On the Role of Duration Prediction and Symbolic Representation for the Evaluation of Synthetic Speech

Caren Brinckmann & Jürgen Trouvain

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

Abstract

In order to determine priorities for the improvement of timing in synthetic speech this study looks at the role of segmental duration prediction and the role of phonological symbolic representation in listeners' preferences. In perception experiments using German speech synthesis, two standard duration models (Klatt rules and CART) were tested. The input to these models consisted of symbolic strings which were either derived from a database or a text-to-speech system. Results of the perception experiments show that different duration models can only be distinguished when the symbolic string is appropriate. Considering the relative importance of the symbolic representation, "post-lexical" segmental rules were investigated with the outcome that listeners differ in their preferences regarding the degree of segmental reduction. As a conclusion, before fine-tuning the duration prediction, it is important to calculate an appropriate phonological symbolic representation in order to improve timing in synthetic speech.

1. Introduction

It is generally accepted that timing plays a crucial role for encoding and decoding speech next to intonation modelling. It is therefore the second major part for prosody modelling. Prosody is indispensable for an adequate reflection of the linguistic information in speech, but also for paralinguistic and extra-linguistic factors. However, the ambitious aims of enlarging the scope of applications to emotions and personality cannot hide the difficulties of acceptance of synthetic speech, even for the default case of reading normal texts.

The prerequisite for appropriate timing in speech synthesis is a high quality model for duration prediction. The performance of a duration model is usually measured by comparing the predicted durations to durations observed in a database of segmented natural speech.

Input to the duration prediction in text-to-speech (TTS) systems is a symbolic representation which consists of phonological information regarding sound segments, syllable structure, lexical stress, phrasal accents and prosodic phrase boundaries.

This study contributes to the following questions:

- Are the differences between predicted and observed durations also perceptible in synthetic speech
  a) if symbolic representation is "natural"-like?
  b) if symbolic representation is not "natural"-like?
- What is the role of symbolic representation for timing in synthetic speech?
- What is the contribution of segmental "post-lexical" rules for the acceptance of synthetic speech?

Our aim is not to fine-tune the duration prediction for a given TTS system, but to determine priorities for improving timing in synthetic speech.

For this study we selected two standard duration prediction methods: the rules developed by Klatt [1], and a statistical machine learning algorithm, the classification and regression tree (CART) [2]. As information source about natural speech we use a German manually labelled speech database. Since not all information that was needed as input for the duration models was present in the database, some further processing was necessary. A German TTS system was used to create stimuli for various perception experiments.

2. Database preparation

2.1. Corpus

The speech database used was the Kiel Corpus of Read Speech [3], which is also known as PhonDat. Most parts consist of single sentences taken from a variety of contexts, e.g. railway information scenarios but also segmentally balanced material, as well as two shorter stories. Two speakers (male speaker *kko* and female speaker *rtd*) read the entire material, 51 other speakers read only part of it. The database is segmented manually, the sound segments are labelled as realised forms, and it is indicated when a realised form deviates from its lexical form. Pauses are labelled as well, and orthographic word boundaries and function words are marked separately. Prosodic information includes lexical stress, phrase accents and phrase boundaries.

2.2. Syllabification

Since syllabic information, which we needed for duration modelling, was missing in the database, a syllabification of the realised utterances had to be performed. The syllabification algorithm defined every vowel as syllable nucleus and every sonorant [m,n,N,l][1] (preceded by a consonant) as potential syllable nucleus. The syllabification of the segments between the established nuclei was based on standard phonological principles such as:

- *Ambisyllabicity*: a consonant in a VCV pattern is allowed to belong either to one or to both syllables, e.g. "raten" (engl. "guess") ['ra:-t@n] vs. "Ratten" (engl. "rats") ['ra_t@n]
- *Obligatory Coda Closing*: syllable coda must be closed after a short, lax vowel (except schwa)
- *Maximal Onset Principle*: put as many consonants into the syllable onset as allowed by phonotactic restrictions.

