AN INTONATION MODEL FOR A GERMAN TEXT-TO-SPEECH´SYSTEM

K Kohler -

Universität Kiel, Institut für Phonetik und digitale Sprachverarbeitung, Kiel, West Germany

1. INTRODUCTION

That appropriate F0 patterns are essential for intelligible and natural-sounding synthetic speech has become common knowledge in speech technology research. What is less commonly realised is the fact that not only the semantically and pragmatically determined macro patterns are crucial but that the segmentally induced micro variations are equally important. Before F0 rules can be implemented in a text-to-speech system of a particular language a linguistically and phonetically motivated intonation model has to be developed which incorporates general assumptions about the generation of pitch in human language and specific descriptive facts about the individual language concerned. Such a model has to integrate the macro and micro F0 aspects. No model described in the literature fulfils this demand.

In order to be an adequate and economic representation of the production and perception of intonation contours in the speech communication process the model cannot work on the basis of an inventory of tunes or melodic elements, as is the case in the Dutch intonation models of various languages (e.g. de Pijper 1983, Adriaens 1984). Such a taxonomic approach misses important generalisations and has to change the model whenever new elements are needed to cope with empirical findings under a variety of conditions (speech rate, semantic, pragmatic and expressive aspects). Intonation models should therefore be developed within a generative framework (e.g. Gårding 1979, Pierrehumbert 1980), which postulates underlying elements to which (partly) ordered rules are applied at a number of different levels to yield an empirically testable F0 output. The phonological, syntactic, semantic, pragmatic and expressive functions of F0 are then not linked to the output contours directly, but to the underlying elements and to the various levels of the rule system. Generative models differ as to the degree of abstraction and to the number and type of levels introduced into the rule system.

The model to be presented here has been developed in the Kiel Phonetics Institute within a research project financed by the German Research Council (DFG, Ko 331/19-1,2,3). It meets the two

requirements set out above by being generative and by incorporating micro-prosodic adjustments. It took its point of departure from the analysis and systematically manipulated resynthesis of natural productions of single-peak terminal utterances (see Kohler 1986, Kohler 1988), but it has since been extended to multi-peak sentences as well as to rising patterns. The model has also been implemented in the INFOVOX text-to-speech-system for German (Carlson & Granström 1976, Carlson, Granström & Hunnicutt 1982). There has been a constant and efficient interaction between the model development and its soft-ware implementation, the former making the latter possible and the latter providing new insight into the intonation structures of German.

2. THE KIEL INTONATION MODEL

The following points summarise the main components of the model.

(1) The highest level for the operation of the intonation rules is the sentence; thus a text has to be decomposed into a sequence of sentences. Although there are data pointing to an intermediate paragraph level between a text and its sentences (e.g. Lehiste 1975, 1979), signalled by prosodic features such as pitch, duration and pausing, this fact is ignored in the model until more details are known.

(2) The next lower level is represented by the phrase which is delimited by syntactic analysis, e.g. theme vs. rheme.

(3) Within each phrase, words that are to be accented have to be marked by a binary STRESS feature . Accentuation in phrases and sentences is not only governed by syntactic considerations, but also by pragmatic ones. Among others, the following cases have to be taken into account:

(a) Function words have -STRESS in the unmarked case; if they are +STRESS they convey contrast or insistence, e.g. 'Er wird nicht kommen können.' vs. '..."können.'

(b) If there are only function words in a sentence the function verb gets the feature +STRESS; if there are several function verbs the first two do, e.g., 'Es <u>wird</u> so <u>gewesen</u> sein müssen'. Further specifications are necessary to cover all the details of function word stressing.

(c) The word order rules of German require the final position of past participles and infinitives of composite tenses in main clauses, e.g. 'Er (wird) hat Peter einen Brief geschrieben (schreiben).' In the unmarked case this uninflected part of the verb is -STRESS; +STRESS again signals contrast or insistence since the verb then receives the final intonation turn

(falling or rising). The rules are more complicated than is suggested by this statement because deaccentuation also seems to depend on the grammatical cohesion between the verb and other components of the VP. When it is weaker, as in the case of an indirect object or an adverbial phrase, compared with a direct object, +STRESS is assigned to the verb, e.g. 'Er hat einem Freund geholfen.', 'Er hat endlich geschrieben.'

(4) Each word has to be analysed with regard to being simple or compounded. German compounds normally receive +STRESS on the first, -STRESS on all other elements, e.g. 'Schornstein-fegermeister'. The -STRESS elements of compounds are, however, not reduced to the same extent as totally unaccented units. To capture this difference, all +STRESS words as well as all -STRESS compound parts are at the same time given the feature +1STRESS. This also applies to the cases of deaccentuation in (3)(c).

