# Have a break! Modelling pauses in German Speech.

Jens Apel, Friedrich Neubarth, Hannes Pirker & Harald Trost
ÖFAI, Austrian Research Institute for Artificial Intelligence
1010 Wien, Austria
jens.apel@gmx.net, {friedrich, hannes}@oefai.at, harald@ai.univie.ac.at

**Abstract**

In the present paper we investigate the distribution of pauses and to some extend boundary tones in natural speech. In particular, we examine the pause duration at various positions depending on features like Part-of-Speech (POS) of the word before the pause or position of the pause in the sentence. The results of this explorative study were used to train a CART-model in order to reach a better performance of pause duration control in a current text-to-speech synthesizer.

## 1. Introduction

Modelling the duration of speech events is a central task in speech synthesis. At first sight this task is mainly concerned with the duration of phonemic segments (sometimes intermediated by higher level categories such as the syllable, see Zellner, 1994). Rule-based speech synthesis systems of an earlier generation were equally concerned with pauses occurring between phonetic segments, whereas more statistically oriented systems used to edge out the explicit modelling of the duration of pauses. The problem of predicting pauses in speech synthesis can be tackled in two seaparate steps: first it has to be decided whether the insertion of a pause is appropriate in a certain position, and then the actual duration of this pause has to be determined.

What we want to show in this paper is how both of these goals, modelling occurrence and duration of pauses can be achieved rather easily by using an existing speech synthesis system with appropriate interfaces (MARY, cf. Schröder & Trouvain 2003) and an annotated corpus of natural speech from a single speaker (The Vienna Prosodic Speech Corpus, cf. Neubarth 2000). Before discussing the details of our approach it is worthwhile to reconsider the linguistic and psychological background of having a break.

Zellner (1994), when discussing classification schemes for pauses in speech, distinguishes two different traditions: (1) a physical/linguistic classification, and (2) a psychological/psycholinguistic classification. The physical/linguistic classification differentiates between intra-segmental pauses and inter-lexical pauses. Intra-segmental pauses appear inside words and are due to the occlusion of the vocal tract in normal speech production. In contrast, inter-lexical pauses tend to appear only between words. In the psychological/ psycholinguistic classification pauses are divided into silent pauses and filled pauses. Silent pauses relate to sections in the speech material where no speech signal appears. Filled pauses on the other hand are voiced sections in the speech material such as drawls, repetitions of utterances or false starts.

The reasons for producing pauses in speech are manyfold. Zellner (1994) put forward that the occurrence of pauses is strongly speaker-dependent. Therefore, attributes like respiration, muscular tone, or the articulation rate trigger the number of pauses. It is also argued that pauses mirror cognitive activity. Goldman-Eisler (1968) indicate that a speaker might produce a pause in order to plan what he wants to say.

In a more linguistically oriented view, pauses also act as a kind of “beacon” for speech. The speaker splits up complex utterances into smaller chunks or segments in order to improve comprehension, or simply to express the structuring or an utterance by prosodic means. Finally, the occurrence and length of pauses depends on the communicative situation. So, for example, it is quite conceivable that if a speaker

has to perform speech in a noisy environment, she will insert more and more significant = longer pauses and hesitations than if she has to read a text under more quiescent conditions.

As just mentioned above, pauses are an imortant prosodic cue for marking boundaries in order to improve comprehension. However, it has to be emphasised that not only pauses but also pitch changes, pre-boundary lengthening, declination reset or intensity shaping are used as prosodic means for the structuring of utterances. In a ToBI oriented approach (cf. Beckmann & Ayers 1994) all these factors are considered as physical manifestations of abstract break indices. If we think about prosodic structuring, the concept of having a break appears to be the more fundamental one. Therefore, beside the realisation of pauses, we will also consider the occurrence of boundary tones in the present study.

