A grammar of intonational units in German digit numbers

Irene Jacobi and **Uwe D. Reichel**

Department of Phonetics and Speech Communication
University of Munich, Schellingstr. 3, 80799 München, Germany
{jacobi,reichelu}@phonetik.uni-muenchen.de

ABSTRACT

This study is concerned with intonation patterns within the restricted domain of German numerals. Regular and non-regular (random) numbers were recorded and F_0-contours extracted. Manual segmentation followed by comparison yielded some general intonation patterns as well as other regularities which were speaker-specific. Stress, repetitions of digits and the position of a digit in a number had an influence on the direction and amplitude of F_0-movements. To test the reliability of fundamental frequency patterns found as being recurrent and distinguishable, F_0 was reduced to a straight line for each digit word and resynthesized with PSOLA for perceptual judgement. The results indicate that three different simple straight-line F_0-units sufficiently represent intonational information within German numbers.

1 Introduction

The importance of intonation for synthesized speech has been shown in various investigations. Naturalness of synthesized speech as well as stress in general are connected with varying F_0 and duration. Like other utterances numerals are prosodically structured. The reason or result beside comprehension is a better memorability. Previous examinations of 6 and 7 digit telephone numbers already revealed speaker-dependent grouping strategies including groups of single digits, tens and hundreds [1]. Our aim was to find intonational units within numbers; therefore, speaker productions of the same grouping strategies had to be analysed, F_0-contours had to be extracted, and detected units had to be judged with the help of a perceptual test.

2 Data and method

First two women and two men were recorded uttering numbers they knew by heart. Usually they were uttered in digit words (*three, four* instead of *thirtyfour*) and structured into groups of two or three digits. Then subjects had to read a sequence of visually ungrouped digits. Though having been instructed to utter the numbers in an ungrouped way none of the subjects was able to do so. Not grouping seems to be unnatural, nevertheless, this is the way many providers present a sequence of spoken digits to their customers. Subsequently, all numbers were presented visually parted into double- or triple-digit groups being the preferred group-quantity.
380 numbers were made up of four, six, and nine digits comprising non-regular (random) and regular digit numbers, where 'regular' means that the numbers contain repetitive digits or successively decreasing or increasing digits (e.g. *347 357 367*). The 4-digit numbers made up the first part of the 6-digit numbers which made up the first part of the 9-digit numbers (e.g. *14 27 / 14 27 90 / 142 790 / 142 790 538*). Numbers were presented in random order and read aloud by the four subjects at least once. Subjects were instructed to first read the number silently and then utter it in the way they would recall it. None of the subjects called into question how the presented visual grouping of the digit sequence should be expressed.

3 Analysis

The F_0-contours of all recorded spoken numbers were extracted, digit words and voiced segments manually segmented and labeled, revealing some general intonation patterns. Grouped digits correspond to an intonation phrase demarcated temporally by short breaks. The standard pronunciation of German single digits is as follows: [nʊl], [a͜ins], [t͜sva͜i] or [t͜svoː], [dra͜i], [fiːɐ̯], [fʏnf], [zɛks], [ˈziːbn̩], [axt], [nɔ͜ʏn]. Reflecting the amount of their voiced sounds digits 6 and 8 tend to have the shortest voiced phases, whereas 9, followed by 0 show the longest. Many pronunciation variants could be found within unstressed digits. Generally, digits were uttered as syllable; even the assumed disyllabic digit word 7 was usually uttered as [ziːm].

3.1 Group-independent patterns

A regularly recurring shape in the course of F_0 resembles an *S*, clearly appearing in the fundamental frequency contour of stressed digits. The *S*-shaped as-

ISBN 1-876346-48-5 © 2003 UAB

cending of F_0 was observed by Delattre [2] as a characteristic form of German F_0-contours. In contrast the F_0-contour of e.g. American English is an opposite S characterized by a descending course. As global intonation phenomenon F_0-movements show the tendency to decline and within a number the gradient of the S-shaped curve declines from initial to medial group, downright reversing within the final group (cf. fig.1-3).

3.2 Stress

As expected, stressed digits show longer syllable durations than unstressed digits. Apparently the duration and with it the length of voicing of an unstressed digit proportional to a group's stressed digit stays unattached from it being the only unstressed digit ahead of a stressed one or having another unstressed digit in the same group. Stressed digits show a rising F_0-contour unless they are a final group's last or a triple-group's middle digit (cf. fig.3). In this case there will be a falling F_0-contour, differing from unstressed very last digits by steeper incline or stronger movement. Additionally, stressed digits tendentially show a steeper gradient within their F_0-contour than unstressed ones. F_0 peaks do not provide the crucial indication of the location of accent. Rather, the strength of movement and the duration of a F_0-section portend on the prominence of the respective digit. Within stressed digits a correlation between the gradient of the F_0-contour and its duration was found.