(5) Within each word or compound element the syllable has to be determined whose vowel is to receive the ±STRESS, +1STRESS specifications. This is done by the word accent assignment rules. All vowels that are not given these feature combinations are -STRESS, -1STRESS.

(6) For each phrase it has to be decided whether the pitch is to rise or to fall towards the end, and if there is to be a rise the model has to choose between a low continuation rise and a high rise, as, e.g., in questions.

(7) In every +STRESS, +1STRESS vowel, except the last one in the phrase, a low and a high F0 point (90 and 130 Hz for a male voice) are fixed. Their temporal alignment is as follows: the low point occurs at the beginning of the accented syllable, i.e. its first consonant; the position of the high point is governed by the phonological quantity (short vs. long) and by the quality (close vs. open) of the accented vowel. The longer the vowel due to quantity and openness, the farther the high F0 point is moved into the vowel. Here the F0 rules interact with the duration rules. The F0 position is further adjusted to the lengthening or shortening of the vowel due to speech rate, but it is independent of the shortening of vowels in polysyllables as against monosyllables, or before voiceless as against voiced consonants. This means that the high F0 point is closer to the consonant in 'eine' vs. 'ein' and in 'leite' vs. 'leide'.

(8) In the last accented vowel of a phrase 3 F0 points are set. Three cases have to be distinguished:

(a) terminal fall: the first two F0 points as in the preceding accented vowels, the third at 80 Hz and at a constant time after the peak F0, depending on speech

rate (the end of the utterance is marked by a further F0 point at 70 Hz);

(b) continuation rise: 85 Hz at the beginning of the accented syllable, 100 Hz at the same point after vowel onset as in (a), and 120 Hz at the end of voicing before the phrase boundary;

(c) question rise: 90 Hz at the beginning of the accented syllable, 100 Hz at the same point after vowel onset as in (b), and 200 Hz at the end of voicing before the phrase boundary.

(9) Every phrase now contains an alternation of base and peak F0 points. In the case of a terminal fall, the series ends in a base point, otherwise in a peak point. The sequence of peaks within a phrase is downstepped by a semitone (=6 % of the F0 value) from each preceding peak. At the beginning of each subsequent phrase, the F0 peak is reset to its original value, and downstepping starts anew. If the downstepped peak goes below 95 Hz, this value is taken. This threshold prevents a descent to F0 values that are too low. The low base points, except the phrase-final one, are adjusted to 3 semitones (= 18 % of the F0 value) below the preceding peak. F0 declination is thus not handled by initially set lines on a logarithmic frequency scale over the time axis with varying steepness according to the utterance length, as is the case in, e.g., the Dutch model. There is also no declination in enumerations of items that all have continuation rises. This treatment of declination corresponds more closely to what is found in production and produces better perceptual results.

(10) To give any F0 peak extra prominence over its neighbours for increased semantic weight its value is increased by ΔF0. This also cancels the effect of the regular downstepping.

(11) If the terminal fall in (8)(a) is followed by -STRESS, 1STRESS vowels the right base point is shifted to the beginning of the last such vowel. This flattens the descent and gives non-initial compound units as well as deaccented words more stress. Thus deaccentuation for emphasis/ contrast and for syntactic constraints can be differentiated, e.g., 'Er hat einen Brief geschrieben.', with a fast final F0 fall over a fixed time after the peak to emphasise 'Brief', or with a slow descent controlled by the following -STRESS, 1STRESS vowel to signal the neutral case without special emphasis on 'Brief'.

(12) The falling peak patterns generated so far are all located around the centre of the accented vowel. This is the central peak position. Two further positions have to be provided for:

(a) early: the high F0 value is shifted leftward to the

beginning of the accented syllable, the left-hand low F0 value to a point preceding the high value by 100 ms, and the right-hand low F0 value - if there is one attached to the same accented vowel - to the original position of the high F0 value;

(b) late: the low F0 value is shifted to the right into the position of the peak value, the latter is shifted rightward by 150 ms in a medium speech rate (or to the beginning of a following unstressed vowel), and the right-hand low value - if there is one attached to the same accented vowel - to a point 100 ms after the peak.

These different peak positions signal different implications: the early peak has the function of summarising and concluding an argument, the central one of starting a new chain of arguments, and the late one of adding further emphasis, as, e.g., in an exclamation over some new discovery.

(13) If an early peak follows a central or late one its left base F0 point is deleted, thus eliminating a dip between two successive peaks (resulting in a "hat pattern") and preventing the left-hand low base point of the early peak from preceding the previous peak.