## 2. Aims of the present study

In the present study we investigate how inter-lexical, silent pauses are distributed in natural speech. Intra-segmental pauses as a result of plosive formation as well as pauses occurring at the end of an utterance, which may have some significance for the higher structuring at the textual level, are not considered here. The basis or our model is a corpus of read speech, which is compared to the pauses predicted by the MARY text-to-speech system. Apart from determining whether a certain structurally determined break index is to be actually realised as a pause, special focus is put on the duration of these pauses. The crucial question is how the length of a pause depends on the surrounding structural properties, such as Part-of-Speech of the word before the pause, different punctuation signs or position of the pause in the sentence. This information is then used to guideline the automatic training of a prediction-model. The main goal behind is to develop a method to improve the performance of pause control in a text-to-speech synthesizer. We investigate which features should be used for such a model and how it should be trained to reach a good performance. We will use modelling by the CART method on both tasks, first training on a Boolean value in order to decide whether a pause has to be realised, and in a second step determining the duration of the pause. In the following we will describe the specific components used in the present study.

## 3. Components

### 3.1 The Vienna prosodic speech corpus

The corpus was established during a project primarily concerned with the modelling of the duration of phonetic segments.[1] It comprises approx. 50.000 phones (1.2 hours of actual speech) of Standard Austrian German spoken by a single non-professional speaker. The corpus consists of 3 different types of read material for different purposes of analysis:

- phonetically balanced text: the classical "Nordwind und Sonne" and "Buttergeschichte" and 300 isolated sentences
- material where the information-structure is controlled for the analysis of the relation between focus structure and prosody.
- 22 contiguous texts from newspaper.

The segmentation and annotation of phonetic labels was done semi-automatically and was also controlled manually. The corpus is also annotated for intonation (ToBI-labels, Fujisaki accent and phrase commands), prominence (manual labels for each syllable) Information on linguistic structure (feet, syllable structure, etc.) is dynamically derived and stored in the data-base.

Since the corpus was designed for a slightly different purpose we could not use the full corpus for our experiments. Due to the short length of many utterances there are not always chances to have a rest, and also reading style inhibits the insertion of pauses as compared to for example free dialogues. After excluding those irrelevant samples we ended up with a subcorpus of 347 sentences with potential pause candidates actually used to train our models.

[1] Segmental Duration in German Speech (SpeeDurCont). This project has been sponsored by the FWF, Grant No. P13224. The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry of Education, Science and Culture.

### 3.2 The MARY Text-to-Speech system

Modelling in speech synthesis requires an automatically generated structural description of an utterance. Since even the boundary tone labels in the Vienna Prosodic Speech Corpus are labeled manually (and therfore rely on subjective judgement) and does not contain break indices, we have to use a different procedure in order to obtain the relevant features for modelling. The MARY-system (Modular Architecture for Research on speech Synthesis) offers an environment perfectly suitable for this purpose. The MARY Text-to-Speech Synthesizer is a System developed at the language technology lab of the German Research Center for Artificial Intelligence (DFKI) and at the Institute of Phonetics at Saarland University. It is organised in a modular way and its internal markups can be accessed as XML at every stage of computation. Of particular interest is the procedure how this system sets pauses within a given text. The MARY system assigns three relevant indices for potential breaks (break index 3, 4 and 6). These indices map onto boundary tones (according to the GToBI annotation scheme, cf. Grice & Baumann 2002), but they also directly correspond to pauses inserted into the string of phonetic symbols. The duration of each pause category is fixed and corresponds to an empirically evaluated mean duration [(3) 100ms, (4) 300ms and (6) 410ms].

The relevant break indices are selected in dependency of the surrounding types of words and punctuation signs. Break index 3 is applied to a position in a sentence after the Vorfeld (i.e. in a matrix clause before the first finite verb in the sentence) and before conjunctions (i.e. and/or). The rule setting break index 3 after the Vorfeld is suppressed if a personal pronoun occupies the position before the verb or if a comma is positioned after the finite verb. Break index 4 is set after punctuations like brackets, commas or quotation marks. Break index 6 is used only for the end of a sentence and is not of concern in this study. As for boundary tones, break index 3 is linked with the boundary tone H-, break index 4 with H-% and break index 6 with boundary tone L-%. Consequently every break index is linked with both a pause and a boundary tone.

