On the Prosody of German Telephone Numbers

Stefan Baumann & Jürgen Trouvain

Institute of Phonetics, University of the Saarland, Saarbrücken, Germany
{baumann,trouvain}@coli.uni-sb.de

Abstract

Spoken telephone numbers are prosodically structured. This is reflected on various levels, such as grouping, wording and accenting. Realisation strategies employed by German speakers are used to model the prosody of telephone number production. In a listening preference test using synthetic speech two strategies used by commercial inquiry systems proved to be less acceptable than the versions based on the proposed models. These models are proposed for use in speech-synthesis-based telephone number inquiry services.

1. Introduction

The fact that the prosody of telephone numbers has been of very little interest in linguistic research leads to the assumption that this topic is widely regarded as a trivial matter. However, a closer investigation shows that the prosodic structure of telephone numbers is far from clear-cut, at least as far as telephone numbers in German are concerned. A brief inquiry in the "Speech Prosody Special Interest Group" [1] and an informal search among colleagues revealed that the same telephone numbers are structured quite differently across the world. It also appears that there are clearer conventions in other countries compared to Germany. Despite this, there is an official standard [2] for structuring *written* telephone numbers in Germany.

1.1. Aims of the Study

The central aim of the study is to explore the prosodic structure of German telephone numbers to an extent that enables us to propose an appropriate model of description and generation. Such a model could e.g. serve to improve current telephone information systems based on synthetic speech.

In order to reach this goal we take three steps. Step one is a production experiment. Here we investigate which grouping or chunking strategies subjects apply to telephone numbers of different sizes and arrangements, which type of wording is used under varying conditions, and how numbers are metrically and intonationally structured. The insights derived from the data lead to a preliminary model (step two), whose acceptability is evaluated in a perception experiment (step three). In this experiment, our prosodically complex model of telephone numbers was tested against two forms of presentation implemented in telephone information services.

1.2. Grouping

The grouping of telephone numbers is peculiar in several respects, regardless of language. First, telephone numbers are usually not treated as a single large unit, unlike other numbers used e.g. in a currency context or in an arithmetical problem ("1356829" rendered as "one million, three hundred and fifty-six thousand, eight hundred and twenty-nine"). Rather, telephone numbers are grouped into smaller units, which can also be regarded as semantic or pragmatic sense units. Preferred sizes of such units are groups of two or three digits; sometimes remaining single digits form a group of their own. Thus, the 7-digit number "1356829" is usually grouped either as "135-68-29" (3-2-2), as "13-56-829" (2-2-3), as "13-56-82-9" (2-2-2-1) or as "1-35-68-29" (1-2-2-2).

Second, the grouping can be predetermined by various "meanings" and structures of particular digit combinations. They may e.g. function as dialling codes (country codes, area codes, telephone company codes), which in Germany start with a "0", have a limited length, a certain position, and often a geographical meaning. They are probably stored in the mental lexicon as a single unit. Furthermore, dialling codes are generally separated from the subscriber number by spaces or dashes. Due to these properties a special status can be attributed to them suggesting a special grouping strategy.

Third, there are combinations of digits within the subscriber numbers which might influence the choice of grouping and the size of units. Some of these are consecutive identical digits ("777"), consecutive ascending or descending digits ("789"), and complete tens or hundreds ("70" or "800"). It is likely that these special features form a unit of their own or, at least, are part of the same unit, which might lead to a conflict with an otherwise preferred grouping strategy. Finally, there are commonly known and/or advertised numbers ("11-8-33"), and a wide range of idiosyncrasies which are unpredictable and might cause deviations from an expected grouping.

1.3. Wording

The above example for the number "1356829" spoken in full, demonstrates that grouping is a prerequisite for wording: A million consists of 7 digits, but it can only be pronounced as a million if the number is treated as a single unit. Since telephone numbers are preferably grouped in chunks of two and three digits, usually just the wording of numbers as tens and hundreds occur, apart from the commonly used alternative to render the numbers digit by digit. Thus, two-digit units ("53") will either be rendered as ones ("fünf drei") or as tens ("dreiundfünfzig"), three-digit units ("539") will either be rendered as ones ("fünf drei neun") or as hundreds ("fünfhundertneununddreissig"). Note that a wording strategy never applies across group boundaries: A 3-digit number, e.g., grouped as "7-22" will never be rendered as a three-figure number ("siebenhundertzweiundzwanzig") but either as "sieben zweiundzwanzig" or as "sieben zwei zwei".

