Learning duration

Friedrich Neubarth, Hannes Pirker, Harald Trost
ÖFAI, Austrian Research Institute for Artificial Intelligence,
1010 Wien, Austria
{friedrich, hannes, harald}@ai.univie.ac.at

Abstract [*]

In this paper, we investigate the possibilities to enhance statistic modelling of segment duration for speech synthesis. In particular we look at the effects of gradually increasing size of training data and at specific problems of phonetic coding. We show that questions arising due to the inherent mismatch between cannon phonemic representation and phonetic realisation are best answered by statistical tests, in many cases it turns out that linguistic intuitions are in fact reflected by the empirical data.

Introduction

The quality of speech synthesis depends on a variety of factors, e.g., signal processing, intonation modelling or control of duration. Sometimes improvements in a certain area may not result in an increased overall acceptance by users. This can be explained by the contrast of occurring errors to the background of a more natural synthesis, which leads to higher expectations. Despite of this perspective, there is a lot to gain in each of these areas, in terms of overall performance of speech synthesis systems as well as for a deeper understanding of natural speech.

One of the levels where many factors are at work is segment duration. It is obvious that we will achieve good results only if we can handle the interplay of all of these factors. However, in this paper we will concentrate on a problem which is located on a very deep level: the disparity of phonological representation and phonetic realisation. In the next section we will present a short overview on the linguistic nature of this problem. Section 2 describes the speech corpus we have established and use in our investigation, section 3 deals with the implications on the prediction of segment duration and section 4 presents a study on how this problem can be tackled in a reasonable way.

1 Basic considerations

1.1 The phonetics-phonology mismatch

In the past few years it became more and more obvious that in the field of linguistics there is a gap between phonology, which often uses phonetic terminology for its own purposes but avoids a look at the “real” world of physical manifestations of sound – and phonetics, which depends in part on abstract, phonological classifications, but focuses on physical entities such as spectral shape, duration and intonation measures.

The problem has been acknowledged earlier, Fant states that “*the speech wave is not a very good image of our abstract notion of speech as a sequence of discrete invariable units selected from a finite inventory.*” [Fant, 1974] These days it was clear that phonetic research meant classifying cues in the acoustic signal, which are neither discrete nor strictly sequential and would rather function as carriers for abstract information than being mani-

[*] This work has been sponsored by the *Fonds zur Förderung der wissenschaftlichen Forschung (FWF),* Grant No. P13224. The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry of Education, Science and Culture.

festations of it. Phonology wanted to build upon the methods developed for phonetics and started to treat phonology and phonetics as compatible components of grammar. Most phonological theories up to today disregard the fundamental difference between these areas. However, it should be clear that a fruitful communication will become possible only if those differences are grounded in the theories and methodologies themselves.

From a technical point of view, one may be tempted to regard this discussion to be of less interest. But if one examines the assumptions behind technical procedures like segmentation, labelling, feature classification, more and more evidence is mounting that some mysterious shortcomings might arise from a too naive acceptance of theoretical base assumptions which prohibit a deeper view on the actual properties of spoken language.

In the context of speech synthesis it will be indispensable to analyse the particular steps one takes when establishing models about prosodic entities. In the following section we outline the considerations which guided the creation of a speech corpus of Standard Austrian German. The main goal was to obtain a better modelling of segment durations for a concatenative synthesiser. One of the concerns is to find an optimal machine-learning technique. Another area of research is the selection of an appropriate feature encoding for the data used by machine learning techniques. Both, technical and linguistic considerations have to be taken into account. We will show that a careful examination of certain properties of the data can lead to enhanced results in the modelling of durations of speech segments.

1.2 Segment duration

In the context of concatenative speech synthesis, it sounds pretty reasonable that the basic aim is to first identify segments of speech. Now, in the light of the previous discussion it becomes clear that segments are a rather hybrid entity. The labels they carry are derived from the abstract notion of a ‘phoneme’. Right from the start, this is not entirely true. On the one hand subparts of phonemes are identified as units (e.g., when occlusion and release/aspiration phase of stops are given separate status), on the other hand certain phoneme combinations are treated as a single unit. Crucially, the symbols are strictly sequentially ordered. This is necessary when we want to deal with segment durations, but it also leads to the fact that there is only a weak correspondence between segments and acoustic events. Segment boundaries tend to be compromises between competing acoustic signatures; sometimes they can only be assigned by subjective auditory control; and sometimes the need to assign a boundary overrides the fact that no boundary can be detected even subjectively. Segmentation errors are therefore methodologically intrinsic. In short, phonetic segmentation suffers from the mismatch of superimposing phonological categories onto an overlapping sequence of phonetic events. However, for concatenative speech synthesis this is all we need: a measure of similarity on a temporal scale. Let us investigate, what the most problematic issues on segmentation are:

- Segment boundaries may not be identifiable properly. Even when an ideal segmentation routine is assumed, there is an intrinsic error rate.
- A whole string of phonemes is matched onto a single phonetic segment, due to (total) assimilation or other deletion processes.
- Certain phonological distinctions are not straightforwardly reflected phonetically, e.g., geminates in German. The question is whether they should be reflected in the set of labels.

2 The speech corpus

Our corpus comprises approx. 50.000 phones (1.2 hours of actual speech) of Standard Austrian German from 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

(Marburg sentences, Berlin sentences and others), parts of which are also found in the PHONDAT corpus [Kohler 1994].

- Material where the information-structure is controlled for the analysis of the relation between focus structure and prosody. The sentences are elicited answers to questions (question-answer pairs). The basic structure of the answer sentences is a control matrix verb and two embedded conjoined infinitival groups with an object.
- 22 connected texts from newspaper.

For the coding of phonetic labels we use a slightly modified SAMPA-notation, adapted from the PHONDAT corpus. Deviations from the standard can be formulated in terms of rules, so the interconnectivity to other databases of German speech is guaranteed. In order to have full control over the canonical transcription, especially the coding of accent and syllabification, we use a specifically designed semi-automatic transcription tool to establish an extendable full-form word list. The canonical phonemic layer derived from this lexicon, which also encodes accent and morphological structure, can then be automatically linked to the actual phonetic segmentation by applying certain alignment rules.

The segmentation and annotation of phonetic labels was performed with STX[1]. We used an HMM based automatic segmentation tool, which had been trained on a manually created subset of the corpus. The results were corrected by hand. For the segmentation certain guidelines were obeyed. Phonemes from the canonical transcription missing in the acoustic signal (due to, e.g., schwa-deletion, stop-assimilation etc.) were explicitly indicated as missing in order to make sure that this information is retrievable.

The corpus is also annotated for intonation (ToBI-labels, Fujisaki accent and phrase commands), and prominence (manual labels for each syllable). The part of the corpus with controlled information structure is also annotated for topic-focus structure and with a rough syntactic tagging. Information on prosodic structure (feet, syllable structure, etc.) is dynamically derived and stored in the database.

All the data is stored in a proprietary format which is structured relationally but offers easy extensions by the ‘tag=value’ structure. This was necessary since the source data was created by different tools, each employing its own file format. The benefits are that it is easy to parse and manipulate the data and to create new outputs to other programs for further computation.

3 The size of training data

As a first step we investigated the effects of the size of training data on the behaviour of a simple and easy to perform machine learning technique like CART (classification and regression trees [Breiman et al., 1984]). As a standard coding scheme we use the feature set proposed by Riedi [1998]. A more detailed discussion on the design of this set will be presented in the next section. We divide the data randomly into 10 subsets of equal size. Each of these subsets is used as a test sample in a 10-fold cross-validation routine. The size of the training set is varied from 1 subset to all of the remaining 9 subsets (i.e., approx. from 5.000 to 45.000 items).

As measures for the performance of the learning algorithm we use the correlation (Cor) between predicted and original duration values in the test set and the root of the mean squared error (RMS). The whole procedure was performed using the CART package `tree` provided with the statistical software R.[2] The control variable of the CART algo-

[1] STX is a workstation for acoustics, speech and signal processing, developed at the Institute for Acoustic Research of the Austrian Academy of Sciences. See: http://www.kfs.oeaw.ac.at/

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

rithm is minimal deviance (mindev), which determines when the algorithm stops splitting nodes, indirectly it controls the number of terminal nodes.

When mindev is too high (at 0.005), the results are relatively poor in general and they do not show any significant difference when varying the size of the training set. When mindev is set to 0.001 (its default), the results seem to become rather stable from 6 training subsets on (approx. 30.000 items).