Furthermore, the MARY system returns various features concerning the words and syntax of the sentences. It returns whether the word could be retrieved from the lexicon or not (« g2p_method »), the Part-of-Speech (POS) of every word in the sentence and the syntactic attachment of every word to the previous word (syn_attach), which reflects the chunk structure of the given utterance. The break indices, as well as the type of syntactic phrase, tones and accents are calculated on the basis of this information.

The « g2p_method » feature (grapheme-to-phoneme method) has the following values : "lexicon" (i.e. the pronunciation of a word is directly retrieved from the lexicon), "compound" (i.e. pronunciation of two or more words from a lexicon are combined with each other), "mtu" (multi-token unit, i.e., numbers or abbreviations are treated as whole entity) and "rules" (i.e. the pronunciation of the word is not found in the lexicon and is therefore built up by grapheme-to-phoneme rules). The « POS » feature is based on the Stuttgart/Tübingen Tagset (STTS, cf. Schiller et al. 1999). The « syn_attach » (syntactic attachment) has the values: '0' (current item is part of the same lowest-level syn_phrase as preceding one), '1' (no attachment to any previous items), '2' (the current item is neither part of the present syntactic phrase nor end of it), '+' (current item is attached one level higher than the preceding one), '-' (current item is attached one level lower than the preceding one), 'u' (current item is attached 2 levels higher than the preceding one) and 'd'(current item is attached 2 levels lower than the preceding one).

In principle, the features described above are used for our training of statistical models. However, the break indices (3 and 4 being relevant) have a special status insofar as they are used as primary indicators of actual pauses. The whole architeture of our pause generation model resembles a cascaded algorithm, the first step being a rule based selection of candidates (corresponding to the break indices assigned by the MARY system), and two following steps based on statistical methods: one determining whether a pause is realised and a second one determining the duration of the positively selected candidates.

## 4. Experimental Design

Within the 347 sentences selected from the corpus the speaker actually made 350 pauses within the utterance. It has to be mentioned that the pauses in the corpus were also manually labelled. They are defined as the absence of any speech signal within a subjectively determinable span of time, not being related to pauses within plosive segments (intra-segmental pauses).

The same sentences were fed into the MARY Text-to-speech system, which returned syntactic features mentioned in section 3.2.

For the present experiment only those pauses were included in the modelling, which were realised both by the speaker and the MARY system. See the next section about precision and recall ratios. Imposing this further selection on the corpus reduces the number of pauses again to a number of 280.

This procedure was performed in order to correlate the duration of pauses produced by the speaker with features provided by the MARY system for the position of the specific pause candidate.

## 5. Evaluation Results

### 5.1 Matching pauses

As shown in Table 1 the natural speaker made 350 pauses. In contrast, the MARY system calculated 610 pauses.

| | Pauses | Tones | Intersection of Pauses & Tones | Set union of pauses & Tones | Tones without Pauses |
|---|---|---|---|---|---|
| MARY-System | 610 | 610 | 610 | 610 | -- |
| Natural speaker | 350 | 499 | 350 | 571 | 221 |
| MARY & natural speaker set union | 280 | 350 | 280 | 404 | 124 |
| Recall | 80.00 % | 70.14 % | 81.29 % | 70.75 % | -- |
| Precision | 45.90 % | 57.38 % | 37.05 % | 66.23 % | -- |
| F-score | 58.33 % | 63.12 % | 50.90 % | 68.42 % | -- |

**Table 1 :** Number of pauses and tones. [2]

[2] In the MARY-system prosodic phrase boundaries are always realised with both, a boundary tone and a pause. Therefore, the field «Tones without pauses» remains undefined.

All together 280 pauses were made by both the natural speaker and the MARY system. This leads to a recall value of 80.00 %, but also to a relatively low precision rate of 45.90 %.

One reason for this low precision rate lays in the fact that the MARY system does not include the position of a pause relative to the beginning of a sentence as a dependent variable. It can be observed that the precision rate differs significantly when different regions within the sentence are analysed separately, i.e. when pauses occuring earlier in the sentence are distinghuished from those occuring later in the sentence.