In order to evaluate the quality of our syllabifier, we tested the algorithm on the German part of the Celex lexical database

[1] Throughout the paper the SAM Phonetic Alphabet (SAMPA) for German is used [4]. For ease of reading all pronunciations are given in square brackets, irrespective of phonemic/lexical/canonical or phonetic/realised status. Syllable boundaries: "-" (ordinary), "_" (ambisyllabic).

[5]. With a score of 97% matching, and with doubts on some cases of syllabification in Celex in mind, we decided that the algorithm had reached acceptable quality.

All labelled sentences of the corpus were then syllabified, irrespective of word boundaries or other morphological, syntactic or prosodic information except the sentence and prosodic phrase boundaries.

2.3. Segmental mappings

In PhonDat, the release phase of a plosive is labelled separately from the closure phase: the closure gets the symbol of the plosive, whereas the release is labelled as [-h] for all plosives. Since the intrinsic duration of the release phase varies considerably for the six German plosives, we decided to mark the releases with different symbols, according to the preceding closure. In our corpus, the plosive closures are therefore labelled as [P_,B_,T_,D_,K_,G_], the plosive releases are labelled as [p,b,t,d,k,g] respectively.

As the sonorants [m,n,N,l] vary in their intrinsic duration depending on whether they are syllabic or not, we introduced the SAMPA convention of [m=,n=,N=,l=] for those segments which were classified as syllable nucleus by our syllabifier.

2.4. Prosodic labels

The lexical stress information given in PhonDat (primary or secondary stress) is attached to the vowel of the stressed syllable. We considered not only the vowel to be stressed, but all segments of the same syllable. Therefore, if our syllabifier worked incorrectly, the stress information was wrong for some segments. Note that function words carry no lexical stress in PhonDat.

In PhonDat, each word was labelled as being accented or not. The accent strength was given on a scale from 0 (unaccented) to 3 (emphatically accented). We kept this division and marked every segment in a word with the labelled accent strength.

Prosodic phrase boundaries are labelled with only one generalised category ("PGn"). Since we considered the distinction between a minor phrase boundary and a major phrase boundary important for duration prediction, we decided to differentiate the phrase boundaries as follows: a major boundary is followed by a pause, a minor boundary is not. Bear in mind that the pause-based definition and the boundary strength definition correspond only roughly, and that this can lead to some serious mistakes (see 5.3).

2.5. Test and training corpus

From the entire text material, 10 sentences of different length were selected randomly as test corpus for the subsequent perception experiments: 4 long (> 30 syllables), 2 medium (around 20 syllables), 4 short (< 10 syllables) sentences. The remaining data formed our training corpus.[2] We opted for the data of the male speaker *kko*. In total, the training corpus read by *kko* consisted of 23,133 realised segments, whereas the test corpus consisted of 661.

[2] The test corpus, the stimuli for the perception experiments and other material related to this study is available via http://www.coli.uni-sb.de/~cabr/ssw4/

3. Duration prediction models

3.1. Factors

First, the following influencing factors for segmental duration were defined in accordance with the factors used by the Klatt rules [1]. Then, for each realised sound segment these factors were extracted from the database. Domains are presented in small capitals, factors in italics, and the possible values in standard font.

- REALISED SEGMENT
  - *segment identity*: modified SAMPA code (see 2.3)
  - *segment type*: vowel, consonant
  - *manner of articulation*: 0 (vowels), plos. closure, plosive release, affricate, fricative, nasal, lateral
- POSITION OF SEGMENT IN SYLLABLE
  - *syllable initial*: yes, no
  - *syllable part*: onset, nucleus, coda, ambisyllabic
- SYLLABLE
  - *lexical stress*: primary, secondary, unstressed
- POSITION OF SYLLABLE IN WORD
  - *word initial*: yes, no
  - *word final*: yes, no
- WORD
  - *part-of-speech*: function word, content word
  - *degree of accentuation*: unaccented (0), partly de-accented (1), accented (2), emphatic (3)
  - *length in syllables*: integer
- POSITION OF WORD IN PHRASE
  - *minor phrase initial*: yes, no
  - *major phrase final*: yes, no
- REALISED PREVIOUS SEGMENT
  - *segment type*: vowel, consonant
  - *manner of articulation*: 0 (vowels), plos. closure, plosive release, affricate, fricative, nasal, lateral
- REALISED FOLLOWING SEGMENT
  - *segment type*: vowel, consonant
  - *manner of articulation*: 0 (vowels), plos. closure, plosive release, affricate, fricative, nasal, lateral
  - *voiced*: 0 (vowels), yes, no
  - *syllable part*: onset, nucleus, coda, ambisyllabic