1.4. Prosodic Structure

Although there is no difference between digits with respect to their semantic weight or importance, there actually *is* a difference with respect to their prosodic weight. In a group of two items the last one is generally more prominent. In

Metrical Phonology [3] this relation is expressed as a weaker-stronger pattern. The accent structure is superimposed on the metrical structure, which means that the stronger item (s) receives a pitch accent whereas the weaker item (w) does not. Pitch accents are characterised by a combination of local pitch movement, greater duration and greater intensity. The type of the pitch accent, whether e.g. falling or rising, can vary e.g. with the position of the group in the utterance. Groups are delimited intonationally by boundary tones, so that a group corresponds to an Intonation Phrase (IP). The overall intonation contour of an utterance depends on both accent type and type of boundary tone. Temporal markers of prosodic phrasing are silent pauses and the lengthening of final syllables.

2. Production Experiment

2.1. Methods

10 students from the General Linguistics Department served as subjects. Each subject was recorded while dictating a total of 30 German telephone numbers to the experimenter. The numbers were printed on index-cards (one number per card) and presented once in randomised order. The subjects were asked to take one card at a time, look at the number in order to become familiar with it, and read it out aloud in such a way that the experimenter can write the number down.

2.2. Material

The subset of data presented here contained telephone numbers of 6 and 7 digits with or without various kinds of (mostly 4-digit) dialling codes (country codes, area codes, codes for mobile phones, for telephone companies, or codes with special call charges). Graphical separators occurred only after or within a dialling code and never within the subscriber number ("0177-36035"). Furthermore, the data included the following special numbers:

- consecutive identical digits ("7747703")
- repeated but non-consecutive identical digits ("9387868")
- consecutive ascending digits ("1234784")
- zeros in different positions ("9702403")
- widely known numbers ("11833")

2.3. Results

2.3.1. Grouping

Only a quarter of all recorded numbers were grouped in the same way across all the subjects. Every single telephone number was realised by the 10 subjects in 4 different versions on average (ranging from 2 to 6 versions). Every speaker had a basic favourite grouping strategy but the strategy could be disturbed by the factors listed in the above special cases.

The expected special status of dialling codes was confirmed by the data. They were almost exclusively grouped as units of their own. However, 11% of the dialling codes were internally divided by minor prosodic boundaries, henceforth referred to as intermediate phrases (ip).

Intonation Phrases (IPs) of up to four (mostly monosyllabic) items were used by the subjects, with a size of two or three items being clearly preferred. In 6-digit subscriber numbers, the strategies 2-2-2 and 3-3 were most frequent and equally distributed. In 7-digit numbers we found three competing grouping strategies, also equally distributed: 3-2-2, 2-2-3, and 2-2-2-1. It was not always clear, though, whether a 2-2 combination should be counted as two groups or as only one group separated by an ip-boundary.

Whenever consecutive identical digits occurred, they were grouped into the same unit in 89% of all cases. They formed a unit of their own ("xx-77-xx") in 48% and were part of the same unit ("x77-xx-xx") in 41% of all cases.

Consecutive ascending digits are represented in 60% of all test numbers in question by a single unit ("1234-xxx") or part of a unit ("xxx-0567").

Zeros often serve as boundary markers. In 57% of all numbers which include a zero, it occurs at the right edge of a group ("xx0-xx"). Group-initially ("xx-0x"), zeros occur in 28% and group-medially ("x0x") in 15 % of all cases.

2.3.2. Wording

The wording of dialling codes as single numbers proved to be the default case (64% of all dialling codes). However, the wording of dialling codes is often predefined. As examples, the code "0800" indicates a special call charge and was rendered by all subjects as "null achthundert", whereas the dialling code "01024", representing a specific telephone company, was rendered as "null zehn vierundzwanzig" by most of the subjects.

Contrary to dialling codes, subscriber numbers are semantically and structurally neutral. In 82% of all cases the numbers were pronounced as single digits, 10% of the numbers included tens, 1% hundreds. In 3% of cases both wording strategies, tens and hundreds, were applied in the same number. Nine out of ten subjects realised the purely repetitive number "2222222" as "seven times the two". As expected, hundreds proved to be the largest figures used. In half of all tens (subscriber numbers only) a combination of two identical digits is involved ("477-22-6" is spoken as "vier siebenundsiebzig - zweiundzwanzig - sechs"), which suggests that two consecutive identical digits not only trigger a grouping in the same unit but also a pronunciation as tens.

