COMPUTER SYNTHESIS OF INTONATION

K.J. Kohler

Institut für Phonetik, Universität Kiel,
Olshausenstr. 40, 2300 Kiel, F.R.G.

ABSTRACT

This paper distinguishes between a global utterance intonation, which is related to the meaning to be conveyed, and local FØ perturbations, which are due to articulatory constraints. Results are presented of computer synthesis and listening experiments, and a set of rules is given for adjusting the same global contour to different segmental strings in German.

INTRODUCTION

Appropriate FØ patterns are essential in the synthesis-by-rule of natural-sounding speech. The analysis of natural speech production has established the following points of FØ control:

1. A global utterance intonation, which is related to the meaning to be conveyed, has to be distinguished from local FØ perturbations, which are due to articulatory constraints.
2. Apart from local adjustments, FØ develops as if there were no voiceless sections, i.e. it continues after a voiceless interruption from a value it would have reached had there been voicing.
3. In the presence of voiceless sections, the timing of FØ has to be such that its characteristics (e.g. the peak value and the indication of an FØ descent) fall within a voiced stretch of speech and can thus be recovered by a listener.

This paper discusses the incorporation of these points into an FØ synthesis-by-rule for German. It limits itself to falling terminal contours containing a single peak.

GLOBAL FØ

In a sentence such as "Sie hat ja gelogen." ("She's been lying."), with focus stress on the syllable lo /lo:/, the FØ peak can be on the syllable ge, preceding the stress, or at the center of the stressed syllable, or at its end (see fig. 1). This shift in the FØ peak position is correlated with a change in meaning from "established" to "new" to "emphatic".

Using the Kiel Phonetics Institute Speech Signal Processing program (SSP, cf. Schäfer, 1982) and the pitch algorithm in it (Schäfer-Vincent, 1983), LPC-based synthesis of the above sentence with a stepwise shift of the FØ peak contour A1A2 along the time axis was carried out, starting from the center peak position (see fig. 1a) and moving it to the left in 6 equal steps as well as to the right in 4 equal steps of 30 ms each. The shifted contour was masked in voiceless stretches, its right-hand branch time-expanded in the left shift, and the FØ transitions smoothed (see fig. 1b). The resulting 11 stimuli were presented once as a series of left to right FØ peak shifts to 33 native German speakers, who had to mark the positions in the series at which they perceived changes. Table I lists the results.

Table I. Number of changes perceived at the positions II-XI of a series of left to right FØ peak shifts in "Sie hat ja gelogen."; 33 subjects.

first change at position

II	III	IV	V	VI
0	1	2	20	10

further changes at position

V	VI	VII	VIII	IX	X	XI
1	3	5	9	8	15	9

The sharp increase in the "first change" score from stimulus IV to stimulus V suggests that a category boundary is transgressed at this point in the series. The greater spread of "further change" answers and the lower second maximum score at stimulus X point to a more gradual change after the first category boundary. Stimulus V is the first in the series that has the FØ peak in the stressed vowel; in stimulus IV it is at the consonant/vowel boundary. In the original stimulus VII, the FØ peak is located about 100 ms after vowel onset. Since there still is a substantial "first change" score at stimulus VI the FØ peak should occur later than the 60 ms mark to convey the meaning "new" unambiguously. But it may be shifted into the vowel by another 30 ms without losing its semantic identity because the next maximum score only occurs at stimulus X. Thus the FØ peak signaling the concept "new" should be located at the center of a phonologically long vowel, i.e. about 100-120 ms into a vowel of 200-250 ms duration at a medium speech rate.

LOCAL FØ

The next question to be answered concerns the manifestation in other segmental strings of an FØ peak contour with the same meaning "new". To investigate these local FØ adjustments of the same global contour a set of naturally produced utterances of the type "Sie malt." /zɪ 'ma:lt/, "Sie macht." /zɪ 'maxt/, "Sie malen." /zɪ 'ma:lən/, "Sie machen." /zɪ 'maxən/, "Sie strickt." /zɪ 'ʃtrɪkt/, "Sie niesen." /zɪ 'ni:zən/ etc., i.e. with long and short, low and high vowels in voiced and voiceless consonantal environments and in mono- and multisyllable words are analyzed with the help of SSP. Then the center FØ peak contour of "Sie malt." is transferred to the other utterances and adjusted according to the various conditioning factors by the following set of ordered rules.