3.2. The Klatt rules for German

A rather simple method for predicting segment durations are the rules developed by Klatt [1] for American English. Klatt rules predict the segmental duration by multiplying the intrinsic duration of a given segment with a context-dependent factor value. The result is then added to a segment-specific minimal duration which also can be multiplied by a context-dependent factor.

To adapt the Klatt rules to German, the inventory of sound segments had to be transferred and sounds occurring in German but not in English had to be integrated, e.g. [y:,2:,9,6,x]. Also, the syllabic sonorants [m=,n=,N=,l=] were distinguished from their non-syllabic counterparts.

Klatt takes the duration of a segment in an *accented* position as its intrinsic duration. In contrast, for our adaptation the *mean* duration of *all* realisations by one speaker is taken as intrinsic duration. In our approach, minimal duration was derived from the magnitude of the intrinsic duration as follows: intrinsic durations were divided into six classes. To each class, one minimal duration was

assigned, ranging from 10 ms to 60 ms. The schwa sounds [@,6] and the syllabic sonorants, which always occur in unstressed position, get additional 20 ms for the minimal duration.

The adaptation of the context-dependent factor values to German was done in two ways. First, a manual-auditive trial-and-error procedure with the help of a German speech synthesiser [6] took place. Second, the predicted durations were compared with the corresponding durations in the training corpus, broken down for each sound and each context-dependent rule. Depending on the differences regarding mean duration, standard deviation, and correlation coefficient, and on the frequency of occurrence in the database, the factor values were adapted iteratively.

3.3. The CART

A CART is a binary branching tree with questions about the influencing factors at the nodes and predicted values at the leaves. The advantages of CARTs are that standard tools for their generation are widely available, and that the computed regression tree is interpretable (in contrast to neural networks). The disadvantage lies in the fact that it needs a large amount of training data.

The first necessary step is to factor out the influence of the intrinsic duration. To do this, the absolute duration values were first converted to z-scores, and the mean and the standard deviation of each sound were stored in a separate file.[3]

The CART was then trained on the training corpus of speaker *kko* with the program "wagon" from the Edinburgh Speech Tools Library [7]. To keep it simple and comparable, we used the same 19 factors as for the Klatt rules (see section 3.1), and did not change the default settings of "wagon" (e.g. minimum of 50 cases in one leaf).

3.4. Performance statistics

When developing a new model for duration prediction in a TTS system, its performance is usually measured by comparing the predicted durations with the observed "original" durations in a database. The performance of the new model is then expressed in terms of error rates such as root mean square error (RMSE) and the correlation coefficient. Thus it can be compared to another duration model.

As can be seen in table 1 the performance of *cart* was always superior to the performance of *klatt* in terms of differences of predicted and observed durations. This is true for both speakers, no matter whether the training or the test part of the corpus was used.

Table 1: Correlation coefficient and RMSE broken down for a) duration model, b) part of corpus, and c) speaker.

	correlation coeff.		RMSE in ms	
	cart	klatt	cart	klatt
kko_train	.89	.82	20.35	25.56
kko_test	.86	.79	22.46	27.41
rtd_train	.84	.79	20.83	23.78
rtd_test	.83	.78	21.40	23.40

[3] A z-score can be converted back easily into the absolute duration value by applying the following formula: absolute duration = (z-score * standard deviation) + mean duration

4. Experiment 1: Perceptual relevance of duration models

4.1. Aims

Even if RMSE and correlation coefficient show a significant difference both between the two duration models and between each model and the original durations: Can the differences be perceived in synthetic speech? Furthermore, if the models are perceptually different: Do listeners really prefer the one that is closest to the original data? In summary, are the performance measurements RMSE and correlation coefficient a good estimate of the listeners' preferences?

