PARAMETRIC DESCRIPTION OF F_0-CONTOURS IN A PROSODIC DATABASE

B. Heuft; T. Portele; F. Höfer; J. Krämer; H. Meyer; M. Rauth; G. Sonntag
Institut für Kommunikationsforschung und Phonetik, University of Bonn, Germany
Poppelsdorfer Allee 47, 53115 Bonn

ABSTRACT

A maximum-based model for parametrization of F_0-contours was developed. A prosodic database is described, which is used for a rule-driven as well as a data-driven approach to synthetic prosody. For each syllable, it contains perceptive, acoustic, and linguistic information.

MOTIVATION

For an investigation of intonation it is necessary to describe a given contour with as few parameters as possible, i.e. only the relevant information should be kept. The parameters must be automatically computable. On the other hand, it should be no problem to generate any possible F_0-contour with these parameters from an adequate input. So there is a double need: First, a model for parametrization has to be developed, and second, a database has to be constructed that contains all information that is supposed to have an influence on the model parameters.

THE MODEL

When developing this model, we assumed the F_0-maxima to be the perceptually most important points of the F_0-contour. This hypothesis was motivated by the experiences made in Bonn using the parametrization with Fujisaki's model [1], which does not predict the position of the F_0-peaks exactly. Further problems exist for the modeling of the utterance-final F_0-decrease. German stress can be signalled by a falling F_0, so quite a lot of unintended stress-shift occured.

Our new method of parametrization is called maximum based description. Each F_0-contour is parametrised describing only its maxima: for each maximum, four parameters are given.

First, the maximum is located precisely in time, relative to the onset of the accented vowel assigned to it. This distance is called *delay*, it is negative when the peak preceeds the vowel onset, positive when it follows.

The second parameter, the height of the maximum (*amplitude*), is described as a percentage value between a top- and a baseline. To simplify the description, these lines are currently kept constant for a given speaker. At a later stage, these lines may be used to modify easily the F_0-range.

The third and the fourth parameter describe the steepness of the contours preceding (*left slope*) and following (*right slope*) the maximum. These slopes are interpreted as sinoidal slopes with different degrees of damping. Fig. 1 explains the four model parameters with a stylized F_0-contour.

Minima are not described explicitly, but they are the crossing points of the contours preceeding and following the maxima. This method of description does not make any presumption about the functions of F_0-maxima; for example, no boundary tones are labelled, maxima at the end of questions are not treated differently to maxima caused by focus.

TEXT CORPUS

First, the corpus consists of three short texts, each about 400 words long. Then, there are 50 wh-questions and yes/no-questions each, with their pertinent answers. Some of the questions are segmentally identical but focus different words. There is further a couple of instructions and categorical questions, as well as hundred simple structured, isolated one-phrase utterances. The corpus was read by three speakers, two female and one male, among them the two speakers of the unit inventories for the speech synthesis system developed in Bonn [8]. The recording was done in the anechoic chamber of the IKP.