2.3.3. Prosodic Structure

The metrical and the intonational structures were used in a very consistent way and are considered in greater detail in the model presented in the next section.

3. A Model for Telephone Number Prosody

Our goal was to propose an appropriate model for the description and generation of German telephone numbers. For now, we will restrict ourselves to a digit-by-digit wording of 7-digit numbers since this size seems to be very frequent nationally and internationally.

In table 1, the three derived prosodic strategies are described and contrasted to two strategies from telephone inquiry systems. The first one (TEL1) follows the suggested writing norm [2] prescribing a pairwise grouping from right to left, the second one (TEL2) groups each digit into a unit of its own. Additionally, TEL1 is characterised by a division of intonation phrases into minor phrases with less salient ip-boundaries marked by shorter pauses.

Table 1: Strategies with groupings and metrical structures.

strategy	grouping	metrical structure
NAT 1	3-2-2	sws - ws - ws
NAT 2	2-2-3	ws - ws - sws
NAT 3	2-2-2-1	ws - ws - ws - s
TEL 1	1-2-2-2	s - s(-)s - s(-)s - s(-)s
TEL 2	1-1-1-1-1-1-1	s - s - s - s - s - s - s

The general prosodic patterns found in the production experiment plus those found in the TEL versions are illustrated in figure 1 and described using the GToBI labelling conventions [4], which are based on ideas of Autosegmental-Metrical Phonology [3].

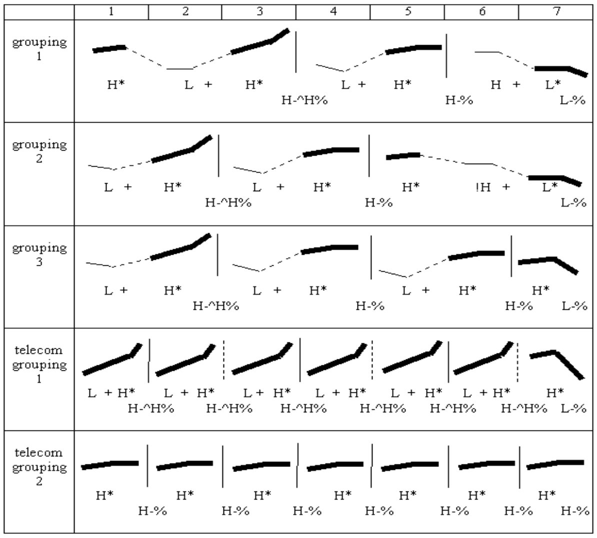
1
2
3
4
5
6
7
grouping 1
H* L + H* H-^H% L + H* H-% H + L* L-%
grouping 2
L + H* H-^H% L + H* H-% H* !H + L* L-%
grouping 3
L + H* H-^H% L + H* H-% L + H* H-% H* L-%
telecom grouping 1
L + H* H-^H% L + H* H-^H% L + H* H-^H% L + H* H-^H% L + H* H-^H% L + H* H-^H% H* L-%
telecom grouping 2
H* H-% H* H-% H* H-% H* H-% H* H-% H* H-% H* H-%

Figure 1: Schematic contours. Lines represent: accented (heavy), unaccented (thin), transitions (dotted), IP-boundaries (solid vertical), ip-boundaries (dotted vertical).

In the production data, secondary (prenuclear) accents are frequently realised as H* pitch accents, and primary (nuclear) accents mostly surface as L+H* in all IPs except the last one. If the final IP consists of only one digit, its nuclear accent is realised as H*, if it consists of two digits, the nucleus is generally realised as H+L*, and if the final IP has three digits, the most frequent realisation of the nuclear accent is a downstepped fall !H+L*. The form of contour at boundaries is similarly clear-cut. The first IP normally ends in a steep rise (H-^H%), subsequent IPs (except the last one) end in a high plateau (symbolised as H-%), and the final IP falls to an extra low pitch (L-%). In terms of global intonation contours, there is a stepwise lowering of register, i.e. rises at boundaries are getting shallower from IP to IP. This overall contour has already been reported on for generic lists in German [5].

