

# LISTENING TO LISTS: STUDYING DURATIONAL PHENOMENA IN ENUMERATIONS

Hannes Pirker and Stefan Kramer

*Austrian Research Institute for Artificial Intelligence (ÖFAI), Vienna, Austria*

## ABSTRACT

A study on durational phenomena in list like enumerations in German is presented. Due to its highly structured and uniform nature this rather specialized utterance type seems especially well-suited for investigating principles of the rhythmical organization of speech. A corpus extracted from radio weather reports is used in order to investigate phenomena like prefinal lengthening and effects of isochrony and prominence. In addition to studying durational phenomena with standard statistical methods, the data also was analyzed using Structural Regression Trees (SRT), a machine learning algorithm.

## 1. INTRODUCTION

The temporal organization of speech has been an issue amongst phoneticians for decades. Numerous phonetic studies dealt with the investigation of (segmental) duration. Typically these works dealt with the isolated influence of single factors on duration and established a whole inventory of possible influences that span from segmental context to linguistic factors (see, e.g., [6] for a summary of factors and [5] for discussion and further references).

In spite of this rich phonetic tradition the specification of segmental duration is still an issue in speech synthesis. In this context not only the factors but also their exact *quantitative* effects have to be determined. Because of the number of factors and their complex interactions this becomes a very difficult task. Nowadays large corpora and statistical approaches are used to tackle this problem. Nevertheless, all these sophisticated methods – including neural networks and machine learning techniques (for German e.g. [5], [13], [14],[10]) – have to struggle with the problem of data sparsity produced by the misbalance between the vast number of possible value combinations and the restricted size of available corpora.

Thus, it is necessary to integrate as much phonetic knowledge as possible in order to sustain the automatic methods and to restrict the number of hypotheses generated by these methods [15].

In this spirit the work presented in this paper is an exploratory study in the context of an upcoming large scale study on segmental duration in Austrian German. Besides the original interest in the investigation of list like enumerations in Austrian German it was used for the development and evaluation of inspection tools, methods of representation, encoding, and processing of the data.

Enumerations are certainly an interesting domain for studying duration. Not only because of their inherent rhythmical organization, but also because of their rigid structuring. The utterances display controlled uniform syntax and semantics in combination with a small vocabulary. This results in the production of uniform and thus more easily comparable utterances.

Also, these enumerations are broadcasted via radio several times a day. This allows for the simple collection of a highly controlled corpus of multiple professional speakers.

From a practical point of view the study also aims for an improvement of the prosodic quality of enumerations within our Concept-to-Speech generation system VieCtoS [1].

## 2. WEATHER-LISTINGS

### 2.1. General description

The enumerations analyzed in this study are mostly standardized listings of weather data for Austria's provincial cities which are broadcasted almost every other hour by the national public radio.

The following example transcript exemplifies the structure of such items:

- |        |                           |               |                      |
|--------|---------------------------|---------------|----------------------|
| (11)   | Wien                      | heiter        | drei Grad,           |
|        | <i>Vienna</i>             | <i>clear</i>  | <i>three degrees</i> |
| (12+3) | St.Pölten und Linz        | heiter        | sechs,               |
|        | <i>St.Pölten and Linz</i> | <i>clear</i>  | <i>six,</i>          |
| (14)   | Salzburg                  | bewölkt       | sechs,               |
|        | <i>Salzburg</i>           | <i>cloudy</i> | <i>six,</i>          |
|        | ...                       |               |                      |
| (19)   | und Klagenfurt            | heiter        | drei Grad.           |
|        | <i>und Klagenfurt</i>     | <i>clear</i>  | <i>three degree.</i> |

### 2.2. Structural issues

These enumeration are organized into *lines*. Each line consists of three blocks: a city's name, information on the clouding, and a temperature value. Usually there is one line for each capital, resulting in 9 lines for a whole *message*.

The structure of the lines is fixed as well as the strict ordering of the capitals. If adjacent cities share the same values, these lines can be conflated as in (12+3). In our corpus less than 10% of the lines contain such a coordination of city names.

Clouding information is highly uniform reflecting both standardization of the message and meteorological conditions. From a total of 17 different types found in the corpus only 10 are observed on a regular basis and almost 50% of all lines in the corpus contain either /heiter/ (*clear*) or /bewölkt/ (*cloudy*).

Temperature values may be accompanied by the word /Grad/ (*degree*) which is always used in the first and last line but is optional elsewhere. Thus, approximately 40% of all lines are terminated with this word.

Very few deviations from this strict scheme are observed, e.g., less than 2% lines bear information on extraordinary wind speeds. This extra information is always appended to the end of a line.