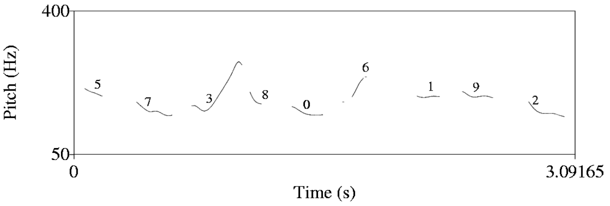

Figure 1: F_0-contour of *573 806 192*.

3.3 Default stress and stress shift

Repetitions of digits and the position of a digit in a number had an influence on the direction and the strength of F_0-movements. Particularly the comparison of regular digit sequences (and thereby occurring shifts of stress) with irregular digit sequences can give information about those F_0-movements that are relevant for prominence and furthermore which occurences are not ascribable to stress. Within irregular sequences usually a double- or triple-digit group's last digit will be the stressed one (cf. fig.1). Within regular sequences containing recurring or regularly ascending or descending digits mostly the irregular and therefore newly introduced digits will be stressed. These digits then show a stronger F_0-movement and a steeper gradient than the unstressed ones (cf. fig.2,3). Generally the F_0-contour of a triple-group's second and third digit aligns with a double-digit group contour. Conspicuous is a slight F_0-break in front of stressed digits within a triple-group. This is not exclusively due to digits' voiceless fricatives or plosives as the time signal indicates a short pause as well. Two speakers tended to put a stress additionally on the first digit of a triple-group. Also speaker-specific was the tendency of one subject to show a F_0-maximum in the final double-groups during the first non-prominent digit.

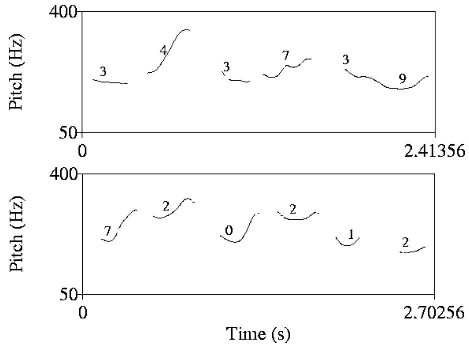

Figure 2: F_0-contours of the double-digit grouped numbers *34 37 39* (top) and *72 02 12* (bottom). Stressed digits are underlined.

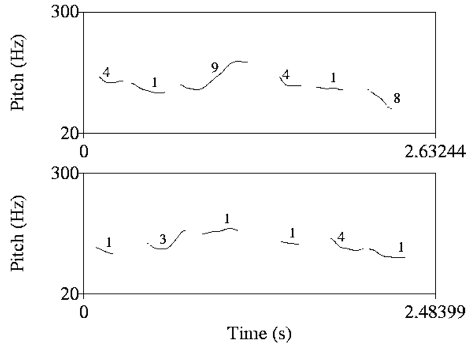

Figure 3: F_0-contours of the triple-grouped numbers *419 418* (top) and *131 141* (bottom). Stressed digits are underlined.

Generally, medial group patterns resemble initial group patterns with only weaker parameter values. Initial and final group patterns don't change with the absence or presence of a medial group.

4 Stylization

As mentioned in 3.2 maxima of the fundamental frequency are no safe index towards the location of accent. If stress cannot be derived by the ascent or decline of F_0 a comparison of the durations of the corresponding F_0-segments or the dimension of movement will still be a better indication than a F_0 peak.

Following these results – at least concerning digit sequences – intonational units should be made up of F_0-patterns differing at least in duration and gradient. For digit numbers the smallest intonational units could be equated with syllables as all digits were uttered mono-syllabically, the only exception being the numeral 7, often uttered mono-syllabically but occasionaly uttered disyllabically. Indicated by an analysis of the same 380 numbers uttered in tens and hundreds, intonational units for number words would be one- or disyllabic sequences comprising the information on the numerical values of the digits that together compose the number's value (e.g. *three-hundred-four-teen*).

