect to some reference point.

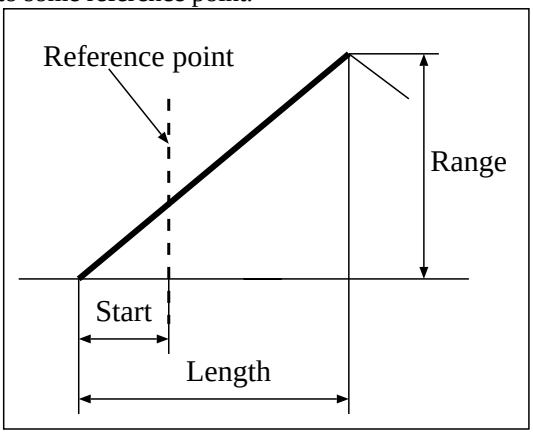

Fig.(1):Parameters used for specifying parts of a contour.

Of course this is formally equivalent to specifying the position of subsequent target points that are connected by linear interpolation. Such a conceptualization (see also [4] and [11]) also seems to be more closely in line with the "spirit" of the tone sequence model because tonal target points instead of stretches of rising and falling contours are conceived as the basic units. Whatever style of conceptualization is chosen, the overall problem remains the same: time alignment and frequency of turning points within the pitch contour have to be specified.

4. INFLUENCING FACTORS

A number of factors related to different levels (see [7] for a summary) may influence the actual form of pitch contours

Segmental context

These influences are often subsumed under the term *microprosody*. Apart from the minor perturbations observable in every naturally occurring F0 -which seem to be neglectable for the moment being - more systematic influences of segmental context onto the pitch frequency can be observed. For instance, Kohler [9] identified a difference between intrinsic F0 values in high and low vowels, as well as influences of prevocalic consonants to the vowel's pitch.
An overview on segmental influences on the timing domain is given in [7]. Simply speaking, the timing of pitch accents depends on the amount of voiced "material" that is available for carrying the pitch information. An accent bearing syllable with voiced on- and offset simply leaves more time for the realization of the pitch movement, peaks are thus positioned later in the syllable than in the context of unvoiced segments.

Prosodic context

Prosodic context may strongly influence the pitch contour. The type and distance of prosodic boundaries, rhythmical organisation, tonal environment, global and local pitch range etc. obviously are of great importance.

Speaking rate

Changes in the speaking rate have to be reflected in the contours as well.

5. PARAMETRIZATION

At the moment we employ the following parametrization for our phonetic model (see Fig.1).

Range

For the specification of pitch ranges a semitone (ST) scale instead of absolute frequency values in *Hz* is used. Thus pitch ranges can "automatically" accommodate if the overall pitch register is changed. Pitch movements usually are in the range of 5 to 8 semitones for accentual patterns.

Reference Point

In order to specify the positioning of a tonal configuration within the time domain a number of possible reference points have been discussed that are usually tied to the accent bearing syllable, e.g., syllable onset, voiced onset, P-centre, end of the vowel or voiced offsets.
We use the onset of the accented syllable's nucleus as general reference point. Starting points are set in a fixed distance to this onset. In the IPO system the different peak-accents are mainly distinguished by different starting points as
shown in fig.(2).

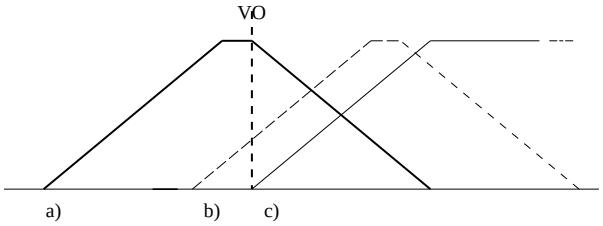

Fig.(2) Temporal alignment of stylized contours in relation to the vowel onset VO (vertical line)

Length

For the specification of the length of rises and falls that are used to encode accentuation also fixed values are employed at the moment (with the exception of some intermediate stretches of variable length that merely span the interval between subsequent accent lending pitch movements).

Declination

The observation that pitch contours generally display a down-drifting tendency throughout an utterance (or intonation phrase) is rather uncontroversially accepted. Nevertheless, this may be modelled in different ways. We follow the simple IPO approach by superimposing a fixed amount of declination by default. For example - with the exception of very short utterances - the height of the intonation baseline uniformly decreases by 5 semi-tones.

6. DEVELOPMENT AND EVALUATION

The evaluation and subsequent refinement of our model is based on several methods.

Achieving close-copy contours

First, neglecting the matching procedure from abstract tones to phonetic parameters, the validity of the phonetic parameters is tested in isolation. It is investigated how closely stylised contours can match naturally occurring intonation contours perceptionally by adapting only global parameters such as declination rate and pitch range to the original utterance.
Both read and spontaneous utterances are evaluated this way. This task is performed by resynthesizing the utterances with both original duration and F0 values as well as with stylized contours that are based on the parameters of our model. The obtained contours are also compared both perceptually and visually with the original F0 contour.