### 2.3. Prosodic structure

The prosodic structure of the utterances is also highly uniform and reflects the semantic chunking within the lines. Especially the intonation follows a highly standardized rise-fall-rise contour which in terms of the prosodic annotation scheme G-ToBI [4] can be described as follows:

H\* L\* H\* H-L%  
Wien heiter drei Grad

Some speakers frequently insert pauses after the city name and after the cloud information. This additional prosodic chunking nevertheless usually does not alter the overall intonation contour.

### 3. THE CORPUS

#### 3.1. Overview

The corpus was recorded from radio broadcasts over a period of several months and comprises approximately 27 minutes of speech from 21 speakers. It contains 92 evaluable messages, 770 lines, 3394 words, and about 18000 phonemes.

#### 3.2. Annotation

The data was transcribed and word boundaries were annotated manually using UCL's *Speech Filing System SFS* [16]. Prosodic information is encoded rather rudimentary: only explicit pauses are marked with a special label /p/ for minor pauses and /b/ for breaks. Segmental boundaries were labelled using an automatic alignment procedure [10]. This automatic labeling has not been completely evaluated yet and thus will not be used in the present study.

#### 3.3. Normalization

In order to compensate for differing speaking rates and allow for a more comprehensible representation normalized durations were used in addition to the original word length. As the city names uniformly appear in all messages for each message their average length per syllable is calculated and used as a normalizing factor for all words within that message. Thus a normalized length of 1.4 means that the length per syllable of this item is 1.4 times the average length per syllable of all 9 cities in that message.

#### 3.4. Representation

Information about the corpus is stored in a relational database that holds information about the speaker and lexical information on the words such as syllable structure, word class and semantic tags. This database is simply realized in Prolog (see [3] for a similar approach). From this Prolog database ASCII files are produced which are usually further manipulated using standard UNIX-tools.

#### 3.5. Analysis tools

In order to facilitate the fast and interactive exploration of a hypothesis a graphical user interface for producing "clickable histograms" (such as Fig.5 or Fig.6) was developed. In addition to visualizing the data and calculating means and standard deviations clicking on a bar produces the sound associated with that token. Thus an immediate auditive inspection of conspicuous items can be performed.

## 4. DURATIONAL PHENOMENA

#### 4.1. General issues

In this section some analyses of durational phenomena are presented which deal with the inspection of effects on word duration. We were interested whether this corpus that only displays minor lexical variation would allow for an investigation of duration without recourse to a finer grained labeling.

#### 4.2. Prefinal Lengthening

It is uncontroversial that major prosodic boundaries have a lengthening effect on preceding items. In our corpus prosodic information was not labelled in much detail. Only pauses that were perceived as phrasal boundaries were marked with a label /b/ while other pauses were labelled /p/. We thus checked to which extent prefinal lengthening can be identified on the basis of word length information and this rudimentary labeling.

It was no surprise that, e.g., city names followed by a /b/ display a lengthening tendency. A more detailed survey of the effect

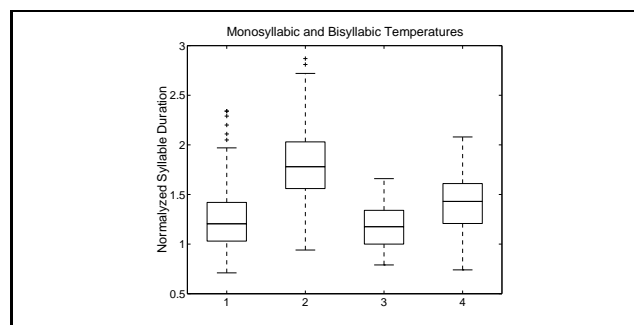


Figure 1: Comparing effects of prefinal lengthening on temperature values. Monosyllabic: (1) followed by /Grad/ (N=240) (2) phrase final (N=309) Bisyllabic: (3) followed by /Grad/ (N=40) (4) phrase final (N=93)

was performed by inspecting the variation in the duration of temperature values. The specification of temperature always is situated at the end of a line and thus at the end of an intonation phrase. Nevertheless some interesting variation exists because 40% of the lines end with the optional word /Grad/. Thus it is possible to compare temperatures that are phrase final proper with items that are separated from the boundary by a single syllable that functions as a well defined spacing element.

In Fig.1 the influence of the /Grad/ on mono- and bisyllabic words is displayed. In terms of absolute word length the presence of /Grad/ triggers a shortening of about 100ms independent of the number of syllables of the temperature value. In terms of normalized length per syllable this results in a decreased effect for polysyllabic words. This uniform absolute additive effect – also observable in words comprising four syllables – may be an evidence that the main scope of prefinal lengthening lies on the ultimate syllable.