To clarify the digit's F_0-contours for further examination the extraction and interpolation of only few basic values from the voiced segments seemed to be expedient. Though t'Hart [4] points out that there's hardly a perceptual difference between the stylization by straight lines or parabolas, this is of course dependent on the amount of values used for interpolation. The preferably sparse amount of values was picked out in such way that after interpolation both a curve and the double turnaround in case of an *S*-shaped contour would be registered. Linear interpolation and an amount of five F_0-values seemed to be advisable. The first and fifth value were means of the first three resp. last three voiced frames as the algorithm calculating the F_0 value should not include in the final values the microprosodic effect on F_0 which arises at the beginning and at the end of the voiced frames due to voicing and devoicing. The other three values were extracted in equidistant steps between the first and fifth value. Stylization like this was very close to the original (cf. fig.4) and since not all F_0 variations are equally important to perception the following aim was thus to further reduce the F_0-contours to the perceptually relevant acoustic features, abstracting away from redundant information. Knowing that stressed segments opposite to unstressed segments should be especially marked (see [3]) a further stylization should accordingly represent in particular stressed segments appropriately. Though the *S*-shape emerges over all speakers, the shifts of directions delimiting the main gradient of the slope often appear in a reduced form revealing their redundancy and leaving a quite straight middle part of the shape, so that further linear regression seemed to be an appropriate solution. Following this the five extracted values were regressed to one straight line.

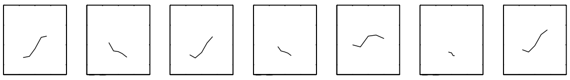

Figure 4: Capturing F_0-movement by interpolating five values.

Comparing the lines, the first digits of a group tendentially show no gradient or a negative one (cp. fig.5,6). An exception are the values of one subject who shows a slightly ascending slope in the final group during the non-prominent first digit (cp. fig.5, cp. 3.3). The scatter plot of the first digits of the triple-groups are diffuse since two speakers tended to stress the first digit in addition to the third digit (cp. fig.6). The inital and medial third digits show a positive gradient ebbing with gaining length of the voiced section. This coherence of the length of a voiced section and its steepness is speaker independent. Whether a stylization with straight lines truly catches the crucial intonational event and the presumptions can be kept, was explored by means of perceptual judgement.

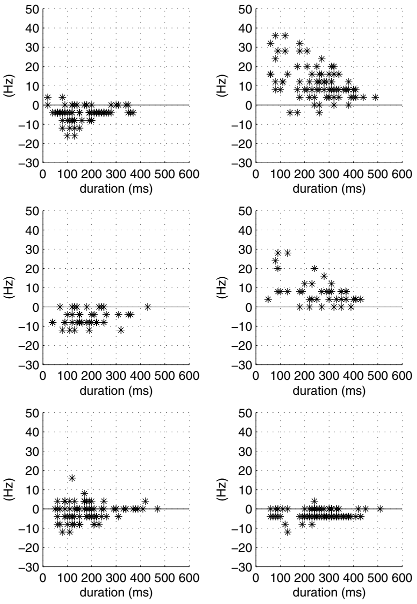

Figure 5: Scatter plots of cross section F_0-gradient of stylized double-grouped unregular numbers, overall speakers. Rows: inital, medial, final group. Columns: 1st, 2nd digit.

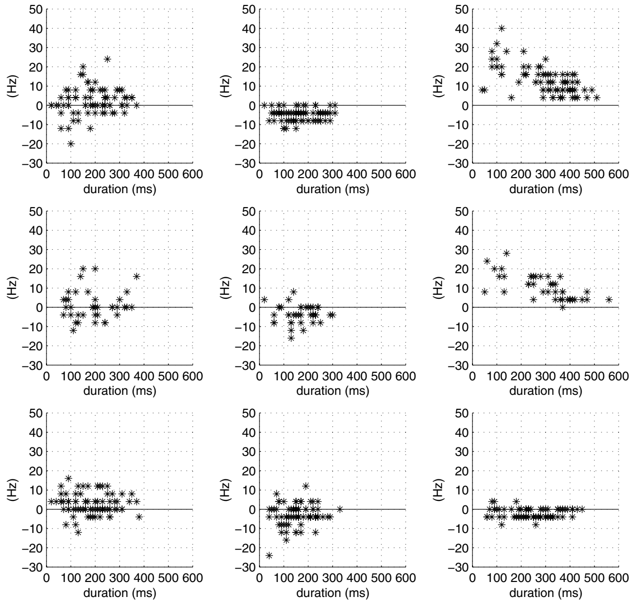

Figure 6: Cross section F_0-gradient of stylized triple-grouped unregular numbers, overall speaker. Rows: inital, medial, final group. Columns: 1st, 2nd, 3rd digit.