To answer these questions, the developed CART and the Klatt model are compared to the re-synthesised "original" realisation of the test corpus by speaker *kko*.

4.2. Methods

Stimulus generation was performed as follows. For each of the 10 sentences of the test corpus (see section 2.5) three different versions were created according to the duration model: a) *cart*, b) *klatt*, and c) the *original* durations as segmented in the database. For each of the resulting 30 stimuli the realised segments and the pause durations are taken directly from the database; the F0 target values were determined by one of the authors by inspecting and measuring the original F0 curve. Thus, everything but the segmental duration is kept the same for each sentence version. The stimuli were then generated with a male voice of the MBROLA diphone synthesis [8] within the framework of the MARY TTS system [6] using segment symbol, segment duration and F0 targets as input.

For every test sentence each version (*original*, *cart*, and *klatt*) was paired with every other version, leading to a set of 6 stimulus pairs for every sentence (both orders for each pair). Every sentence formed a block, and within each block the stimulus pairs were randomly ordered. 9 undergraduate students of phonetics and/or computational linguistics, who are native speakers of German, served as subjects. The 9 subjects listened to every stimulus pair only once via loudspeakers, and had to decide within 5 seconds, which stimulus they preferred (forced choice). The sentences were given in written form. The whole test took about 20 minutes.

4.3. Results

As shown in table 2, on the whole the *original* duration was significantly preferred to both *cart* and *klatt*. Compared to the *original* durations, *cart* received a higher or equal score for four sentences (three of those are long sentences). *klatt* never scored higher than the *original* durations, but was judged equal for two sentences (one long, one short).

Table 2: Scores of preference experiment 1. n=180.

duration model	scores	significance
original	**126**	significant
cart	54	($p \leq 0.001$)
original	**129**	significant
klatt	51	($p \leq 0.001$)
cart	**108**	significant
klatt	72	($p \leq 0.01$)

If compared directly, *cart* is significantly preferable to *klatt*. *klatt* received the worst two scores for two long sentences. The four cases where *klatt* scored slightly higher or equal to *cart* are three short sentences and one medium-length sentence.

4.4. Interpretation

The results from experiment 1 show that the differences in RMSE and correlation coefficient are indeed perceived by people listening to synthesised speech. Furthermore, listeners prefer the duration model that is closer to the original data, in our case the *cart*.

In summary, the performance measures for duration prediction, RMSE and correlation coefficient, adequately reflect the preference of TTS users - as long as everything but the segmental duration is kept as close to the original as possible.

5. Experiment 2: The role of the symbolic string

5.1. Aims

Experiment 1 showed that two different duration models can be distinguished by the listeners - on condition that the input to the duration prediction is optimal. However, even the best TTS systems produce errors, and they can occur at several different stages before segmental duration is calculated. Some examples, taken from our test corpus, illustrate possible errors:

- wrong word pronunciation: "abends" (engl. "at night") ['a:-be:nts] instead of ['a:-b@nts] or ['a:-bn=ts]
- unnatural phrasing, esp. for longer utterances
- wrong/strange lexical stress assignment: "Telefonhörer" (engl. "telephone receiver") ['te:-le:-fo:n-"h2:-r6] instead of [tE-l@-'fo:n-"h2:-r6]
- inappropriate accent placement: "Doris fährt zu weit links." instead of "Doris fährt zu weit links." (engl. "Doris drives too far on the left.")
- ignoring post-lexical phonological processes:
  - Schwa deletion: "richten" ['rIC-t@n] instead of ['rIC-tn=]
  - assimilation: "gegen" ['ge:-g@n] instead of ['ge:-gN=]
  - glottal stop deletion: "(indem) Sie auf (die)" [zi:-?aUf] instead of [zi:-aUf]
  - consonant deletion in clusters "mit dem" [mIt-de:m] instead of [mI-de:m]

The second experiment deals with this problem of a potentially sub-optimal input to the duration prediction. The two questions to be answered are:

- Is the difference between two duration models still perceptible if the symbolic string is slightly deficient?
- Which is more important: an optimal symbolic string or an optimal duration prediction?