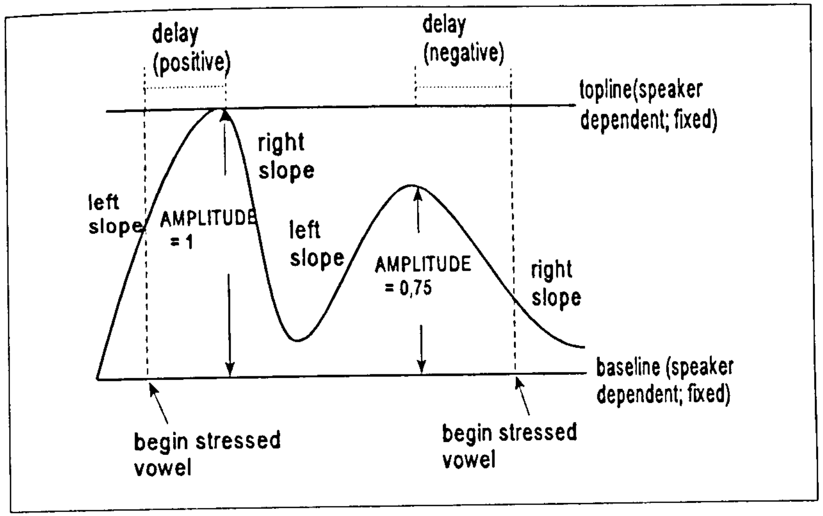

Figure1: The model parameters

ACOUSTIC INFORMATION

Model parameters: An algorithm was developed, that automatically extracts the model parameters mentioned above from a raw F_0-slope [2]. As additional input it needs information about phrase boundaries and maxima locations and the onset time of the accented vowel. The algorithm first determines the *delay* from the maximum position, second the *amplitude* relative to top- and baseline. Finally, the F_0-slope left and right of each maxima are approximated as $\cos^2$ functions. The optimization uses the least square error, which is increased next to the maxima, thus ascribing them a bigger importance.

In most cases, no auditive differences appeared between the original and the parametrized and resynthesized contours; the functional aspects of a contour certainly remained unchanged. A quantitative analysis showed a high correlation between original and parametric contours (see fig.2).

In the database, the model parameters are labelled only for syllables associated with an F_0-maximum. A very important and also difficult task in the future will be to determine these syllables using only the text-input into a synthesis-system.

Duration: The database was segmented automatically [3]. Syllable duration was determined using the segmentation output. Moreover, the duration of the coda (i.e. the sounds following the syllable nucleus), the syllable onset (i.e. the sounds preceding the syllable nucleus) and the duration of the nucleus itself are given.

Boundary-types: Two types of prosodic boundaries are distinguished: progredient and non-progredient. The distinction was made using exclusively the slope of the F_0 at the previously perceptively determined boundaries (see below). Rising contours are labelled as progredient, falling contours are labelled as non-progredient.

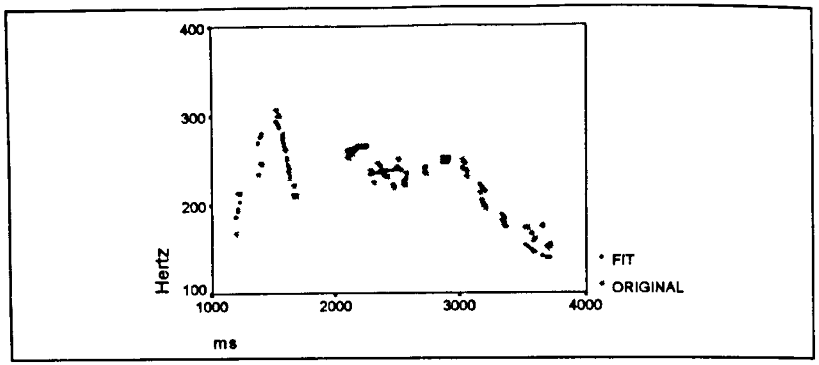

Figure 2: Original and parametrized F_0-contours of the randomly chosen utterance:" Die Schüssel mit Äpfeln haben wir auf dem Küchentisch gedeckt"; female speaker

In the database, all information is given in respect to syllables: each syllable contains information about its distance to the following boundary and about type and strength of this boundary.

PERCEPTIVE INFORMATION

Syllable-prominence: The perceptive prominence of each syllable was rated by three subjects. The rating was done on a graphical scale ranging from 0 to 31. The adequacy of this method has been proved by Fant & Kruckenberg [4]. The subjects could listen to an utterance as often as they needed to, and they could determine the size of the presentation frame. We found a high inter-subject correlation ($\rho = 7,9$).

Place of boundaries: The boundaries were set by two labellers experienced in prosody work. They consulted each other in problematic cases, but most of the labelling was done by one person alone. Both the auditive and the instrumental analysis (graphical presentation of F_0-slope and waveform) were used. In cases of doubt, the auditive impression was of prime importance.