Controlled modifications

Another line of experiments mainly deals with the study of segmental influences onto the pitch contour. For instance, experiments in the style of [9] are performed in order to quantify these influences. By synthesizing an utterance with a number of systematically changed intonation contours a test set is produced. Test subjects

have to rate randomly selected pairs from this set with respect to their perceptual identity or their perceived relative prominence etc. In another task the test-subject may interactively alter a single parameter (e.g. pitch height) of a speech sample until it is rated perceptually identical to a given target utterance. This line of tasks also is performed with utterances that differ in their segmental string (e.g. comparing high and low vowels in order to verify and quantify the influence of so-called intrinsic pitch).
Nevertheless, results of these fine grained experiments have to be taken with care. For example, in our experiments test subjects were able to perceptionally distinguish minor variations in the pitch height when comparing samples directly.
One should not automatically conclude though that such distinctions have to be integrated into the stylisation model as they do not necessarily contribute to an increase in global naturalness.
We are more confident about the usefulness of further investigations in the domain of segmentally triggered alignment modifications. At the moment we aim at the development of a model in the style of [14] that takes for instance the voicing characteristics of syllables into account in order to specify peak positions.
Also, the possible interaction of pitch with duration and intensity surely becomes a problem when performing experiments that deal with prosodic features on a very fine grained level. When comparing the "prominence" different segmental strings, variations in duration or even intensity might be exploited by the test subject if no clear intonational cues is present.

Functional rating

As complete perceptional equivalence in general cannot be obtained by a stylisation model, the evaluation has to rely on some notion of functional similarity. Nevertheless, the exact definition of this measure remains problematic. As the 'naive' listener usually lacks a functional categorization of prosodic features it is difficult to assign clear-cut functional tagging to different tonal configuration. This works quite well in the case of rising tones that are used in order to signalize questions in German. These contours are systematically applied to different sentences and subjects judge whether the obtained utterances are perceived as questions (as a matter of fact the modelling of question intonation within the IPO system soon turned out to be inadequate in some contexts). On the other hand [6] and [4] indicate that subjects are especially sensitive to differences in prepausal contours. Thus, though it is more difficult to obtain perceptive equivalence in these contexts, the functional rating seems to be easier to perform. Another rough functional category is established by Kohler ([9]) who claims that falling contours on accented syllables ("early peak" in his terminology) signal a functional category of establishment/finality. Nevertheless, it is a problem to define more fine grained functional categories. Rating placement and relative prominence of accents may be performed, but more detailed categorizations, often including attentional features such as "doubt" or "conspiciousness" are rather difficult to establish.

7. INTERFACING

In the course of integrating a model for phonetically interpreting a phonological description into a speech synthesis framework the specification of the appropriate interface it not a trivial task.
On the one hand it has to be decided which prosodic variations may be viewed as totally phonetically triggered (e.g. merely segmental influences on the time alignment) and which require access to other sources of information. The list of potential influencing factors (in section 4) shows that the scope of the ToBI annotation cannot suffice in order to fully specify the prosodic features of an utterance. E.g., prominence levels are not directly encoded within the ToBI annotation but have to be supplied separately. Changes in the overall register, declination resets etc. have to be encoded as well. Supplementing the bare ToBI transcriptions with a representation of tone scaling features (see e.g. [8]) surely is necessary.
Being embedded into a Concept-to-Speech system [2] the scope of the "phonetic interpreter" has to be carefully defined. For example, hat contours sometimes are viewed as "allotonic variants" of two subsequent pitch accents in order to avoid a tone-clash. It has to be defined whether this tone-clash resolution should be performed on the phonological level already, or be subject to the phonetic module. The same holds true for the integration of influences from higher levels such as information-structure. House [7], e.g., reports that topic-initiality appears to delay F0 peaks. Should this be included into a extended phonological description (e.g. introducing a new category "delayed") or should the phonetic interpreter itself make use of all upper-level information?
In the long run, the interplay within the component that specifies durational properties also has to be considered, as duration and pitch may influence each other, i.e., it is not a priori clear that the phonetic interpreation of tonal specifications always may rely on a preceeding complete specification of segmental durations.

8. IMPLEMENTATION

At the moment the technical setup in which our study is performed consists of freely available software only. F0 analysis, labelling of speech data and TD-PSOLA resynthesis are undertaken with the powerful SFS system from UCL. Also the MBROLA speech synthesizer [5] is used for producing speech samples. A number of *perl*-scripts are used in order to perform "prosody transplantation" from natural speech to synthesized samples. A testing framework has been implemented that comprises the production of sets of test tokens with systematically varying prosodic parameters, their randomized presentation to the test-subject as well as the logging of the results. Also a version, that allows for the interactive modification of parameters is available.