(1) The FØ peak is positioned on the time axis in such a way that, within the same speech rate, the timing of the FØ peak contour as such stays the same (independently of the segmental timing), provided the listener receives a clear indication of the central FØ rise-fall and its peak value on the stressed word. This means that the FØ peak occurs at the same absolute time of about 100 ms after stressed vowel onset in all phono-

logically long vowels, and in phonologically short vowels if they are followed by a voiced consonant or another syllable to signal the FØ descent; if, in the latter case, the FØ peak falls inside a voiceless consonant and is thus not recoverable by a listener, it is brought forward to the end of the short vowel. In the case of only a voiceless consonant following a short vowel in a monosyllabic word, 30-50 ms have to be provided for both an FØ rise and an FØ fall inside the vowel, otherwise the terminal nature of the global intonation would not be signaled to a listener.

(2) The whole FØ peak contour in its original duration from the beginning of /m/ to the end of /l/ in "Sie malt." is transferred to the new segmental string and time-locked to the peak point as fixed in (1).

(3) The right-hand branch of the FØ peak contour is time-expanded to fit multisyllable words.

(4) To account for vowel-intrinsic FØ differences, the FØ peak contour is expanded in the frequency domain for high vowels in such a way that the FØ value at the vowel center (STRESSPOINT) is raised by a constant factor for male and female speakers. The program then interpolates FØ between this point and the time markers at the beginning and the end of the peak contour according to the proportion $(n_1/n)\Delta F\emptyset$, where n = total number of FØ values between one of the time markers and the stresspoint, n_1 = number of FØ values from the time marker to the FØ value to be changed. Thus the FØ values at the boundaries of the peak contour stay the same, whereas the others between a boundary and the stresspoint are changed in steps to a maximum $\Delta F\emptyset$ at the vowel center. This procedure guarantees a maximal intrinsic vowel influence on FØ at the vowel target and takes the preceding and following coarticulation into account, providing weaker effects at vowel onset and offset.

(5) There is an increase of FØ at the transition from a voiceless fricative to a voiced sonorant through a strengthening of the Bernoulli effect, and the opposite at the reversed transition. This consonantal effect on FØ is accounted for by frequency expansion or compression according to the same formula as in (4) with the beginning/end of a vowel after/before a fricative as the stresspoint and the vowel center as a time marker. This way the effect is strongest at the vowel boundary and disappears at the vowel center.

(6) The FØ contour for "Sie" is transferred separately.

(7) FØ is masked in voiceless stretches.

After LPC-based synthesis of the rule-generated FØ patterns over the various segmental strings, the utterances are arranged in pairs, and listeners have to judge whether they represent the same intonation pattern. The procedure of FØ manipulation and auditory evaluation is repeated until listeners accept all the pairs as belonging to the same pitch category. The result is a speaker-independent set of rules which generates all the local FØ perturbations in the same global pattern, starting from one basic contour.

REFERENCES

Schäfer, K. (1982). "Concepts in the SSP programme," Arb. Inst. Phonet. Univ. Kiel (AIPUK) 18, 18-31, 110-121.

Schäfer-Vincent, K. (1983). "Pitch period detection and chaining: method and evaluation," Phonetica 40, 177-202.