LINGUISTIC INFORMATION

Linguistic accentability: The syllables with a nucleus consisting of a schwa-sound or a vocalic /r/ as well as the German suffix /IC/ were labelled as unaccentable, all other syllables were labelled accentable.

Syllable structure: The number of sounds in each syllable as well as the number of sounds in syllable onset and coda is determined. Additionally, information about the type of articulation of the sounds in onset, nucleus and coda is given.

Boundary prominence: For both the rising and the falling boundary type, four degrees of boundary prominence are distinguished. The distinction is made after the perceptual determination of the boundary place. The lowest level was labelled when a prosodic phrase occured within a sentence. The second level describes boundaries after rising or falling wh-questions. The third level was labelled at the end of a declaration or an decision question. The highest level of prominence is labelled at the end of a text or an isolated sentence or an echo question.

CONTEXT INFORMATION

As mentioned earlier, the information about each syllables distance to a following boundary is given. Furthermore, the distance to the preceding and the following F_0-maxima are labelled (in syllables and in milliseconds) as well as the number of voiced sounds following the syllable-nucleus up to the next nucleus.

CURRENT APPLICATIONS

The database is used in two ways: in a data-driven system and to develop rules for duration and intonation. The data-driven system (a neural network) generates the model parameters and the syllable duration in one step [5]. As input it uses most of the parameters represented in the database.

In the rule-based prosody systems, duration and F_0 are generated in two steps. The duration is generated with a syllable-based model. A model to predict syllable durations was worked out using the new database [6]. It uses information about utterance-finality, number of sounds within a syllable, perceptive prominence and syllable structure. This duration control is already implemented in the Bonn speech synthesis system.

As for F_0, a preliminary model has been implemented as well. It predicts mean values for the model parameters *delay*, *amplitude* and *slope* using information about boundary distance, the type of boundary, the place of the syllable within a given utterance and three degrees of syllable prominence. Although this F_0-generation is very simple, its output is acceptable. This is at least a hint to the apropriateness of our description.

The database is further used to investigate more extensively the relations between perceptive and acoustic prominence [7].

CONCLUSION

A database wich includes perceptual acoustic and linguistic information proved to be a valuable research tool. A new method of parametrization shows good results for resynthesis of natural speech. If it is really adequate for the generation of prosody will have to be proved by future research.

REFERENCES

[1]Möbius, B. (1993): *Ein quantitatives Modell der deutschen Intonation - Analyse und Synthese von Grundfrequenzverläufen.* Tübingen: Niemeyer

[2]Portele, Th.; Krämer, J.; Heuft, B.; Sonntag, G.(1995): Parametrisierung von Grundfrequenzkonturen. *Fortschritte der Akustik-DAGA'95*, Bad Honnef

[3] Wesenick, M.B.; Schiel, F. (1994): Applying Speech Verification to a Large Data Base of German to obtain a Statistical Survey about Rules of Pronunciation. *Proc. ICSLP'94* pp 279-282

[4] Fant, G. & Kruckenberg, A. (1989): Preliminaries to the study of Swedish prose reading and reading style. STL-QPSR 2/1989, pp. 42-45

[5] Portele, T.; Reuter, A.; Heuft, B. (1995): Generating synthetic prosody with a neural network. Submitted to: Eurospeech'95

[6] Meyer, H.; Portele, T.; Heuft, B. (1995): Ein Silbendauermodell für die Sprachsynthese. Fortschritte der Akustik, DAGA'95, Bad Honnef.

[7] Heuft, B.; Portele, T.; Höfer, F.; Meyer, H.; Rauth, M. (1995): Betonungsstufen von Silben und ihre Beziehung zum Sprachsignal. *Fortschritte der Akustik-DAGA'95*, Bad Honnef.

[8] Portele, T.; Heuft, B.; Höfer, F.; Meyer, H.; Horst, W.(1994): A New High Quality Speech Synthesis System for German. *Progress and Prospects of Speech Research and Technology*, München