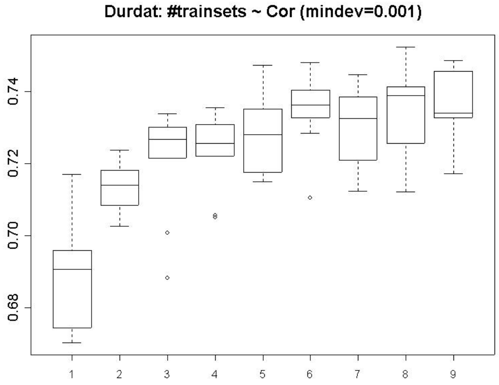

Figure 1 : correlation coefficient for different amount of training data

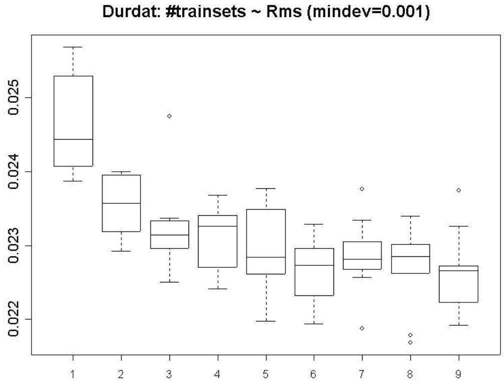

Figure 2: RMS values coefficient for different amount of training data

If we increase the granularity of the CART tree by lowering the mindev value to 0.0005, it can be seen that the correlation and error measures improve in general and the point where further improvements become marginal is reached already with a training set consisting of 4-5 subsets.

Using logarithmic time encoding instead of linear time values did not lead to any significantly different results. If we look at smaller test sets, in particular if the size of the test set is reduced in a way that the ratio between training and test set is kept to a constant value of 9:1, the picture changes slightly: the curve of improvement becomes flatter, marginal improvement is reached later at a size of 7 subsets for the training set.

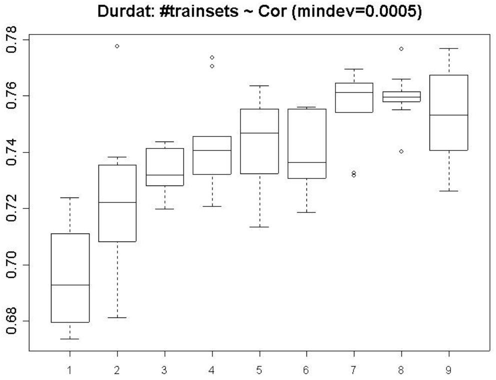

Figure 3: correlation coefficient for different amount of training data

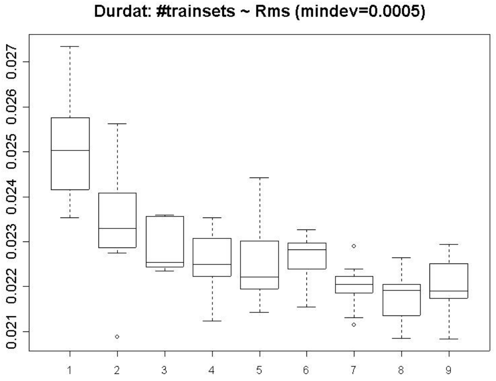

Figure 4: RMS values coefficient for different amount of training data

This justifies the need for a larger set of data, especially when we also want to investigate the behaviour of subsets of a corpus, as we will present in the following section.

4 Phonetic coding

In this section we will address two problems arising in this context: of coding of phonetic items, namely deletion and geminates.

The first problem is how to deal with segments, which are phonetically not realised, though being part of the canonic phonemic representation. There are 3 prominent manifestations: total assimilation of stops, an effect which also occurs across word boundaries (e.g. "*mit dem*": [mIdem] instead of [mitdem]), rather short combinations of vowels+a-schwas (e.g., "*der*": [d6] instead of [de6], "*für*": [fY] instead of [fY6]) and schwa-deletion after nasals and liquids (e.g., "*reden*": [Re:dn] instead of [Re:d@n]).[3] Similar phenomena have been studied recently by Brinckmann & Trouvain [2001].

In Figure 5 the duration of both stop segments (`dur(C1)+dur(C2)`) is plotted against the duration of the second consonant (`dur(C2)`). The majority of instances is distributed neatly around the line of equal duration, and there is a considerably large class where the duration of the second consonant equals the sum of durations (i.e., the first consonant is assimilated: `dur(C1)=0`). We definitely do not want to learn 0-durations. So the empty segment is left out from the training set. Assuming that we have an independent algorithm which decides whether assimilation takes place or not, the question is where to represent the information that a neighbouring segment is empty.