4. Perception Experiment

The perception experiment aimed to compare the three strategies found in the production experiment with natural speech (NAT1,2,3) with the two above mentioned strategies in telephone inquiry systems (TEL1,2). In a direct comparison listeners were asked in a forced choice test which of two consecutive stimuli they preferred. The hypothesis was that the strategy with hardly any prosodic structure (TEL2) would never be preferred, and that the unnaturally realised phone inquiry versions were inferior to the "natural" counterparts. This would be in line with the findings of a similar test [6] with "neutral", "continuant" and "terminator" intonation.

4.1. Methods

Three 7-digit numbers without an area code were selected. To guarantee a "default" mode, special cases as listed in section 2.2 were avoided. Additionally, the digits "0" and "7" are excluded for reasons of special status and the number of syllables ("sie-ben"), respectively.

Stimuli were generated with an MBROLA diphone speech synthesiser [7] using the German female diphone set. To determine the fundamental frequency, a starting and a terminating value in Hz was chosen for each syllable which was based on our recordings and corresponded to the five strategies in the proposed model. For the selection of the syllable and segment durations two factors were considered: On the one hand the speaking rate of the telephone inquiry voices where only accented syllables occur should be matched, on the other hand deaccented syllables had to be shortened. A native speaker produced the three numbers so that her syllable durations could serve as an orientation. All synthesised digits were checked and corrected by the authors so that they corresponded to the prosodic conditions to be modelled. Mean syllable durations were 340 ms for unaccented and 480 ms for accented digits. Pause durations were 450 ms at an IP-boundary, and 150 ms at the additional ip-boundary for strategy TEL1.

Each token to be judged consisted of the following components: a beep tone - 1 sec silence - stimulus 1 - 1 sec silence - stimulus 2 - 2 sec response time. Each stimulus pair occurred twice, once in each order. In total there were 42 tokens (3 numbers X 2 orders X 7 pairs) plus three warm-up dummy tokens at the very beginning. All tokens were randomised. The test took 12 minutes in total.

Eight German-native speakers (partly those from the production experiment) were told to evaluate artificially generated spoken phone numbers played over loudspeakers in an office environment. They were asked to indicate immediately after each token whether they preferred the first or the second rendering.

4.2. Results

The results presented in table 2 confirm our hypothesis that "natural" strategies were preferred over the "telecom" strategies. However, the hypothesis that TEL2 (with reduced prosody) is inferior to TEL1 (over-accented prosody) had to be rejected. The comparable rating of the two strategies was due to divided preferences across listeners rather than to inconsistency within listeners.

Table 2: Percentage of preferences of tested strategies (bold) over TEL1,2 (italic); (per cell n=48).

	NAT 1	**NAT 2**	**NAT 3**	**TEL 1**
TEL 1	92%	92%	90%	-
TEL 2	92%	94%	94%	44%

4.3. Interpretations

Informal comments by the listeners support the interpretations that the clear grouping of TEL2 and the natural-like intonation of TEL1 were appreciated, whereas the overall monotony and the lack of a terminal contour in TEL2 and the exaggerated pitch range of TEL1 were felt to be disadvantageous. Obvioulsy, too much pitch variation can be inferior to none, but a balanced intonation seems to be best, which is in line with our hypothesis and the findings in [6]. Deficiencies of the NAT versions feature some rhythmic weaknesses, e.g. a too fast tempo in three-digit IPs. A too *slow* tempo caused by accented syllables throughout and many pauses in the TEL versions (speech rate incl. pauses in syll/sec: 1.9 for NAT1,2 vs. 1.4 for TEL1 and 1.2 for TEL2) could be another reason for their relative inacceptability.

5. Discussion

One immediate impact of our study concerns the automatically spoken phone numbers of telephone inquiry systems. Even if natural, non-manipulated speech is used for the single digits, a quasi non-existent prosody can hamper the information quality so much that some numbers are simply not intelligible. In a stress situation, the "2222222" (car emergency call) with a limited phrasing and accent structure and no terminating intonation reportedly led to a pure guess of how many times "2" had to be dialled.