(14) The F0 peak values are increased by a factor of 1.08 for close vowels. This takes care of the vowel-intrinsic micro F0 in non-rising patterns.

(15) The F0 value at the beginning of the accented vowel after voiceless obstruent (+sonorant) consonants is raised by 15 Hz, and this ΔF0 is then linearly decreased to 0 at the next F0 point. This takes care of the contextual F0 increase after voiceless stops and fricatives. A similar adjustment before following obstruents has not been incorporated into the model yet, because the empirical facts have not been sufficiently analysed.

(16) Between the generated points, F0 is interpolated, e.g. linearly, which is the simplest procedure and results in a very close perceptual approximation of naturally produced pitch patterns.

(17) Finally, F0 is masked in voiceless sections of the speech signal. The generation of F0 contours thus proceeds on the assumption that they develop independently of the voiced/voiceless decision in segments. There are physiological grounds for this assumption (different muscles responsible for tensing the vocal folds and for separating them).

3. THE INFOVOX IMPLEMENTATION

Provided there is a powerful grapheme-to-phoneme module in a text-to-speech system most of the components of the Kiel intonation model can be implemented. The segmentation of the text into sentences and further into words provides no problems. Function words have to be marked in the lexicon for other purposes as well (duration and segmental reduction). The deaccenting of past participles and infinitives is far more tricky because it requires a syntactic component. The same applies to the segmentation of sentences into phrases. Commas according to punctuation rules help, but are not sufficient; they can, however, always be supplemented by subsequently introduced additional commas at phrase boundaries.

Falling and rising pitch patterns can be associated with punctuation marks . ! and , ? respectively. This is only a modest start, because question word and yes/no questions, for instance, have to be given different tunes in their unmarked cases, which further stresses the need for the incorporation of at least a rudimentary syntactic component into the text-to-speech system. Whereas the late peak may be associated with ! as a first approximation (although a separation of exclamations from commands will be necessary and will have to be handled by a syntactic module), there is at present no way to generate early F0 peaks automatically from text. Additional control symbols have to be introduced manually. Since in this case, pragmatic aspects are involved that go beyond the sentence no simple solution will be possible. Similarly, increased prominence of accented syllables can only be achieved by manually introduced stress level markers.

The marking of parts of compounds and of stressed vowels in words is very efficient in the INFOVOX German grapheme rules, which have been developed by the Kiel Institute, so that the assignment of the F0 values, as specified in the Kiel Intonation Model, to these focal points works very well. The output, as you can judge for yourselves by listening to the following longish prose text, sounds quite natural and does not have the dull ring about it, which has been attributed to the synthesis on the basis of the Dutch model. This is due, in particular, to the greater variety of patterns, the introduction of microprosodies, which add further diversity, to less temporal regularity in declination, and to the use of central rather than early F0 peaks. A very elaborate system of duration rules, which will be detailed elsewhere, adds to the naturalness of the speech output. Further research has to concentrate on improvements at the segmental and coarticulation levels. Greater naturalness can

also be achieved by slight random variations of speech rate, of pauses between phrases and sentences, and of the first F0 peak value in each phrase, because these random changes counteract the repetitiveness of the same patterns and are a more faithful representation of natural speech.

4. REFERENCES

L M H ADRIAENS, 'A preliminary description of German intonation', IPO Annual Progress Report 19 p36-41 (IPO, Eindhoven), (1984)

R CARLSON & B GRANSTRÖM, 'A text-to-speech system based entirely on rules', Proc. IEEE Intern. Conference on Acoustics, Speech and Signal Processing (Philadelphia), p686-688, (1976)

R CARLSON, B GRANSTRÖM & S HUNNICUTT, 'A multi-language text-to-speech module', Proc. ICASSP 82 (Paris), Vol. 3 p1604-1607, (1982)

E GÅRDING, 'Sentence intonation in Swedish', Phonetica 36 p207-215, (1979)

K J KOHLER, 'Macro and micro F0 in the synthesis of intonation', to appear in: Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech. (M Beckman & J Kingston, Eds.) (Cambridge University Press, Cambridge), (1988)

K J KOHLER, 'Computer synthesis of intonation', 12th Intern. Congr. Acoustics, A6-6 (Toronto), (1986)

I LEHISTE, 'The phonetic structure of paragraphs', in Structure and Process in Speech Perception (A Cohen & S G Nooteboom, Eds.) (Springer, Berlin), p195-203, (1975)

J B PIERREHUMBERT, The Phonology and Phonetics of English Intonation. Doctoral dissertation M.I.T., Cambridge, Mass. (1980)

J R DE PIJPER, Modelling British English Intonation. (Foris Publications, Dordrecht), (1983)