Of 290 pauses set in the sentences by the MARY system before the 7th word, only 74 match with the natural speaker. In contrast, of 320 pauses set by the MARY system after the 7th word, 206 match with the natural speaker. The statistical analysis shows that these 2 populations (i.e. the population of pauses before the 7th and the population of pauses after the 7th word in a sentence) differ significantly with $p<.05$. The threshold value (7th word) was obtained heuristically.

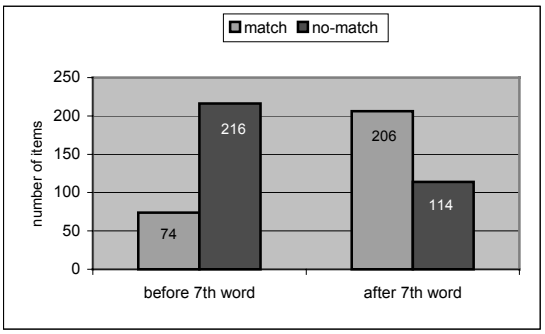


**Figure 1** Diagram of the number of words preceding the pause.

Similar influences of quantitaive factors were found by Zwonik and Cummins (2003), who show that the probability of a pause being shorter than 300ms rises if both the phrase preceding and following pause are short. In the present study we could not replicate the same effect for the following number of words or syllables. Another reason for the low precision rate is the fact that the natural speaker only rarely inserts a pause after the Vorfeld. From

165 candidates in this position, he realises a pause in only 24 instances, which results in 141 mismatches with the MARY system (compare Figure 2).

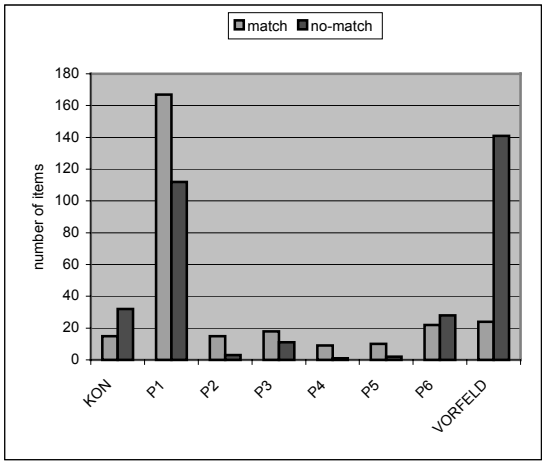


**Figure 2 :** Diagram showing the motivation for the MARY system to set a break index.
KON = Before conjunctions; P1 = « , »;
P2 = « : »; P3= « - «; P4= « ) »; P5= « ( »;
P6 = «"»; VORFELD = Vorfeld;

Other factors like Part-of-Speech of the word before the pause, type of syntactic phrase before the pause or the number of words after the pause did not show such clear results. On the other hand, the natural speaker inserts pauses where the MARY system does not assign a break index. One of the reasons for these unmatched pauses is that the chunker of the MARY system does not discern between prepositional phrases directly attached to the previous noun phrase and prepositional phrases attached higher in the sentence structure. In cases where the *Vorfeld* is occupied by a complex phrase consisting of a noun phrase plus additional prepositional phrases, the MARY system fails to assign a break. On the other hand, the length of the phrase in the *Vorfeld* encourages the natural speaker to insert a pauses at this location.

As mentioned in the introductory section, pauses are not the only indicator for phrase breaks. Boundary tones appear even more eagerly at an abstract phrase break. For a given break index (higher than 3) MARY always assigns both, a boundary tone and a pause. As the data for the natural speaker shows, pauses always co-occur with boundary tones. In contrast, boundary tones can stand in isolation without a pause. The data of the present study shows (see Table 1) that in addition to pauses with boundary tones, the natural speaker also has boundary tones without a pause at positions where the MARY system detects a phrase break.

Consequently, the F-score of the set union of pauses and tones is much better than the F-score of only pauses (68.42 % vs. 58.33%). This indicates that the convention implied by the MARY system to include both a pause and a boundary tone at a phrase break is not supported by our data. Instead, in 221 cases, the natural speaker has a boundary tone without inserting a pause.