5.2. Methods

The stimulus generation was similar to the previous experiment. Four different versions for each of the 10 test sentences were created. Two versions used the symbolic string calculated by the German TTS system MARY [6] as input for the two duration models. We refer to these two versions as *tts.cart* and *tts.klatt*, respectively. The pause duration and the F0 targets of *tts.cart* and *tts.klatt* were calculated directly by the TTS system.

The other two versions used the symbolic string from the database as input for the two durational models. Those two versions are named *orig.cart* and *orig.klatt*. F0 targets for *orig.cart* and *orig.klatt* were calculated by the TTS system from the original symbolic string containing the original phrase accents as realised by speaker *kko*. The pause durations were taken directly from the database.

The synthesis and the procedure for the listening test were the same as in the previous experiment. Three of the nine subjects had also participated in experiment 1.

5.3. Results

Results given in table 3 reveal that *tts.klatt* and *tts.cart* are not significantly distinct in the listeners' preferences. In contrast, significant differences were found when the symbolic strings differ in their origin: stimuli with the *original* symbolic string are always preferred to those with the symbolic string generated by *tts* irrespective of which duration model was used.

Table 3: Scores of preference experiment 2. n=180.

symbol string	duration model	scores	significance
tts	*cart*	83	not significant
tts	*klatt*	97	
tts	*cart*	48	significant
orig	*cart*	**132**	($p \leq 0.001$)
tts	*cart*	50	significant
orig	*klatt*	**130**	($p \leq 0.001$)

For most sentences the stimuli based on the *original* were the clear "winners" over the TTS versions. Interestingly, for one sentence the stimuli with the *original* symbolic string resulted in a clearly worse score. In this case, a prosodic phrase boundary reflecting a syntactic clause boundary was marked as a major boundary in the *tts* version, whereas in the *original* version a phrase boundary without a following pause was labelled as a minor boundary.

5.4. Interpretation

The results show that the difference between *cart* and *klatt* is no longer perceptible if the symbolic string is not optimal. The strong preference of the *original* symbolic strings to the ones generated by the TTS system suggests that the difference between the two duration models is masked by the deficient TTS string. Therefore, it can be concluded that the correct prediction of the symbolic string by the system is a crucial component for timing.

6. Experiment 3: "Post-lexical" rules

6.1. Aims

One of the things that can go wrong in predicting the "correct" symbolic string is the modification of the lexical segmental string. Usually, the lexical form of a word is looked

up in a lexicon or derived through grapheme-to-phoneme rules. In most systems, this lexical segmental string is then subjected to "post-lexical" segmental modification rules. E.g. the lexical form ['ha:-b@n] of the German word "haben" (engl. "have") is often reduced to ['ha:-bm=] in natural speech.

But before developing any sophisticated methods to model these post-lexical processes for our TTS system, we posed the following question: Is there a perceptual difference between the following forms of the segmental string:

- the lexical form,
- the natural form as produced by our selected speaker,
- the form predicted by a simple set of known post-lexical rules for German, e.g. [9].

Furthermore, if there is a perceptual difference, we want to determine: Which form is preferred? If the original form is preferred, then it is worthwhile modelling the natural post-lexical processes as closely as possible. However, as suggested by Portele [10], not all listeners necessarily prefer the original, i.e. the more reduced form.

6.2. Methods

For every sentence in the test corpus, three versions were prepared which differed only on the segmental level: the *original* form, the *lexical* form, and *post-lexical* form. All three versions were given the same F0-target values and pause durations as the *orig*-versions of experiment 2, and used the CART to predict the segment durations. The *original* form is therefore the same as *orig.cart* in experiment 2.