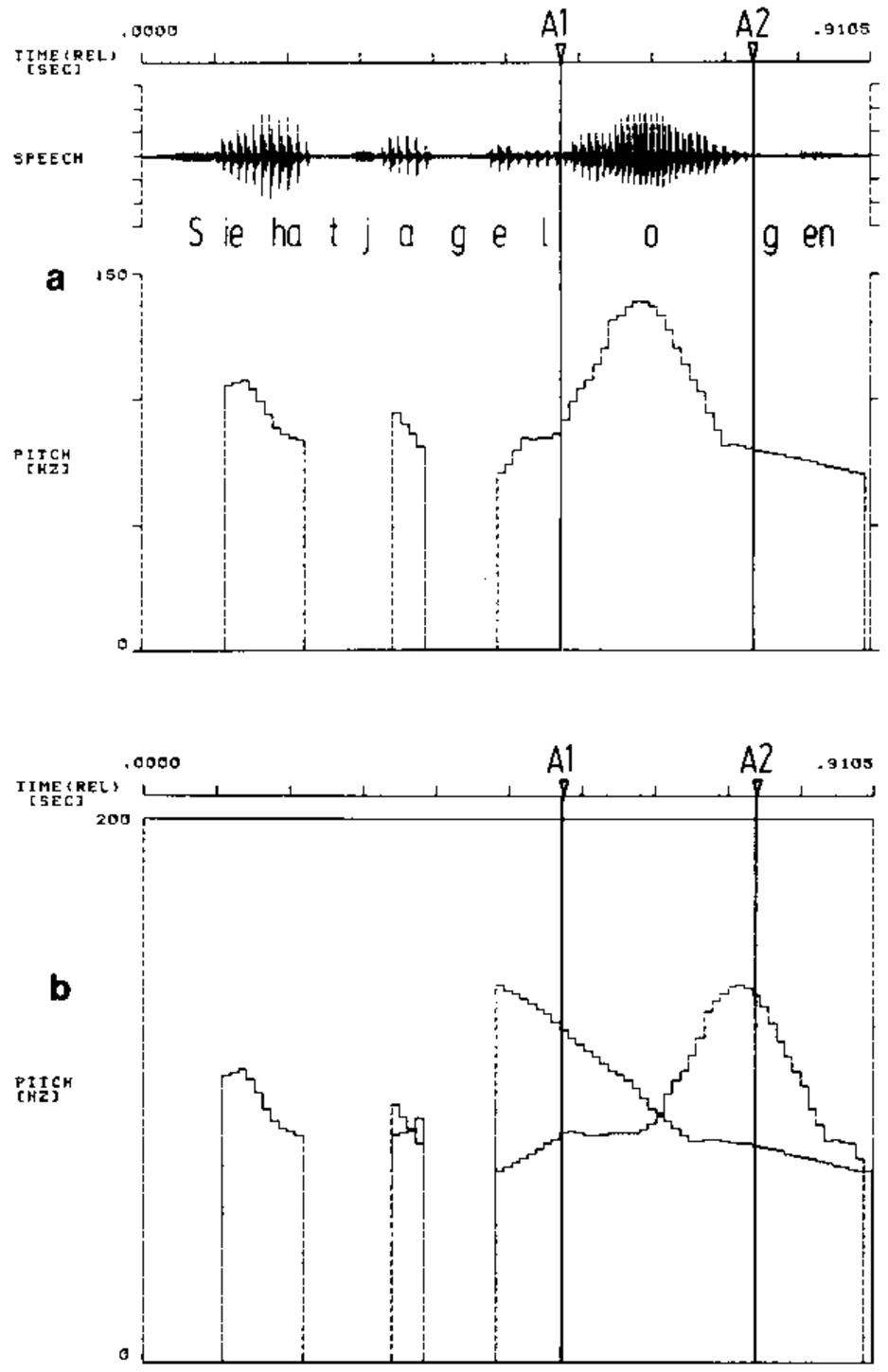

Fig. 1
(a) Speech wave and fundamental frequency (center peak) in the naturally produced utterance "Sie hat ja gelogen." The end contour (on the syllable gen) was added by FØ parameter manipulation because the analysis did not provide it. The time markers A1, A2 delimit the FØ peak contour (coinciding approximately with /o:/) which was shifted left and right.
(b) The left- and right-most positions of the shifted FØ peak contour on the same time scale as in (a), approximating the natural productions of early and late peaks, respectively.