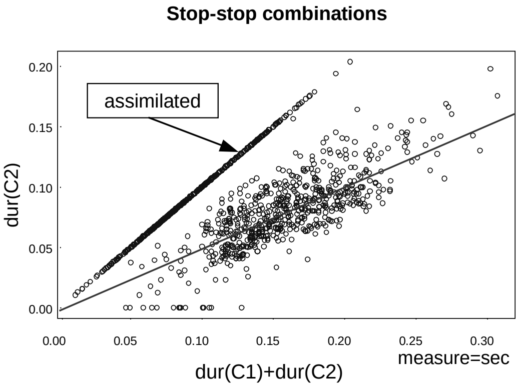

Figure 5: Scatter plot of stop-stop combinations

The following problems for feature coding arise in the context of phonetically not realised segments ('empty segments'). As the identity of neighbouring segments quite strongly influences the duration of sounds, it has to be decided, whether an empty segment should still be treated as a neighbour or rather be ignored and whether the information that a segment has an empty neighbour should be coded as a special feature. An illustration, how this problem can be tackled will be presented below.

Figure 6 shows the distribution of vowels and a-schwas in the same fashion as before. Here we see that only very short sequences of these combinations assimilate (indicated by the vertical line at 100ms). One could in principle apply a different strategy and code all these combinations as unique segments, however, this would enlarge the feature space for coding the identity of vowels.

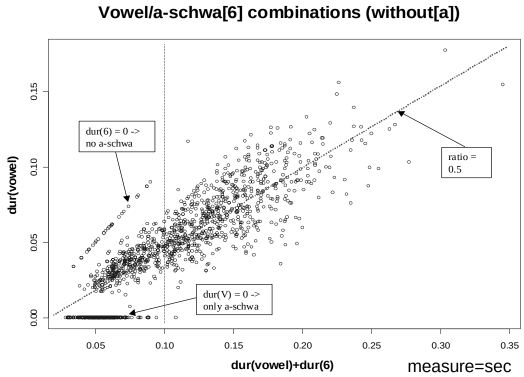

Figure 6: Scatter plot of vowel / a-schwa combinations

[3] We are grateful to an anonymous reviewer to point out that the origin of these mismatches is not straightforward. While stop-stop assimilations are undisputedly a phonetic process (domain independent, not predictable by rules), this is not true for the 2 other processes. Defective vowel/a-schwa sequences are an artifact of labelling, and schwa-deletion could in principle be incorporated in the canonic representation. Still, one has to deal with these effects in one or the other way, an evaluation of possible procedures will be presented in the remainder of this paper.

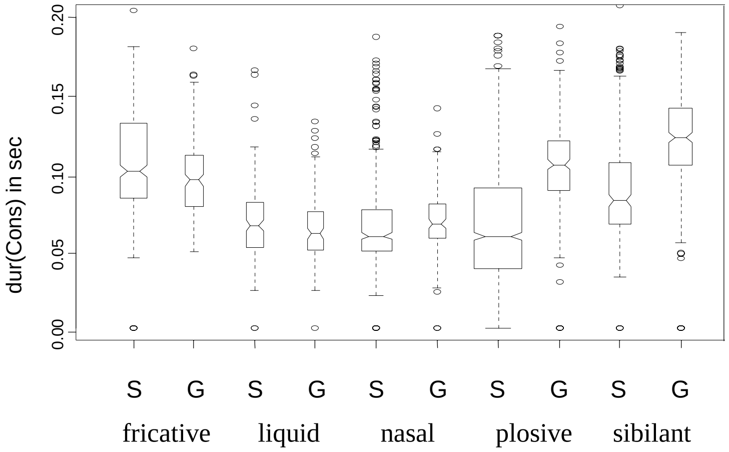

Figure 7: duration of simplex(S) and geminate(G) consonants.

A second area of phonetic/phonological concern are geminates (often tagged as ambisyllabic consonants, cf. [Moebius and v. Santen, 1996]). In Standard German there is no total agreement whether they phonetically exist.