For the *lexical* form the entries given in PhonDat's lexicon were used. Please note that these do *not* refer to the realised forms of the spoken utterances we used so far - the realised forms are used in *original*. In the PhonDat lexicon, each syllable starts with an onset consonant, i.e. a glottal stop is coded before a vowel if there is no other consonant in the onset. Furthermore, the morpho-phonological change of [r] to its schwa form [6] is already considered. In contrast to the labelling conventions in PhonDat, we regarded a plosive as closure plus release, i.e. a plosive release is not an insertion.

In order to get the *post-lexical* form, the lexical form was subjected to the following set of four post-lexical rules (in that order):

1. Delete every glottal stop, except at the very beginning of a major phrase, and in a lexically stressed syllable (function words carry no lexical stress in PhonDat).
2. Delete every schwa, that is followed by a sonorant [n,m,l] in the same syllable. Then, the sonorant becomes the nucleus of the syllable and is changed into a syllabic consonant [n=,m=,l=].
3. Assimilate every syllabic [n=] to the place of articulation of the preceding plosive or sonorant: [n=] preceded by [p,b,P_,B_,m] becomes [m=], and [n=] preceded by [k,g,K_,G_,N] becomes [N=].
4. Delete every plosive release [p,b,t,d,k,g] that is followed by a consonant.

Evaluating the power of these four post-lexical rules a descriptive statistics was applied on the training corpus containing 26,322 lexical segments in total. Table 4 gives the frequency of deletions, replacements and the cases without any modification of the lexical segment. In total, 92.3% of the modifications are correctly modelled by only four post-lexical rules. 84.5% of the segments in the training corpus keep their lexical form. So, if we consider the lexical form as the base-line model of post-lexical rules, it correctly predicts 84.5% of our training corpus.

Table 4: Percentage of replacements, deletions and unchanged segments, broken down to occurrences in the database and correctly predicted by post-lexical rules.

	original	correctly predicted
no changes	84.5%	96.1%
deletions	12.1%	74.9%
replacements	3.4%	60.0%
total	100%	92.3%

The synthesis and the procedure for the listening test were the same as in the previous experiments. Four of the nine subjects also participated in experiment 2, and two of them participated in experiment 1 as well.

6.3. Results

Table 5 shows that post-lexical *rules* are preferred to the *lexical* form, but the difference is only marginally significant ($p = 0.053$). Neither the comparison of *original* and *lexical* form nor the comparison of *original* form and post-lexical *rules* show a significant difference.

Table 5: Scores of preference experiment 3. n=180.

segmental string	scores	significance
original	94	not significant
lexicon	86	
original	91	not significant
rules	89	
lexicon	77	marginally significant
rules	**103**	($p = 0.053$)

However, taking a closer look at the data, we saw that we can group the subjects according to their preference of *original* vs. *lexical* form: 5 subjects prefer the *original* to the *lexical* form (group 1), whereas 4 subjects do not (group 2). If we treat those groups separately, we obtain the results presented in table 6.

Table 6: Scores of preference experiment 3 pooled over two listener groups: group 1 (n = 100) and group 2 (n = 80) as described above. Significant preferences are marked with * ($p \leq 0.05$).

segmental string	group 1	group 2
original	**61** *	33
lexicon	39	47
original	57	34
rules	43	46
lexicon	38	39
rules	**62** *	41

Group 1 significantly prefers both the *original* form and the post-lexical *rules* to the *lexical* form, whereas group 2 does not prefer *original* to *lexical* form and does not distinguish

between post-lexical *rules* and *lexical* form. For *original* form vs. post-lexical *rules*, both groups show no significant preference.

6.4. Interpretation

Hearers of synthetic speech tested here fall into two groups: The first group clearly rejects the lexical form, which might sound too unnatural to them, but makes no difference between the original form and post-lexical rules. The second group makes no difference between the lexical form and post-lexical rules, but rather dislikes the original form, which might be too reduced for them. Thus, it appears that using post-lexical rules will satisfy both groups.

7. Summary & Conclusions

The purpose of this study was to find out whether the differences in the durations predicted by duration models to those found in natural speech data are perceptually relevant in synthetic speech. Subsequently the role of the symbolic representation which serves as the input to a duration model was investigated.