## 5.2 Duration of pauses

Having determined the occurrence of a pause in a given environment, the question arises how long the duration of such a pause will be. As already presented the MARY system provides information about various features used in the internal processing. As a first step, let us analyse the environment of pauses using the following features:

- G2p-method of the word before the pause
- G2p-method of the word after the pause
- The POS of the word preceding the pause
- The syntactic attachment of the word preceding the pause
- The syntactic phrase preceding the pause
- The number of words preceding the pause
- The number of syllables preceding the pause
- The number of words after the pause
- The motivation for the MARY system to set a break index

In Table 2 the duration of the pauses are shown according to the features listed above. The differences between the factors of the features g2p-method of the word after the pause, POS, syntactic-phrase, syntactic attachment, number of words before the pause, number of syllables before the pause and the motivation for the MARY system to set a break index is significant with a level of significance of $p < .05$ (see also Figure 3 - Figure 9).

| Features | Subfeatures | Mean Dur / SD |
|---|---|---|
| G2p-method (word before pause) | Rules (rules)<br>No-Rules (lexicon,mtu, compound) | 322.46 ms (187.33 ms)<br>292.22 ms (192.33 ms) |
| G2p-method (word after pause) | Rules<br>No-Rules | 177.38 ms (171.99 ms)<br>302.26 ms (190.85 ms) |
| Part of Speech | S1<br>S2<br>S3 | 387.23 ms (175.56 ms)<br>304.46 ms (198.36 ms)<br>237.46 ms (176.42 ms) |
| Syntactic Attachment | A0 (0)<br>A1 (1)<br>A2 (+)<br>A3 (-) | 260.39 ms (186.84 ms)<br>335.35 ms (189.12 ms)<br>385.12 ms (162.29 ms)<br>199.34 ms (258.26 ms) |
| Syntactic – Phrase | Y1<br>Y2 | 214.47 ms (165.96 ms)<br>335.25 ms (191.31 ms) |
| Break index | Breakindex 3<br>Breakindex 4 | 219.00 ms (171.56 ms)<br>307.96 ms (192.22 ms) |
| Number of words Before the pause (Position-w) | Wo 1 (<=4)<br>Wo 2 (5-7)<br>Wo 3 (>=8) | 152.69 ms (122.43 ms)<br>242.46 ms (176.05 ms)<br>345.98 ms (188.35 ms) |
| Number of syllables Before the pause (Position-s) | Sy1 (<=12)<br>Sy1 (>=13) | 168.54 ms (128.48 ms)<br>341.20 ms (190.30 ms) |
| Number of words After the pause | L7 (<= 7)<br>L8_15 (8-15)<br>L16_23 (16-23)<br>L24 (>=24) | 275.17 ms (176.54 ms)<br>315.55 ms (199.30 ms)<br>315.44 ms (213.17 ms)<br>277.23 ms (228.10 ms) |
| Motivation for the MARY system to set a pause | Pu1 [ , : ) - ]<br>Pu2 [ ( ‘ “ ]<br>Before KON<br>Before VORFELD | 326.64 ms (191.73 ms)<br>185.92.ms (147.22 ms)<br>311.29 ms (156.11 ms)<br>161.31 ms (157.37 ms) |
| All | | 295.57 ms (191.38 ms) |

**Table 2 :** Mean duration of pauses (in ms) and standard deviation.

The results also show that the duration of pauses for break index 4, which is uniformly set to 300ms by MARY does not significantly differ ($p = 0.26$) from the duration of pauses actually produced by the natural speaker in these positions (MEAN 307.96 ms). In contrast MARY's duration of 100ms for break index 3 differs significantly ($p < .01$) from the duration of pauses the natural speaker produced (MEAN 219.00 ms ).