In certain Austrian dialects it is quite obvious that there are geminates, whereas other dialects such as Carinthian abandoned geminates altogether, but then the preceding vowels are realised long ('*Ersatzdehnung'*), thus preserving the metrical template. The question is, whether geminates are phonetically detectable in Standard Austrian German and if so, whether it makes sense to include it as an additional feature.

Figure 7 displays the duration values for consonants in potential gemination sites, i.e. after an accented vowel, which is not followed by a separate coda-consonant. The graph is split according to phoneme classes, in each class the first vertical plot represents the simplex forms and the second the geminates. It can be seen, that the difference between the two is not significant in most classes[4], however, there is an absolutely clear difference with plosives and sibilants.

Let us return to the setup of a test on various hypotheses how to encode these phonetic or phonological peculiarities. Again we use the feature set proposed by Riedi [1998]. It has a 2-dimensional matrix of features coding the phoneme type (A*n* and B*n*) for the actual segment and the first and second neighbour on both sides[5], one feature for coding the segment's status in the syllable structure (sc) plus number and position within the syllable (sn, sp), five levels of accent (ac) and number and position of hierarchically lower items on the word (wn, wp), foot (fn, fp) and sentence/phrase (po) level. The test is designed as such that CART (with mindev=0.001) is applied to the full set of data using 10-fold cross-validation. The relevant test cases are listed in Table 1.

Cases *a2* and *a4* mark neighbours with 0-duration as a separate feature. Cases *a1/a2* [6]

[4] Overlapping notches in the `boxplot` indicate lack of significance.

[5] A-features in consonants roughly encode 'source' (e.g., *unvoiced plosive, nasal*), B-features 'place' (e.g., *bilabial, palatal*). A1 refers to phoneme-class information of the actual segment, A9/A0 refer to the left neighbors, A2/A3 to the right neighbors.

[6] 'Default coding' means that no alternative features are introduced and the information about neighboring segments is directly drawn from the

Code	Method
a1	Default
a2	Default + marking non-realization of neighbors as special feature
a3	Skipping empty segments
a4	Skipping empty segments + marking non-realization of neighbors as special feature
a5	Leaving out features: '*fn/fp*'
a6	Coding geminates in all '*An*' features ('*A9-A3*')
a7	Coding geminates in feature '*A1*'
a8	Coding geminates as special feature ('*ge*')
a9	Coding geminates in feature '*syllable structure*' ('*sc*')
a10	More refined coding of syllable structure in '*sc*'
a11	Leaving out features: '*wn/wp*'
a12	Leaving out features: '*wn/wp/fn/fp*'

Table 1: Coding schemes

versus *a3/a4* exploit the options of encoding neighbours as described in Fn.6. Cases *a6-a9* deal with options to encode geminates. *a10* is an attempt to provide the algorithm with more detailed information about syllable structure. *a5*, *a11* and *a12* skip the information about structure on the level of word and foot. The correlation and RMS values are shown in the following two figures:

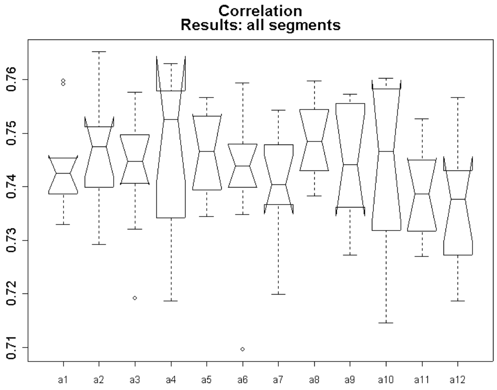

Figure 8: Correlation of predicted durations from CART on different coding schemes.

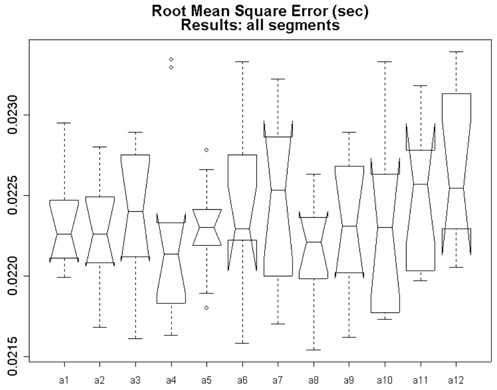

Figure 9: RMS error of predicted durations from CART on different coding schemes

The picture emerging from these results is that some of the options differ slightly from the default setup, but statistical significance is not reached in any of the cases. What can be said is that leaving out the features 'wn/wp/fn/fp' decreases the quality of the prediction and that coding syllable structure in a more refined way does not improve the results.