ISBN 1-876346-48-5 © 2003 UAB

5 Perceptual Judgement and Interpretation

To sound natural and to be perceived as a movement instead of a pitch leap a F_0-change must persist for a certain time span [5]. Whether the rough approximation of the fundamental frequency contours by means of regression lines brings about perceptual losses compared to the original was checked via perceptual judgement. Using headphones 17 native listeners who were either trained phoneticians, students or staff of the IPSK or persons without any relation to phonetics judged the resynthesized data compared to the original utterances by means of an ABX-test.

5.1 Stimuli

First the original utterances were isolated from their spectral features with the help of *Praat*, then the F_0-course was stylized followed by resynthesis on the basis of PSOLA. In an informal test the synthesized utterances were then intensely compared perceptually to the original and proved themselves to be not equivalent. The perceptual deficiencies were tentatively remedied by changing the regression lines' degree of gradient within only the stressed segments.

Basis for the test were 48 digit sequences that were presented in a particular combination to the listeners for judgment. The sequences were made up of 16 recorded originals, their 16 by means of regression lines mathematically stylized versions and a further 16 where the lines had been perceptually revised (example fig.7). The originals were a 4-digit, a 9-digit and two 6-digit numbers spoken by each speaker once. Each presented stimulus was made up of three digit sequences containing at least one original and either the mathematically or perceptively stylized version(s) of that original so that the third of the presented digit rows was identical either to the first or the second stimulus.

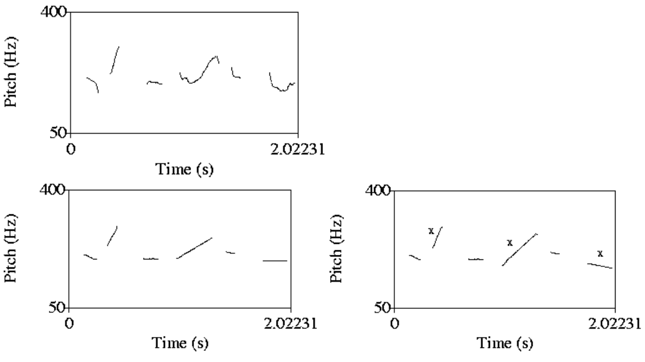

Figure 7: *28 13 64*. Top: original F_0-contour; bottom: stylized contours, perceptually changed lines marked with an x.

Each stimulus appeared in four different environments resulting in 128 different rows of stimuli which were arranged in random order. The first 12 stimuli were repeated at the end so that the final sequence for judgement consisted of 140 stimuli.

5.2 Perceptual results

All 17 sample distributions were Gaussian and below an 0.01 level of significance. None of the stylizations could be safely differentiated from the original as differentiation of mathematically stylized and original stimuli amounted to an average of 64.7% and differentiation of perceptually stylized and originals to 57%. The results of a t-test showed that the differentiation values of the two F_0-stylizations diverged four times (highly) significantly. For each of the significantly diverging courses steeper slopes than the regression lines of the five extracted values were identified worse, indicating different perceptual weight of the S-shape-segments.

6 Discussion

Since differentiation of stylized and original stimuli was poor, and bearing in mind that even a differentiation would say nothing with regard to the acceptability of straight line fundamental frequency contours, a small number of different simple straight-line fundamental frequency units seems to be sufficient to represent the intonational information in German numerals. This would be three straight F_0-lines differing in duration and gradient: short duration with a neutral gradient for unstressed digits and longer duration with a rising respectively falling gradient for stressed digits. The intonation contours of unstressed digits are not crucial for perception and especially the intonation contours of a sequence of two unstressed digits could be further simplified for speech synthesis.

REFERENCES

[1] Baumann, S., Trouvain, J. "On the Prosody of German Telephone Numbers". *Eurospeech*, Scandinavia, 2001.

[2] Delattre, P. *Comparing the Phonetic Features of English, German, Spanish and French.* Heidelberg: Julius Groos Verlag, 1965.

[3] van Santen, J., Moebius, B. "Modelling pitch accent curves". *Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications*, Athens, 1997.

[4] t'Hart, J. "F0 stylization in speech: Straight lines versus parabolas". *JASA 90,6*, 1991.

[5] t'Hart, J., Collier, R., Cohen, A. *A perceptual study of intonation.* Cambridge, U.K.: Cambridge University Press, 1990.

ISBN 1-876346-48-5 © 2003 UAB