The results in experiment 1 show that the differences between segment durations observed in natural speech as labelled in the PhonDat database and segment durations predicted by duration models also reflect perceptually relevant differences. This is equally true for both models, the CART-based model and the adapted Klatt rules, if we compare those with copy-synthesised original sentences. This difference is also confirmed by the perceptual comparison between both duration models, CART being preferred over the Klatt rules.

A very important restriction to the generalisability of this finding in experiment 1 is shown by the outcome of experiment 2. The preference score differences between the CART and the Klatt duration model were neutralised as soon as the symbolic input to the duration model was calculated by a TTS system rather than transferred from a segmentally and prosodically labelled database. It became clear that a good timing prediction starts at the phonological level, before the actual duration prediction. Many timing problems occur at a phonological-symbolic level with relevant consequences for the phonetic-durational level.

Apart from the pronunciation lexicon and the prosodic structure, post-lexical rules also have an effect on the calculated symbolic string. To find out more about the role of post-lexical phonological processes for speech timing, rather simple post-lexical rules were applied in experiment 3. These rules consider schwa deletion, plosive reduction in consonant clusters, nasal assimilation and glottal stop deletion, and cover more than 90% of all processes in the training corpus. On the one hand, the results of this last experiment show that the synthetic utterances based on these rules score better than the utterances with the full lexical forms, although this difference is only marginally significant; on the other hand, it can be seen that not every listener preferred the versions with the original segment reductions. Apparently, there are different listening preferences just as there are diverse speaker strategies in speech production.

It can be presumed that factors other than segmental ones are responsible for the superiority of symbolic strings derived from natural speech compared to calculated symbolic strings. The prediction of location and type of pitch accents and phrase boundary tones as well as the prediction of location and strength of prosodic phrase boundaries might play a central role in calculating a symbolic string that leads to an acceptable synthesis on the timing level. This is also true for correct word pronunciations in the lexicon and for the correct assignment of the lexical stress.

We consider it to be vital integrating the above mentioned phonological points for speech database annotation, be it for exploiting the audio data as a synthetic voice, or as a basis for quantifying data, as in this study. Apart from mapping the segment inventory from the TTS to the database and vice versa, syllabification and a quite detailed definition of boundary strength is required.

We conclude, that for an improvement of timing in synthetic speech, paying more attention to the various linguistic interrelationships leading to an appropriate phonological symbolic representation is essential both on the segmental and the prosodic level.

Acknowledgements

We would like to thank Ralf Benzmüller (G-DATA Software) for his encouragement to carry out this research, and Marc Schröder and Bill Barry for their comments on an earlier draft of this paper.

8. References

[1] Klatt, D.H. Synthesis by rule of segmental durations in English sentences. In Lindblom, B. & Öhmann, S. (eds), *Frontiers of Speech Communication Research*. 287-299, 1979.

[2] Riley, M. Tree-based modelling of segmental duration. In G. Bailly, C. Benoit, and T.R. Sawallis (eds), *Talking Machines: Theories, Models, Designs*, 265-273, 1992.

[3] *The Kiel Corpus of Read Speech, Vol. 1* (CD-ROM). Institut für Phonetik und Digitale Sprachverarbeitung, University of Kiel. 1994.

[4] *SAMPA Phonetic Alphabet for German*. http://www.phon.ucl.ac.uk/home/sampa/german.htm

[5] Baayen, R.H., Piepenbrock, R., & Gulikers, L. *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995

[6] Schröder, M. & Trouvain, J. The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *This volume*.

[7] Taylor, P., Caley, R., Black, A. & King, S. *Edinburgh Speech Tools Library. System Documentation.* CSTR, University of Edinburgh. http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/, 1999.

[8] *MBROLA synthesis*. http://tcts.fpms.ac.be/synthesis/mbrola.html

[9] Kohler, K.J. *Einführung in die Phonetik des Deutschen*. Erich Schmidt Verlag: Berlin. 2nd ed. 1995.

[10] Portele, Th. Reduktionen in der einheitsbasierten Sprachsynthese. *Fortschritte der Akustik. DAGA 97 Kiel*. 1997.