MODELING SEGMENTAL DURATION IN GERMAN TEXT-TO-SPEECH SYNTHESIS

Bernd Möbius and Jan van Santen

Speech Synthesis Research Department
Bell Laboratories, Murray Hill, NJ, USA

ABSTRACT

This paper reports on the construction of a model for segmental duration in German. The model predicts the durations of speech sounds in various textual, prosodic, and segmental contexts. It has been implemented in the German version of the Bell Labs text-to-speech system [18, 12]. The construction of the duration system was made efficient by the use of an interactive statistical analysis package that incorporates the approach outlined in [23]. The results are stored in tables in a format that can be directly interpreted by the TTS duration module. Tables are constructed in two phases: inferential-statistical analysis of the speech corpus, and parameter estimation. The overall correlation between observed and predicted segmental durations is .896.

1. INTRODUCTION

The primary goal of this study was to analyze and model durational patterns of natural speech in order to achieve an improved naturalness of synthetic speech. In natural speech, segmental duration is strongly context dependent. For instance, in our German speech database we observed instantiations of the vowel [e] that were as short as 35 ms in the word "jetzt" and as long as 252 ms in the word "Herren". Among the most important contextual factors are the position of the word in the utterance, the accent status of the word, syllable stress, and the segmental context. These factors and the levels on them jointly define a large feature space. The task of the duration component of a text-to-speech (TTS) system is to reliably predict the duration of every phone depending on its feature vector. An additional requirement is that the feature vector be computable from text.

The prevalent type of duration model is a sequential rule system as proposed by Klatt [7, 8]. Starting from some intrinsic value, the duration of a segment is modified by successively applied rules. Models of this type have been developed for several languages including American English [1, 13], Swedish [4], German [9], French [2], and Brazilian Portuguese [17]. When large speech databases and the computational means for analyzing these corpora became available, new approaches were proposed based on, for example, Classification and Regression Trees (CART) [14, 15] and neural networks [3]. It has been shown, however, that even huge amounts of training data cannot exhaustively cover all possible feature vectors [23]. An alternative method, manual database construction, is only feasible if the factorial space is not too large. But in the duration analysis for our English TTS system a minimum of 17,500 distinct feature vectors were observed [22]. Since the factorial scheme for German bears some resemblance to the one for English, the number of distinct feature vectors can be assumed to be in the same order of magnitude, making a manual database construction impractical. Rare vectors cannot simply be ignored because the combined frequency of rare vectors almost guarantees the occurrence of at least one unseen vector in any given sentence. Thus, the duration model has to be capable of predicting – by some form of extrapolation from observed feature vectors – durations for vectors insufficiently represented in the training material. CART based methods are known for poorly coping with data sparsity, because they lack this extrapolation capability. Extrapolation is complicated by interactions between the factors. Factor interactions prevent simple additive regression models [6], which have good extrapolation properties, from being an efficient solution. This assertion holds even though the interactions are often regular in the sense that the effects of one factor do not reverse the effect of another factor.

The solution proposed by van Santen [19, 20, 23] is the application of a broad class of arithmetic models, *sums-of-products models*. This approach takes advantage of the fact that most interactions are regular which allows describing these interactions with equations consisting of sums and products. Addition and multiplication are sufficiently well-behaved mathematically to estimate parameter values even if the frequency distribution of feature vectors in the database is skewed. This method has recently been shown to be superior to CART-based approaches: It needs far fewer training data to reach asymptotic performance; this asymptotic performance is better than for CART; difference in performance grows with the discrepancy between training and test data; adding more training data does not improve the performance of CART-based approaches [11]. Van Santen's method has been applied to American English [22, 23] and Mandarin Chinese [16], and we used it in the present study. We will now describe the construction of the duration system in more detail.