The following figures illustrate the statistic distribution of duration of pauses according to the selected features. All duration measurements are in ms.[3]

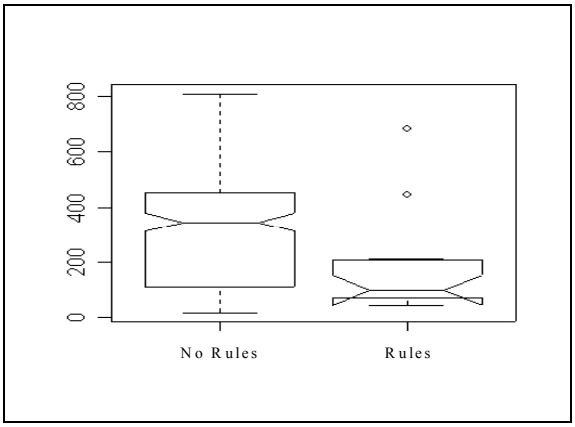


**Figure 3 :** Boxplot of the feature « g2p-method of the word after the pause ».
NoRules = lexicon, compound or mtu;
Rules = transcription obtained by rules

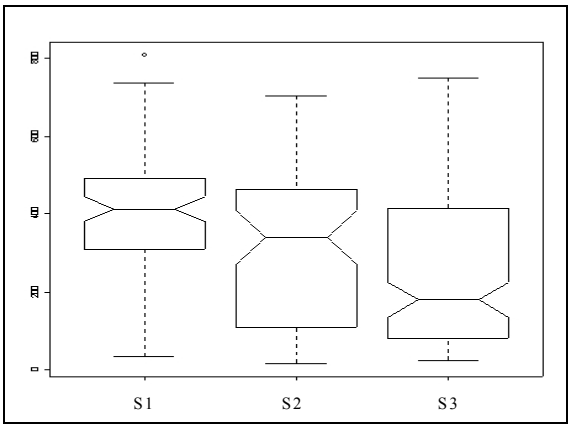


**Figure 4 :** Boxplot of the feature « Part-of-Speech ».

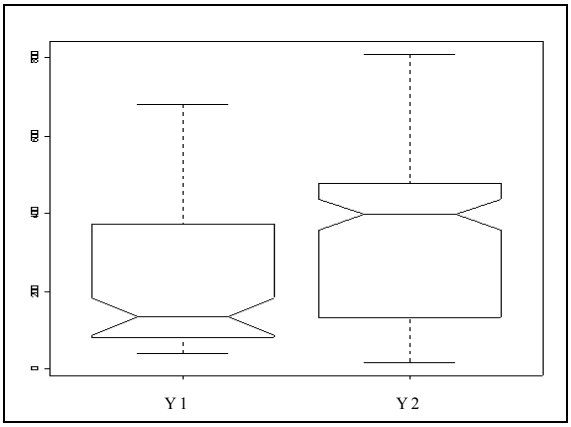


**Figure 5 :** Boxplot of the feature « Syntactic – Phrase »
Y1 = {AP,AVP,CNP,NP}
Y2 = {CAP,_,MPN,PP,VZ}

[3] Ad Figure 4 :
S1 = {APPR,PPER,VAFIN,VAINF,VAPP, VMPP,VMFIN,VVFIN,VVINF,VVPP}
S2 = {ADJD,FM,NE,PRF,PROAV,PTKVZ}
S3 = {ADJA,ADV,APPRART,APZR,ART, NN,PIS,PPOSAT,VMINF}
The clustering of the POS tags into 3 classes was motivated heuristically by looking at the statistic distribution of pause durations of each POS-tag.