Also, conventional unrestricted text-to-speech synthesis can benefit from a natural-like prosodic rendering of phone numbers. This is demonstrated in [8], a system which is able to process multiple number formats in an appropriate manner. Prerequisites for a differentiated processing of digits are considered in speech mark-up languages, e.g. [9]. Different individual strategies, different special numbers, but also various strategies in different cultures can be integrated with these tools. This requires a distinction between default and non-default cases. To achieve this, results of our production experiment can help to provide constraints and rules for special cases (section 2.2). The constraints might be ranked differently for different speakers. There could be basic grouping strategy constraints for different length telephone numbers (e.g. 3-2-2 for 7-digit numbers), a sameness constraint ("group consecutive identical digits together"), and a series constraint ("group consecutive ascending or descending digits together"). Once the grouping is done, there have to be rules sensitive for deaccentuation phenomena. This is because the metrical structure of a group may change due to a change in information structure: If the same digit is repeated (as in "387868"), its second and subsequent occurrences represent given information. A given item is likely to be deaccented, which requires a stress-shift. However, a prerequisite for this rule is the occurrence of the digits in nuclear position, which in turn is dependent on the choice of wording strategy: In "38-78-68", a shift of the accent leftwards applies only in a digit-by-digit reading ("DREI acht - SIEben acht - SECHS acht"), but not when read as tens ("acht-und-DREI-ssig, acht-und-SIEB-zig, acht-und-SECH-zig"), since the supposed default pronunciation of tens in German requires prominences on the first digit in each IP, which coincides with the nucleus.

Dialling codes are generally known by speakers and are presumably treated holistically. Thus, speakers rarely deviate from the predefined wording. For technological applications "meaningful" digit combinations like dialling codes or other common numbers should be considered.

There are anecdotal reports that well-known numbers are not recognised (immediately) if they are produced with a different grouping and/or a different wording from the representation in one's own mind. Additionally, people pointed out that they deliberately search for "Eselsbrücken" (mnemonic aids) which have a strong influence on grouping, wording and accenting. This also explains the wish for numbers which are easy to remember as in our special cases such as "7774777" or "1234353".

It can be assumed that people do not necessarily structure a number (especially a phone number) that has to be stored in long-term memory in the same way they structure a number for short-term memory. Recall experiments with auditorily presented lists of 9-digit numbers [10] proved that a 3-3-3 grouping of the listed items led to a better performance than no grouping. This effect probably plays an important role for production strategies and perceptual preferences. Moreover, first and especially final items in a group are better stored in the short-term memory than medial ones. The metrical pattern we found in our data is structured so that the group-final digit is metrically represented with a strong beat, and the group-medial digit with a weak beat. It can be hypothesised that retention of list items in short-term memory is not only dependent on the position in the group but also on the metrical weight. This would mean that initial but weak items in groups of two instead of three as in [10] are recalled worse.

In contrast to ordinary spoken language, numbers do not contain redundancy, i.e. listeners have to decode the information of *all* items to understand the whole message. Thus, there is a contrast between information load and prosodic coding. Of course, numbers are omnipresent, occurring as bank account numbers, ID numbers, sums of money etc. This study on the prosody of telephone numbers integrates questions of wording, grouping and accenting. It can be regarded as a contribution to research on spoken numbers in general, which surely play an important role in dialogue systems.

6. References

[1] ISCA Speech Prosody Special Interest Group http://groups.yahoo.com/group/sprosig/messages

[2] Sonderdruck DIN 5008. Deutsches Institut für Normung e.V. Berlin (ed.). Berlin, Wien, Zürich, 1996.

[3] Ladd, D.R., *Intonational Phonology*. CUP, 1996.

[4] Grice, M., S. Baumann & R. Benzmüller (to appear). German ToBI. In Jun, Sun-Ah (ed.) *Prosodic Typology and Transcription: A Unified Approach*.

[5] Wunderlich, D., Der Ton macht die Melodie - Zur Phonologie der Intonation des Deutschen. In Altmann, H. (ed.) *Intonationsforschungen*. Tübingen, 1-40, 1988.

[6] Waterworth, J.A., Effect of intonation form and pause duration of automatic telephone number announcements on subjective preference and memory performance. *Applied Ergonomics* 14(1):39-42, 1983.

[7] http://tcts.fpms.ac.be/synthesis/mbrola.html

[8] http://www.bell-labs.com/project/tts/digtalk.html

[9] http://www.bell-labs.com/project/tts/sable.html

[10] Frankish, C., Intonation and auditory grouping in immediate serial recall. *Applied Cognitive Psychology* 9 (Special Issue): 5-22, 1995.