In order to check whether the presence of an explicit pause /b/ is a not only sufficient but also necessary condition for triggering the lengthening, the variation of /Grad/ itself was analyzed. /Grad/ always is situated at the end of a line. Nevertheless sometimes subsequent lines are connected without any intervening pause. It was studied, whether the duration of /Grad/ changes in these cases. It also was studied whether the duration of /Grad/ is different at the very end of a whole message (19) because an additional slowing down was suspected in this special position. Fig.2 shows that no significant variation was found between standard line breaks (left-most column), message terminating positions (middle column) and positions where line breaks lack a pause (right column).

The results indicate that – at least in the case of enumerations – overt pauses are unnecessary for indicating major prosodic boundaries.

#### 4.3. Quantitative influences

When listening to enumerations, they are often perceived as rhythmical. Though we do not aim for a definition of acoustic correlates of rhythm – which is a rather controversial subject – we tested some hypotheses on isochrony effects. Isochronic tendencies should be observable by inspecting prosodic constituents with different quantities of segmental material.

When comparing mono- and polysyllabic words the latter typically display shorter duration per syllable. Nevertheless the comparison on this basis is still problematic because timing differences also can be caused by differing syllabic weights.

Therefore we aimed at the analysis of cases where identical

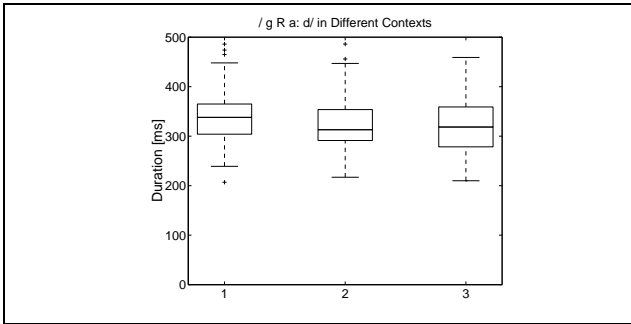


Figure 2: Comparing the length of /Grad/. (1) at linebreak /b/ (N=181), (2) at message end (N=84), (3) others (N=40)

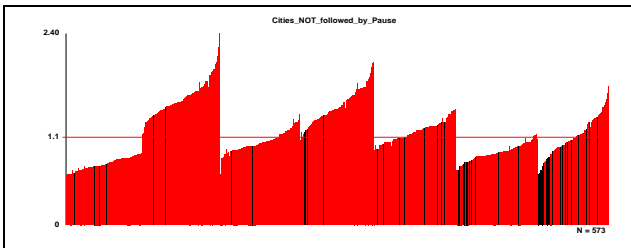


Figure 3: City token not followed by a /b/. Token which are positioned in front of /und/ (*and*) are printed black.

words are embedded in constituents of different size. In the enumeration corpus this can be observed in cases of coordinated city names like in example (12+3). As a hypothesis, city names should be significantly shorter when uttered within such complex nominal phrases.

(12+3) *St.Pölten und Linz* heiter sechs,  
*St.Pölten and Linz* clear six,

Though this is a rather simple hypothesis, it is not trivial to test this condition. There is a strong tendency to place a major prosodic phrase boundary after /Linz/. Thus the whole prosodic structure of the line is altered and the second city undergoes prefinal lengthening which conflicts with the hypothesized shortening. In order to avoid influences from prefinal lengthening effects only words coming first in the coordinated structure were considered in Fig.3. The duration distribution of some cities is displayed where items from the initial positions of coordinated phrases (black) can be compared to all other token of that city name that are not followed by a phrase boundary /b/.

In Fig.4 a direct comparison of durations of city names found

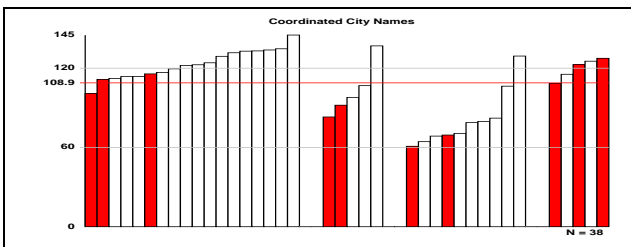


Figure 4: Selection of city items from coordinated phrases (grey: first element in coordination, white: second element)

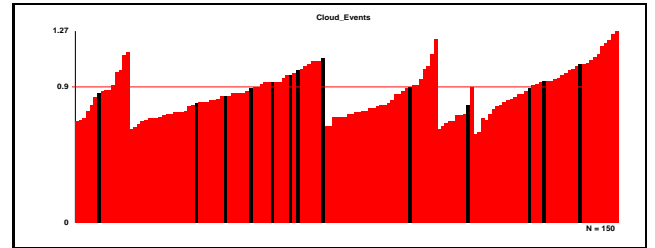


Figure 5: Normalized durations of cloud tokens (e.g.,/clear/, /cloudy/, /rain/) of speaker jwh: items which are explicitly labelled as accented are printed black.