9. CONCLUSION

The stylization system proposed in this paper can be viewed as a combined approach in the tradition of both

the KIM and the IPO model hopefully integrating the best of both worlds. As in the KIM model, we aim at a more detailed investigation of the role of segmental context. Although being more refined in this respect, the KIM's overall coverage is somehow smaller than that of the IPO model. Most of the phonetic studies documented in [9] have been performed with utterances that contained one accent in utterance final position only. The existence of an ontology of three different peak forms was assumed first. Under this theoretical assumption prototypical samples for the three peak forms where produced by a trained speaker and phonetically analysed afterwards. This is to say that the KIM model was especially designed to sustain a not so widely accepted intonational theory. Our handling of utterances with several accents and the implementation of declination is inspired by the IPO model. At the same time some flaws of the IPO model had to be eliminated, e.g., a modelling of question contours that was not fully satisfying . An important feature of our system is its connection to a widely accepted phonological theory, namely the tone sequence model. It is thus aligned with ongoing linguistic research on German intonation. The use of a stylization model for interpreting phonological descriptions makes the ToBI system applicable to speech synthesis. It also allows for an evaluation and verification of the the labelling of speech corpora with tonal information.

ACKNOWLEDGEMENTS

The work reported here has been carried out within the project „Phonology-Acoustics-Conversion for Concept-to-Speech“ (FWF P10822) funded by the Austrian *Fonds zur Förderung der wissenschaftlichen Forschung.* Financial support for the Austrian Research Institute for Artificial Intelligence is provided by the Austrian *Federal Ministry of Science and Transport.*

REFERENCES

[1] Adriaens L.M.H.: Ein Modell deutscher Intonation, Dissertation, Technische Universiteit Eindhooven, 1991.

[2] Alter K., Buchberger E., Matiasek J., Niklfeld G., Trost H.: VIECTOS: The Vienna Concept-to-Speech System, in Gibbon D.(ed.), Natural Language Processing and Speech Technology, Mouton de Gruyter, Berlin, 1996.

[3] Beckman M.E., Ayers G.M.: Guidelines for ToBI Labelling, Ohio, 1994.

[4] Bruce G., Granström B., Gustafson K., Horne M., House D., Lastow B., Touati P.: Speech Synthesis in Spoken Dialogue Research, in Proceedings of the 4th European Conference on Speech Communication and Technology, Madrid, Spain, Vol.2,pp.1169-73, 1995.

[5] Dutoit T.: An Introduction to Text-to-Speech Synthesis, Kluwer Academic Publishers, Boston/Dordrecht/London, Text, Speech and Language Technology, Vol. 3, 1997.

[6] Hermes D.J.: Measuring the Perceptual Similarity of Pitch Contours, in Proceedings of the 4th European Conference on Speech Communication and Technology, Madrid, Spain, Vol.III,pp.2051-54, 1995.

[7] House J., Wichmann A.: Investigating Peak Timing in Naturally-Occurring Speech: From Segmental Constraints to Discourse Structure, in Hazan V., et al.(eds.), Speech Hearing and Language: Work in Progress, Department of Phonetics and Linguistics, University College London, 1996.

[8] Kerkhoff J., Rietveld T.: The Generation of Prosody in the Nijmegen Rule Oriented Speech Synthesis System, in Proceedings of the 4th European Conference on Speech Communication and Technology, Madrid, Spain, Vol.III, pp.1831-34, 1995.

[9] Kohler K.(ed.): Arbeitsberichte Nr.25, Institut für Phonetik und Digitale Sprachverarbeitung, Universität Kiel, 1991.

[10] Möbius B., Pätzold M., Hess W.: Analysis and synthesis of German F0 contours by means of Fujisaki's model, in Speech-Communication, North-Holland, (13), pp. 53-61, 1993.

[11] Möhler G.: Rule Based Generation of Fundamental Frequency Contours for German Utterances, Proc. of 2nd SPEAK! Workshop, Darmstadt, Germany, 1995.

[12] Pierrehumbert J.B.: The Phonology and Phonetics of English Intonation, Ph.D. Thesis, MIT, Cambridge, 1980.

[13] Reyelt M., Grice M., Benzmüller R., Mayer J., Batliner A.: Prosodische Etikettierung des Deutschen mit ToBI, in Gibbon D.(ed.), Natural Language Processing and Speech Technology, Mouton de Gruyter, Berlin, p.144-155, 1996.

[14] Rietveld T., Gussenhoven C.: Aligning Pitch Targets in Speech Synthesis: Effects of Syllable Structure, Journal of Phonetics, 23, 375-385, 1995.

[15] Traber C.: F0 generation with a database of natural F0 patterns and with a neural network, in Bailly G. & Benoit C.(eds.), Talking Machines, Amsterdam, New York, Oxford: North-Holland, pp.287-304, 1992.