2. SPEECH DATABASE

Our analysis of segmental durations in natural speech is based on the Kiel Corpus of Read Speech, recorded and manually segmented

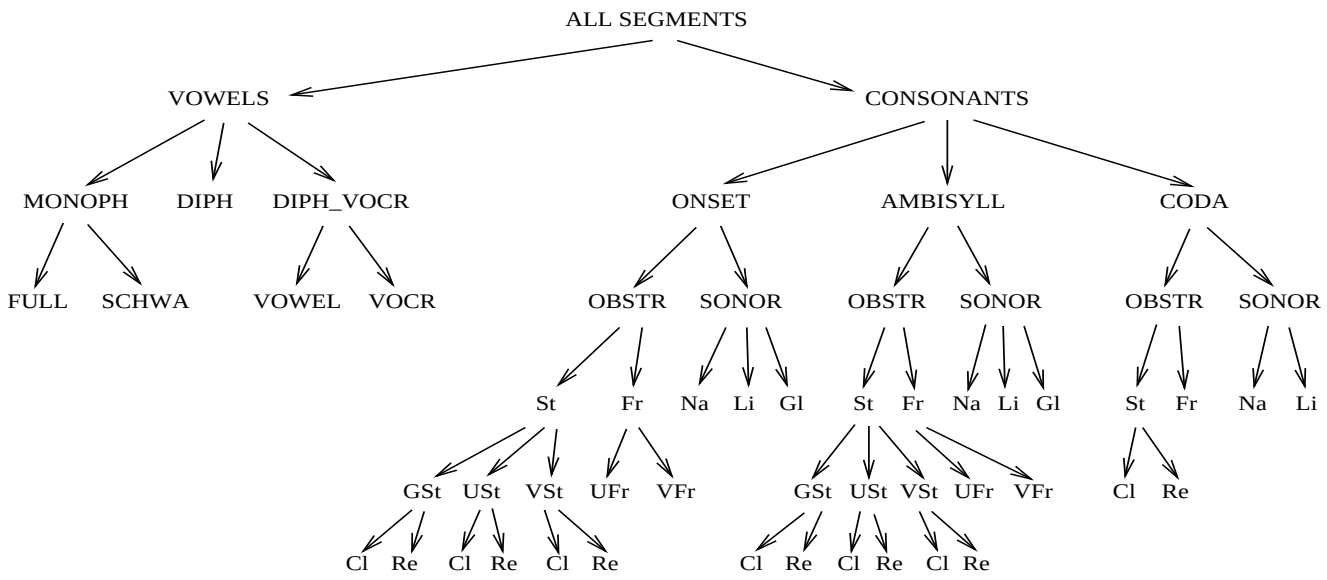

Figure 1: Category tree of the German duration system; MONOPH = monophthongs, DIPH = diphthongs, DIPH_VOCR = diphthongs involving [ɐ], VOCR = [ɐ], AMBISYLL = ambisyllabic, OBSTR = obstruents, SONOR = sonorants, St = stops, Fr = fricatives, Na = nasals, Li = liquids, Gl = glides, GSt = glottal stops, USt/VSt/UFr/VFr = unvoiced/voiced stops/fricatives, Cl = stop closure, Re = stop release.

at the Kiel phonetics institute and published on CDROM [5]. The disk contains speech and label files; the latter provide: orthography; 'canonical' transcription according to standard German pronunciation; transcription of actually realized speech (see [10] for details). Two speakers produced the entire text material. We selected the renditions of the male speaker 'k61'. Some sanity and consistency checks were performed on the labeled data; for instance, all utterance-initial stop closure data were excluded from the analysis. The database ultimately yielded a total of 23,490 segments: 6,991 vowels and 16499 consonants. We computed feature vectors for all the segments in the database. The following factors were included in the annotation:

- segment identity
- segment type; levels: front, mid, and back vowels, voiced and unvoiced stops and fricatives, nasals, liquids, glides, silence
- word class; levels: function word, content word, compound
- position of phrase in utterance
- phrase length (in number of words)
- position of word in phrase; levels: initial, medial, final
- word length (in number of syllables)
- position of syllable in word; levels: initial, medial, final
- stress; levels: primary, secondary, unstressed
- segment position in syllable; levels: onset, nucleus, coda
- segmental context; levels: identities of first, second, and third segments to left and right
- segment type context; levels: type of first, second, and third segments to left and right
- boundary type; levels: phrase, word, syllable, no boundary to left and right
- context segment cluster; levels: (e.g.) voiceless obstruents in coda, empty onset, diphthong nucleus, etc., to left and right

It is important to note that this database was not optimal for the purpose of duration system construction, because no attempt was made to cover the greatest number of distinct feature vectors. By contrast, in their study of Mandarin Chinese duration, Shih and Ao [16] used greedy methods [21] to select a few hundred sentences that covered the same set of feature vector *types* as the much larger set of 15,000 sentences from which the sentences were drawn.