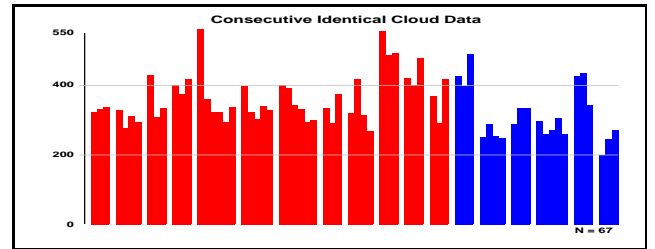


Figure 6: Comparing absolute durations of identical cloud-tokens found in adjacent lines.

in both first and second positions (i.e., in front or after the coordinating /and/) is performed. Though the selection of second items (white) was strictly restricted to words not followed by a /b/, the figure indicates that there is still a tendency for longer duration of the second item. We rate this as additional evidence that checking for the existence of /b/ is not sufficient for indicating phrasal boundaries.

#### 4.4. Prominence

Due to its uniform information structure and intonational realization the enumeration corpus does not include many explorable examples for analyzing the prominence lending function of duration and intonation. Nevertheless, because of their relative low frequency these cases provoke even more interest.

Some speakers deviate from the standard stylized rise-fall-rise pattern in cases like (14). In order to highlight the only new information in the line the cloud information is assigned a prominence lending L+H\* pitch accent while the temperature value is deaccented.

(12+3) *St.Pölten und Linz* heiter sechs,  
*St.Pölten and Linz* clear six,  
 (14) *Salzburg* **bewölkt** sechs,  
*Salzburg* cloudy six,

For a part of the corpus these special pattern were labelled. Fig. 5 shows on how the duration of these extra prominent cloud items (black) compare to other items of their type. It shows that a tendency for longer durations is observable.

We also wondered whether in parallel to the highlighting of new information also deaccenting or reduction of repeated old information can be observed. Therefore messages where the very same cloud information was uttered more than twice in series were extracted. The results in Fig.6 show that this hypothesis did not hold. Also, the auditive evaluation of intonation and duration of these cloud items did not show any perceivable deaccenting tendencies.

## 5. USING A MACHINE LEARNING METHOD

### 5.1. Motivation

As it was demonstrated in section 4.3 the problem of possibly complex interactions between factors can be observed even in relatively simple investigations.

By using a machine learning method for the task of modeling duration and comparing the results with the insights gained from manual inspection both methods can be evaluated. In order to be able to “learn from the learner” the inspectability of the results was an important factor.

### 5.2. Methodology

The method of Structural Regression Trees (SRT) ([1]) was used, which constructs theories for the prediction of numerical values from examples and relational background knowledge. SRT offers the full power and flexibility of first-order logic, provides a rich vocabulary for the user, and produces trees that are both good predictors and interpretable.

SRT can be viewed as an “upgrade” of CART (*Classification and Regression Trees* [2]) to handle relational data. Like CART, SRT makes use of error-complexity pruning for regression problems. The trees in the sequence of pruned trees are evaluated by cross-validation.

Different descriptor-sets were tested. The resulting SRT theories were quantitatively evaluated by means of 10-fold cross-validation, results were summarized in terms of RMSE (the root of the mean squared error of the generated theory on unseen cases) and  $r$  (the correlation coefficient).

### 5.3. Results

The overall best descriptor-set yielded RMSE=0.26 and  $r=0.76$ . Note that the RMSE of the default theory always predicting the average duration is 0.41. So, the result is distinctly better than the default theory.

Moreover, the resulting theories were inspected qualitatively in order to check their plausibility and view the computer’s “insights” in the light of priorly performed investigations on durational effects [2].

To just give an impression on theories SRT produced:

```
duration(A, B) :-
  ( sem_cat(A, temp) ->
    ( succ(A, C),
      sem_cat(C, bound) ->
        ( nr_sylls(A, 1) ->
          B is 1.79845
        ; B is 1.10538
        )
      ; B is 1.25703
    )
  )
  ...
```

In the example, we make use of the operator (+P -> +Q ; +R), meaning “if  $P$  then  $Q$  else  $R$ ”. When applying a theory, we interpret this operator in a way such that not only the first solution of  $P$  is explored.