What about empty segments and geminates? It might be useful to test these classes separately. Let us start with geminates and reduce the test-set to only potential geminate consonants as described in Figure 7. With this subclass correlation is generally higher, but especially case *a8* and *a9* seem to make an improvement, *a8* (coding geminates as a special feature) more than *a9* (coding geminates as ambisyllabic in the syllable structure feature 'sc'.) Cases *a6* and *a7* which code geminates as special phoneme types in 'A1' do not improve the results.

canonic phoneme string. 'Skipping empty segments' then implies that only segments with a phonetic realization (hence, duration) are taken as neighbors.

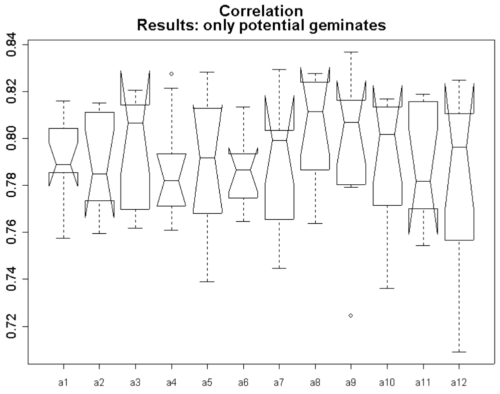

Figure 10: Correlation of predicted durations from CART on different coding schemes

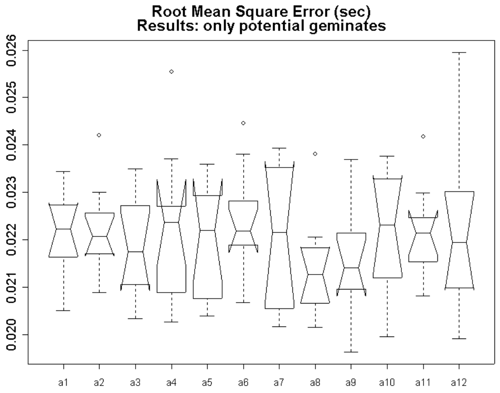

Figure 11: RMS error of predicted durations from CART on different coding schemes

The effect of encoding empty neighbours is even less significant. The results show that deleting the information about empty neighbours raises correlation but also the RMS error increases, which is rather surprising and seems to indicate that information about neighbours is necessary in order to keep the feature space balanced. An additional feature for ‘empty neighbour’ does not improve the results at all.

Conclusion

In this paper we tried to show that when applying machine learning techniques to a corpus of speech data, it makes sense to carefully design the setup of features encoded in the training corpus. With respect to the size of training data, it turned out that using a larger corpus (around 50.000 phonemes) leads to better performance, especially when analysing smaller classes of linguistic items. Many problems cannot be solved in a straightforward way. In our work on Standard Austrian German it was shown that if the feature space is altered in order to reflect the special status of geminates and segments which have a phonetically not realised canonical neighbour, the results slightly improve, but not in a statistically significant way. Future research will have to apply the same methods on higher level entities like metrical structure, phonological phrasing and focus-structure.

References

Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. (1984) *Classification and Regression Trees*. The Wadsworth Statistics/Probability Series, Wadsworth International Group, Belmont, CA.

Brinckmann C. and Trouvain J. (2001) *On the Role of Duration Prediction and Symbolic Representation for the Evaluation of Synthetic Speech*. In “Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis”, Scotland.

Fant, G. (1974) *Analysis and synthesis of speech processes*. In Malmberg, B. “Manual of Phonetics”, North-Holland Publishing Comp., Amsterdam, London, pp. 173-277.

Kohler K. (1994) *Lexica of the Kiel PHONDAT Corpus, Read Speech*. Institut für Phonetik und Digitale Sprachverarbeitung, Universität Kiel, Vol.I, Arbeitsberichte Nr.27.

Moebius B. and Santen J.van (1996) *Modelling Segmental Duration in German Text-to-Speech Synthesis*. Proc. of ICSLP'96, Vol.4, Philadelphia, PA, pp. 2395-98.

Riedi, M.P. (1998) *Controlling segmental duration in speech synthesis systems*. PhD thesis, TIK-Schriftenreihe Nr. 26, ETH Zürich.