3. CATEGORY TREE

Constructing the duration system requires two major steps: setting up a category tree that splits up the factorial space, typically in terms of broad phoneme classes and intra-syllabic location, and selecting a particular sums-of-products model for each leaf of the tree.

Figure 1 shows the category tree of the German duration system. The tree represents a factorial scheme, i.e., the set of factors and distinctions on these factors that are known or expected to have a significant impact on segmental durations. Knowledge-based distinctions in the factorial scheme rely on three types of empirical information:

1. **Conventional distinctions** based on phonetic and phonological features assigned to the segments, e.g., between vowels and consonants or between continuant and abrupt consonants. The underlying criteria for these distinctions are language independent.

Nucleus	ə	ɪ	ʏ	ʊ	ɐ	ɔ	e	a	œ						
	64	89	99	101	103	105	113	116	125						
Nucleus	i:	u:	y:	e:	o:	ø:	ɛ:	a:	aɪ	aʊ	ɔʏ				
	106	115	132	141	147	171	175	178	150	150	153				
Onset	dCl	gCl	bCl	ʔCl	tCl	kCl	pCl	bRe	dRe	gRe	ʔRe	tRe	pRe	kRe	
	36	38	52	53	71	91	119	11	11	18	10	11	12	22	
Onset	v	z	ʒ	h	ç	f	s	ʃ	n	m	l	r	j		
	66	92	93	51	80	90	91	92	52	57	63	71	71		
Ambisyll	dCl	gCl	bCl	ʔCl	tCl	kCl	pCl	bRe	dRe	gRe	ʔRe	tRe	pRe	kRe	
	48	50	60	30	55	60	74	11	11	13	10	24	25	36	
Ambisyll	ʒ	v	z	h	x	ç	f	s	ʃ	n	m	ŋ	r	l	j
	55	59	71	52	72	90	96	98	105	56	71	75	46	51	84
Coda	kCl	tCl	pCl	pRe	tRe	kRe	ç	x	f	s	ʃ	n	ŋ	m	r	l
	47	47	59	13	15	16	84	92	96	116	132	77	80	85	51	64

Table 1: Corrected means [in ms] for all segments in the database.

2. **Qualitative observations** as reported in the (sparse) research literature on segmental duration in German, such as: "Utterance-final lengthening affects the final two syllables only if the penultimate syllable is stressed, otherwise only the final syllable is affected" [9].
3. **Exploratory studies.** In a pilot experiment we found that the single most important segmental context factor for vowel duration was whether or not the syllable coda was empty, in other words: whether the vowel was the nucleus of an open or a closed syllable. The segmental composition of the coda was significantly less important.

Since our main goal is to develop a duration module as part of a TTS system, an important additional requirement in setting up the category tree was that the factors can be computed from text by the text analysis components of the system. The tree structure reflects a compromise between the attempt to obtain homogeneous classes by fine sub-categorization and retaining a reasonable number of observations at each leaf of the tree. Note the we use *homogeneity* not in the sense of the cases at a leaf having similar durations (minimal variance, as in CART), but in the sense that the same factors have the same effects on these cases, so that their behavior can be captured by one and the same sums-of-products model. The following categorical distinctions were made:

Vowels vs. consonants. This distinction is rather obvious and based on well-established phonetic and phonological knowledge, e.g., the observation that some factors like stress and speaking rate have, quantitatively speaking, very different effects on vowels than on consonants.

Vocalic distinctions. Vowels were sub-categorized into central vowels (schwa), diphthongs, and full (non-central) monophthongs. An additional distinction was made for diphthongs that involve the low central vowel [ɐ] as a result of [r] vocalization. Whereas 'regular' diphthongs [aɪ, aʊ, ɔʏ] are each treated as one segment in the acoustic inventory of the TTS system, diphthongs involving [ɐ] are generated by concatenating two segments; thus, durations have to be assigned to both components of these diphthongs.

Consonantal distinctions. The top level distinction among the consonants was based on the location in the syllable. