INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6, 21–31, 2003

The Role of Duration Models and Symbolic Representation for Timing in Synthetic Speech

CAREN BRINCKMANN AND JÜRGEN TROUVAIN
Institute of Phonetics, University of the Saarland, Saarbrücken, Germany

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 the perceptual quality of a text-to-speech system. In perception experiments using German speech synthesis, two standard duration models (Klatt rules and CART) were tested. The input to these models consisted of a symbolic representation which was 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 representation 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 derive an appropriate phonological symbolic representation in order to improve timing in synthetic speech.

Keywords: speech synthesis, timing, evaluation, duration modelling

1. Introduction

It is generally accepted that, next to intonation, timing plays a crucial role for encoding and decoding speech. Prosody is indispensable for an adequate reflection of the linguistic information in speech, but also for paralinguistic (e.g. emotions and attitude) and extralinguistic factors such as personality. 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. Alternatively, perception experiments can be performed where listeners judge the quality of different timing manipulations.

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 optimal?
  b. if symbolic representation is not optimal (i.e. derived by a TTS system)?
- What is the role of symbolic representation for timing in synthetic speech?
- What is the contribution of segmental post-lexical rules as one frequent factor influencing timing for the acceptance of synthetic speech?

There is certainly no doubt that improvements in the duration model improves the timing performance of a TTS system. However, our goal is to have a look beyond this "module", something which has received little attention so far: we examine the effect of the *input* to the duration prediction rather than the quality of the processing *within* this module for improvements

with respect to timing. Thus, 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 (1979), and a statistical machine learning algorithm, the classification and regression tree (CART) (Riley, 1992). 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 three perception experiments. The first experiment concentrates on the question whether differences between the segmental durations of natural speech and those segmental durations predicted by two different duration models are perceptible in synthetic speech. Since the symbolic input to the duration model can be seen as optimal in the first experiment, the follow-up experiment asks whether these differences are still perceptible in synthetic speech when the symbolic input to the duration model is *not* optimal, as it is often the case in TTS systems. Since it has proved that a sub-optimal symbolic input can mask the differences between different duration models, the third experiment aims at improving the prediction of the symbolic representation. Here, we focus on segmental modifications (post-lexical rules) as one of several factors influencing segmental durations.

2. Database Preparation

2.1. Corpus

The speech database used was the Kiel Corpus of Read Speech (IPDS, 1994), 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" (English "guess") ['ra:-t@n] vs. "Ratten" (English "rats") ['ra_t@n]
- *Obligatory Coda*: syllable 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 (Baayen, et al., 1995). 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. For reasons of possible re-syllabification[2] in connected speech this syllabification was carried out 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, for each word the accent strength is 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,[3] a minor boundary is not.

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 (see Appendix). The remaining data formed our training corpus.[4] In total, the training corpus read by *kko* consisted of 23,133 realised segments, whereas the test corpus consisted of 661.

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 (Klatt, 1979). Then, for each realised sound segment these factors were extracted from the database. For each domain the factors are presented 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), plosive 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 deaccented (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), plosive closure, plosive release, affricate, fricative, nasal, lateral

- Realised Following Segment

  segment type: vowel, consonant
  manner of articulation: 0 (vowels), plosive 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 (1979) 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 can also 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 achieved by a trial-and-error procedure. First, a manual-auditive procedure with the help of the German speech synthesiser "Mary" (Schröder and Trouvain, 2001) 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. There are other duration prediction methods which successfully handle the data sparsity problem such as the sums-of-products models (e.g. van Santen, 1994) or multivariate adaptive regression splines (MARS) (Riedi, 1997). However, we opted for CART as easily available data-driven method, because our main point is to explore the influence of two different conditions, namely optimal or sub-optimal symbolic input, on the perceptual distinctness of two different duration models.

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.[5] Two CARTs were then trained on the training corpus of speaker *kko* and *rtd*, respectively, with the program "wagon" from the Edinburgh Speech Tools Library (Taylor et al., 1999). To keep it simple and comparable, we used the same 19 factors as for the Klatt rules (see Section 3.1).

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. Interestingly, the correlation coefficient as well as the RMSE for *klatt* is lower on the test set of speaker *rtd* than on the training set of the same speaker. It could be argued that *klatt* generalises better from the training data to unseen data of one speaker, but it is still inferior to *cart*. In the following three perception experiments only the data of speaker *kko* was used.

One might think that an objective measurement as simple as the just presented performance statistics is sufficient as an evaluation measurement and is as good

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	0.89	0.82	20.35	25.56
kko_test	0.86	0.79	22.46	27.41
rtd_train	0.84	0.79	20.83	23.78
rtd_test	0.83	0.78	21.40	23.40

as perceptual judgements. We assume that this shall be true in many cases. However, to our knowledge there is no general evidence that comparisons of the duration output to duration data of single productions of a single speaker bring the same or nearly the same results as preference tests with actual listeners. We do know that there is variability in timing within speakers, between speakers, and between different text types (to mention only three factors). There are also reasons to assume that listening to speech in unfavourable conditions, such as synthetic speech, is preferred at a slower tempo than the conversational rate in natural speech, which affects timing properties of synthetic speech. Therefore, in our view perception data of several listeners are more meaningful than production data of just one speaker.

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 Appendix) 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 symbolic string of the realised segments as well as the pause durations are taken directly from the database;[6] 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 (Dutoit et al., 1996) within the framework of the MARY TTS system (Schröder and Trouvain, 2001) using segment symbol, segment duration and F0 targets as input.[7]

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: butt1, cn019, s053; one is short: s1011). *klatt* never scored higher than the *original* durations, but was judged equal for two sentences (one long: cn019, one short: s1011).

If compared directly, *cart* is significantly preferable to *klatt*. *klatt* received the worst two scores for two long sentences (cn015, s053). The four cases where *klatt* scored slightly higher or equal to *cart* are three short sentences (mr026, tk024, s1011) and one medium-length sentence (ko014). Looking at individual sentences more closely revealed that the "objective" performance statistics (RMSE and correlation coefficient) predict some of the "subjective" perception

Table 2. Number of preferred stimuli (scores) for three comparisons of Experiment 1 (9 subjects × 10 sentences × 2 directions = 180 judgements per comparison).

	Duration model	Scores	Significance
First comparison	*original*	**126**	Significant
	cart	54	($p \leq 0.001$)
Second comparison	*original*	**129**	Significant
	klatt	51	($p \leq 0.001$)
Third comparison	*cart*	**108**	Significant
	klatt	72	($p \leq 0.01$)

judgements correctly. However, in some cases the predictions were incorrect: high RMSE values and low correlation coefficients do not always lead to significantly worse preference scores. For example in one sentence (mr026) the values of RMSE and correlation coefficient would clearly predict a preference for *cart* but the perception experiment clearly shows a preference for *klatt*.

4.4. Interpretation

The results from Experiment 1 show that the overall 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 of the whole test corpus for duration prediction, RMSE and correlation coefficient, adequately reflect the general preference of TTS users—as long as the input to the duration model is optimal. Nevertheless single sentences can deviate from this pattern which shows that the "objective" performance statistics cannot fully replace "subjective" perception experiments with TTS systems.

5. Experiment 2: The Role of the Symbolic Representation

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" (English "at night") ['a:-be:nts] instead of ['a:-b@nts] or ['a:-b=nts]
- unnatural phrasing, esp. for longer utterances
- wrong/strange lexical stress assignment: "Telefonhörer" (English "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." (English "Doris drives too far on the left.")
- ignoring post-lexical phonological processes:
  - Schwa deletion: "richten" ['rIC-t@n] instead of ['rIC-t=n]
  - assimilation: "gegen" ['ge:-g@n] instead of ['ge:-g=N]
  - 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 three questions to be answered are:

1. Is the difference between the two duration models still perceptible if the symbolic representation is "noisy"?
2. What is the effect of a "noisy" input to the superior duration model compared to an optimal input?
3. What is more important, the quality of the input or the quality of the model?

If the quality of the *input* is more important, then an optimal input plus inferior duration model should not perform worse than noisy input plus superior duration model. In contrast, if the quality of the *duration model* is more important, then a noisy input plus superior duration model should not perform worse than an optimal input plus inferior duration model.

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 representation calculated by the German TTS system MARY (Schröder and Trouvain, 2001) as input for the two duration models. We refer to these two versions as *tts.cart* and *tts.klatt*, respectively. The pause durations and the F0 targets of *tts.cart* and *tts.klatt* were calculated directly by the TTS system.

The other two versions used the symbolic representation 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 representation containing the original phrase accents as realised by speaker *kko*. The pause durations were taken directly from the database.

Table 3. Number of preferred stimuli (scores) for three comparisons of Experiment 2. (9 subjects × 10 sentences × 2 directions = 180 judgements per comparison).

	Symbolic representation	Duration model	Scores	Significance
First comparison	*tts*	*cart*	83	Not significant
	tts	*klatt*	97	
Second comparison	*tts*	*cart*	48	Significant
	orig	*cart*	**132**	($p \leq 0.001$)
Third comparison	*tts*	*cart*	50	Significant
	orig	*klatt*	**130**	($p \leq 0.001$)

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 representations differ in their origin: stimuli with the *original* symbolic representation are always preferred to those with the symbolic representation generated by *tts* irrespective of which duration model was used.

For most sentences the stimuli based on the *original* were the clear "winners" over both TTS versions, no matter whether *cart* or *klatt* was used as duration model. Interestingly, for one sentence (ko023) the stimuli with the *original* symbolic representation 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 representation is not optimal. The strong preference of the *original* symbolic representations to the ones generated by the TTS system suggests that the difference between the two duration models is masked by the deficient TTS representation. The clear advantage of *cart* opposed to *klatt* as duration model is only visible when the input to the duration model is optimal. Therefore, it can be concluded that the correct prediction of the symbolic representation by the system is a crucial component for timing.

6. Experiment 3: "Post-Lexical" Rules

6.1. Aims

Things that can go wrong in predicting the "correct" symbolic representation are mainly the following (cf. 5.1): grapheme-to-phoneme conversion, lexical stress assignment, phrasing, accent placement, and post-lexical segmental processes. Although wrong phonemes, wrong lexical stress, and inappropriate placement of pitch accents can have an enormous negative impact on the acceptability and also the timing of synthetic speech, we consider these types of errors as relatively infrequent in our stimulus corpus, and therefore not central for our small-sized experimental setup. Very important for modelling timing is phrasing, i.e. predicting the location and possibly the strength of prosodic phrase boundaries. This implies the prediction of pause duration between and within sentences, the modelling of phrase-final lengthening and the assignment of boundary tones. However, since there are very few phrase boundaries in our stimulus corpus, phrasing did not seem the central timing component in our sentence-based investigation. Sound segments in connected speech undergo segmental modifications which can be modelled as post-lexical rules. In German these processes are quite frequent. As can be seen in Table 4 (column 2), the PhonDat database reveals that 15.5% of all lexical segments are changed (either deleted or replaced by another segment). Since these modifications appear relatively often in natural speech we decided to have a closer look at the post-lexical rules for German.

Usually, the lexical form of a word is looked up in a lexicon or derived through grapheme-to-phoneme

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

rules. In most systems, this lexical segmental string is then subjected to post-lexical segmental modification rules. For example the lexical form ['ha:-b@n] of the German word "haben" (English "have") is often reduced to ['ha:-b=m] 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. Kohler, 1995)?

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, Portele (1997) suggested that 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 the *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 with two exceptions: keep glottal stops (a) at the very beginning of a major phrase, (b) 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 post-lexical processes 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 baseline model of post-lexical rules, it correctly predicts 84.5% of our training corpus.

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

Table 5. Number of preferred stimuli (scores) of Experiment 3 (9 subjects × 10 sentences × 2 directions = 180 judgements per comparison).

Segmental string	Scores	Significance
Original	94	Not significant
Lexicon	86	
Original	91	Not significant
Rules	89	
Lexicon	77	Marginally significant
Rules	**103**	($p = 0.053$)

Table 6. Number of preferred stimuli (scores) of Experiment 3 (cf. Table 5) pooled over two listener groups: group 1 ($n = 100$) and group 2 ($n = 80$) as described above.

Segmental string	Group 1	Group 2
Original	**61***	33
Lexicon	39	47
Original	57	34
Rules	43	46
Lexicon	38	39
Rules	**62***	41

Significant preferences are marked with * ($p \leq 0.05$).

original form and post-lexical *rules* show a significant difference.

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.

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 this limited set of post-lexical rules could satisfy both groups.

7. Summary and 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 of 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. It might seem trivial to state that a good timing prediction requires a good computation of the input to the duration model. In our view, however, the dependence of timing prediction on the symbolic representation is often underestimated in "state-of-the-art" TTS systems. If we pose the question "what is more important for a good timing in a TTS system, an optimal input or an optimal processing in a duration prediction model?" then the answer given in the results of the experiments described here goes clearly to the quality of symbolic representation as input.

Apart from the pronunciation lexicon and the prosodic structure, post-lexical rules also have an effect on the calculated symbolic representation. 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

representations derived from natural speech compared to calculated symbolic representations. 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 representation 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 these 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.

Appendix

The 10 sentences of the test corpus are listed below. Sentence names are the ones from the PhonDat corpus.

butt1: Vor einem Laden stand bereits um sieben Uhr eine beachtliche Menschenmenge, denn man hatte dort am Abend vorher auf einem Schild schon lesen können, daß frische Butter eingetroffen sei.
cn019: Beachten Sie bitte folgende Anweisungen: Nehmen Sie den Telefonhörer ab, werfen Sie das Geld in den Münzspeicher, und wählen Sie die Nummer des Teilnehmers.
cn015: Der gesuchte Weg erscheint auf dem Stadtplan in roten Leuchtpunkten, indem Sie auf die Taste mit dem entsprechenden Namen drücken.
s053: Ich will einen Tagesausflug nach Nürnberg machen und entweder nachmittags gegen vier Uhr oder mit der letzten Verbindung abends zurückkommen.
ko023: Weil zwei Parlamentssitze frei werden, ist eine Nachwahl erforderlich.
ko014: Wir wollen früh fahren und die jüngeren Kinder aus den Ferien zurückholen.
s1011: Du hilfst mir beim Fernstudium.
tk024: Er wird ihr ewig treu bleiben.
mr090: Danach kannst Du Dich wirklich richten.
mr026: Doris fährt zu weit links.

Acknowledgments

We would like to thank Ralf Benzmüller (G-DATA Software) for his encouragement to carry out this research, and Marc Schröder, Bill Barry, Martine Grice, and three anonymous reviewers for their comments.

Notes

1. Throughout the paper the SAM Phonetic Alphabet (SAMPA) for German is used (Wells, 1996). For ease of reading all pronunciations are given in square brackets, irrespective of phonemic/lexical/canonical or phonetic/realised status. Syllable boundaries: "-" (ordinary), "‗" (ambisyllabic).
2. One PhonDat example is the German sentence "Am Himmel ziehen die Wolken" (English literally 'In the sky move the clouds.') where the lexical phonemic structure of the trisyllabic word sequence "ziehen die" would be /ts i: - @ n - d i:/. The realisation by one speaker was a disyllabic [tsi-ni] where [n] changed its position from the last segment of the first word (syllable coda) to the first segment of the following word (syllable onset).
3. A pause here is defined as a silent interval rather than a perceived pause.
4. The stimuli for the perception experiments and other material related to this study are available via http://www.coli.uni-sb.de/~cabr/ssw4/.
5. A z-score can be converted back easily into the absolute duration value by applying the following formula: absolute duration = (z-score $\times$ standard deviation) + mean duration.
6. The reasons for taking pause durations from the database instead of predicting them are: (a) only few intra-sentence pauses had to be modelled in our test corpus; (b) the break strength assumed here distinguishes only between minor boundary (without pause) and major boundary (with pause); (c) within-sentence pause durations are fixed (200 ms) in Klatt's model.
7. The spectrum information for all stimuli comes from the MBROLA voice and not from the database speaker.

References

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

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and van der Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesisers free of use for non commercial purposes. *Proceedings ICSLP 96*. Philadelphia, PA, pp. 1393–1396.

IPDS (1994). The Kiel Corpus of Read Speech. Vol. I. Kiel-CD#1. Retrieved April 10, 2002, from http://www.ipds.uni-kiel.de/publikationen/cd1.en.html.

Klatt, D.H. (1979). Synthesis by rule of segmental durations in English sentences. In B. Lindblom and S. Öhmann (Eds.), *Frontiers*

of Speech Communication Research. London, New York, San Francisco: Academic Press, pp. 287–299.

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

Portele, Th. (1997). Reduktionen in der einheitenbasierten Sprachsynthese. *Proceedings Fortschritte der Akustik, DAGA 97*. Kiel, Germany, pp. 386–387.

Riedi, M. (1997). Modeling segmental duration with multivariate adaptive regression splines. *Proceedings Eurospeech 97*. pp. 2627–2630.

Riley, M. (1992). Tree-based modelling of segmental duration. In G. Bailly, C. Benoit, and T.R. Sawallis (Eds.), *Talking Machines: Theories, Models, Designs*. Amsterdam, London, New York, Tokyo: North-Holland, pp. 265–273.

Schröder, M. and Trouvain, J. (2001). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *Proceedings 4th ISCA Workshop on Speech Synthesis*. Pitlochry, Scotland, pp. 131–136.

Taylor, P., Caley, R., Black, A.W., and King, S. (1999). Edinburgh Speech Tools Library. System Documentation Edition 1.2. Retrieved April 10, 2002, from http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/.

van Santen, J.P.H. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language 8*. pp. 95–128.

Wells, J.C. (1996). SAMPA for German. Retrieved April 10, 2002, from http://www.phon.ucl.ac.uk/home/sampa/german.htm.