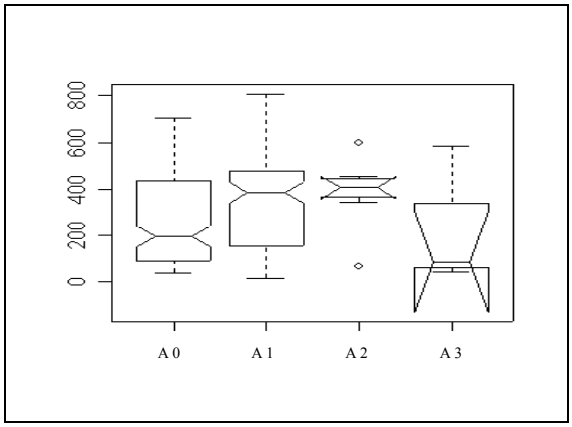


**Figure 6 :** Boxplot of the feature « syntactic attachment of the word preceding the pause ».

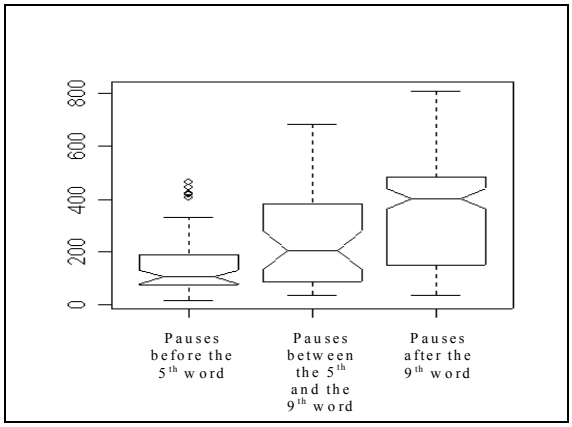


**Figure 7 :** Boxplot of the feature « number of words preceding the pause ».

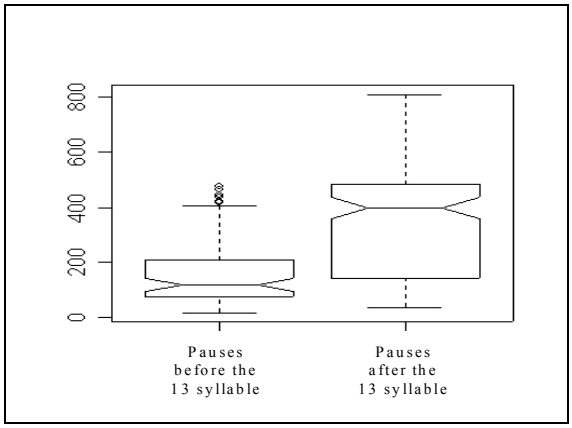


**Figure 8 :** Boxplot of the feature « number of syllables preceding the pause ».

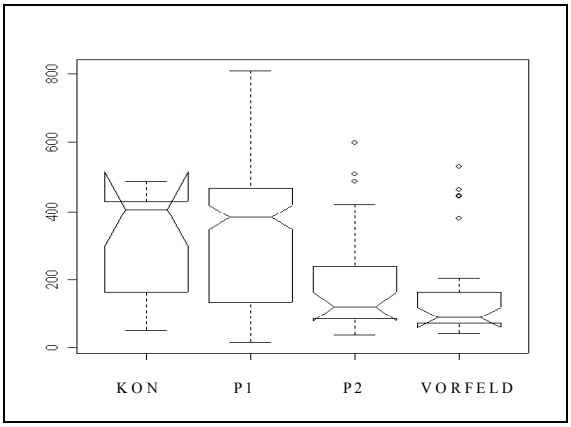


**Figure 9 :** Boxplot of the feature « motivation for the MARY system to set a break index »
KON = Before conjunctions; P1 = { , : ) - };
P2 = { ( ‘ “ }; VORFELD = ‘Vorfeld’

### 5.3 Modelling pauses

For the determination whether a pause occurs in a given position and the prediction of the duration of this pause we used the well known machine learning technique CART [classification and regression trees (Breiman, L., Friedman, J., Olshen, R., and Stone, C. J., 1984)]. As training data we use those sentences of the SpeeDurCont-corpus, where pauses occuring in the natural speech are also predicted by the MARY system. This data is divided randomly into 10 equally sized subsets. Each of these subsets was used as a test sample in a 10-fold cross validation routine. For the evaluation of the performance of this machine learning algorithm we used the correlation value (cor) between the predicted and the original values of the test sample. For training and evaluation we used the CART tree package provided by the statistical tool R[4]. In a preliminary step the candidates for a pause (each position between two words) are filtered against the criterion that a break index 3 or 4 is provided by the MARY algorithm. Then, the occurrence of a pause and its duration are learned.