$B$  is the predicted length in normalized form (1.0 is the average syllable length of the city-names of a message). The fragment deals with the prediction of the duration of temperature values ( $\text{sem\_cat} = \text{'temp'}$ ). As can be seen, SRT “detected” the preboundary lengthening effect (check whether the successor  $C$  is of type ‘bound’) and is sensitive for the number of syllables per word ( $\text{nr\_sylls}$ ).

On the other hand, “higher level” features (e.g. a ‘last\_line\_predicate’ in order to test for possible ritardando

effect in lines such as (19) or a test on coordinated city-names such as in (12+3)) turned out to be mostly ignored.

## 6. SUMMARY

The work presented in this paper has been conducted as a pilot study for an investigation on segmental duration and provided valuable insights on both the durational properties of list enumerations in German as well as on the possibilities and limitations of methods and encodings used.

The results gained from the analysis will help to tailor the set of factors used in the future. It also was demonstrated that prosodic labels have to be integrated in a more sophisticated way.

The adaption of the machine learning task proved to be straightforward and the first results appear promising both in terms of accuracy and inspectability.

## ACKNOWLEDGMENTS

This work has been sponsored by the *Fonds zur Förderung der wissenschaftlichen Forschung (FWF)*, Grant No. P13224. Financial support for ÖFAI is provided by the Austrian Federal Ministry of Science and Transport. Parts of this study have been performed during a short time scientific mission at the IKP-Bonn which was sponsored by the COST 258 action. Special thanks to Johannes Matiasek and Karlheinz Stöber for sharing their software and for ongoing support.

## REFERENCES

- [1] Alter K., Buchberger E., Matiasek J., Niklfeld G., Trost H.: VieCtoS: The Vienna Concept-to-Speech System, in Gibbon D. (ed.): *Natural Language Processing and Speech Technology*, Mouton de Gruyter, Berlin, 1996.
- [2] Breiman L., Friedman J.H., Olshen R.A., Stone C.J.: *Classification and Regression Trees*, Wadsworth International Group, Belmont, CA, The Wadsworth Statistics/Probability Series, 1984.
- [3] Draxler C.: *Introduction to the VerbMobil-PhonDat Database of Spoken German*, Practical Applications of Prolog Conference 95, Paris, 1995.
- [4] Grice M., Reyelt M., Benzmüller R., Mayer J., Batliner A.: Consistency in Transcription and Labelling of German Intonation with GToBI, Proc. of ICSLP’96, Philadelphia, PA, pp.1716-1719, 1996.
- [5] Huber K.: *Messung und Modellierung der Segmentdauer für die Synthese deutscher Lautsprache*, Zürich: ETH, Dissertation ETH Nr. 9535, 1991.
- [6] Klatt D.H.: Linguistic uses of segmental duration in English: acoustic and perceptual evidence, *Journal of the Acoustical Society of America (JASA)*, Vol. 59, Nr. 5, pp. 1208-1221, 1976.
- [7] Kramer S.: *Structural Regression Trees*, Proc. of the 13th National Conference on AI (AAAI-96), Menlo Park, 1996.
- [8] Moebius B., Santen J.van: *Modelling Segmental Duration in German Text-to-Speech Synthesis*, Proc. of ICSLP’96, Philadelphia, PA, Vol.4, pp.2395-98, 1996.
- [9] Pirker H.: *Dauerphänomene in listenförmigen Aufzählungen*, in Proceedings of DAGA98, Zürich, 1998.
- [10] Stöber, K.: *Additional use of phoneme duration hypotheses in automatic speech segmentation*, in Proceedings of ICSLP’98, Sydney, 1998.
- [11] Riedi M.: *A Neural-Network-Based Model of Segmental Duration for Speech Synthesis*, in Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH 95), Madrid, Spain, Vol.1, pp.599-602, 1995.
- [12] Riedi M.: *Modeling Segmental Duration with Multivariate Adaptive Regression Splines*, in 5th European Conference on Speech Communication and Technology (EUROSPEECH 97), Rhodes, Greece, ESCA, Vol.5, pp.2627-2630, 1997.
- [13] Santen J.van: *Prosodic Modeling in Text-to-Speech Synthesis*, in 5th European Conference on Speech Communication and Technology (EUROSPEECH 97), Rhodes, Greece, ESCA, Keynote Speech, KN 19, 1997.
- [14] University College London: *Speech Filing System SFS* <http://www.phon.ucl.ac.uk:80/resource/sfs.html>