For the first part of the learning procedure (i.e. determining whether a pause occurs or not) the two features (1) Motivation for the MARY system to set a break index and (2) number of

[4] R is available under the terms of the Free Software Foundation’s GNU General Public License, see: http://www.r-project.org

syllables from the beginning of the sentence until the pause produce the best results (cor = 0.46002). The feature "Motivation for the MARY system to set a break index" has the values 'pause before conjunction', 'pause after Vorfeld' and seven different punctuation signs [i.e. , - ) ( : ' " ]. The feature « numbers of syllables » was pooled into pauses before the 10th syllable and pauses after the 10th syllable. This number was also suggested by E. Zvonik et al.(2002), as a border for the production of short (before the 10th syllable) and longer pauses (after the 10th syllable).

In the second part of the learning procedure (i.e. determination of the duration of a pause) we used the following features (for subfeatures see Table 2):

- G2p-method of the word before the pause
- The syntactic attachment of the word preceding the pause
- The syntactic phrase preceding the pause
- The number of words preceding the pause
- The motivation for the MARY system to set a break index.

This assortment of features leads to the best results with cor = 0.42.[5] Evaluating the results shows still a rather low correlation. Two strategies, using a larger corpus for the machine learning technique and rethinking of the features used for training may be applied in future research in order to arrive at higher correlation values.

## 5. Conclusion

In the present study we evaluated the occurrence of pauses and the length of pauses made by a natural speaker against a current speech-synthesizer (MARY). In comparison to the MARY system, the natural speaker realises significantly fewer pauses at prosodic phrase boundaries, and rather relies on prosodic means like final lengthening and boundary tones.

Though these results indicate that the speech synthesis system uses pauses to a much greater extent than the natural speaker, it will be necessary to perform perception experiments in order to evaluate if and in which contexts this additional redundancy in MARY's current setting is to be avoided.

Furthermore, we have shown that the natural speaker makes significant differences in pause duration depending on the surrounding environment. We used this data for training a CART-based predictor for pause durations. Using this method, we were able to set very fine graded pauses according to the structural environment the pause would occur in.

[5] When comparing these results with correlation coefficents usually achieved when using similar techniques to the prediction of the duration of phonetic segments the result seems rather poor. But it has to be kept in mind, that there are much fewer trainung-samples in this case, and that the durarion of pauses displays a much higher variance than a speech-sound with its clear limitations on minimal and maximal inherent duration.

## References


Beckman, M. E., Ayers, G. M. *Guidelines for ToBI Labelling*. Online MS and accompanying files, 1994. Available at: http://www.ling.ohio-state.edu/phonetics/E_ToBI

Breiman, L., Friedman, J., Olshen, R., and Stone, C. J., *Classification and Regression Trees,* Chapman and Hall, New York, 1984.

Goldman-Eisler, *Psycholinguistics: Experiments in spontaneous speech*. NY: Academic Press. 1968.

Grice, Martine, Baumann, S., "Deutsche Intonation und GToBI." *Linguistische Berichte 191,* pp. 267-298, 2002.

Neubarth F., Alter K., Pirker H., Rieder E., Trost H.: "The Vienna Prosodic Speech Corpus: Purpose, Content and Encoding," in Zuehlke W. et al. (eds.), *Konvens 2000 – Sprachkommunikation*, VDE Verlag, Berlin, pp.191-96, 2000.

Schröder, Mark, Trouvain, Jürgen: "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching," *International Journal of Speech Technology, (6)* pp. 365-377, 2003.

Schiller, A., Teufel, S., Stöckert, C., Thielen, C. *Guidelines für das Tagging deutscher Textcorpora*. University of Stuttgart / University of Tübingen, 1999.

Zellner, B., "Pauses and the temporal structure of speech," in E.Kellner (Ed.) *Fundamentals of speech synthesis and speech recognition,* pp.41-62. 1994.

E. Zvonik and F. Cummins, "Pause Duration and Variability in Read Texts," in *Proceedings of ICSLP'02,* pp.1109-1112. 2002.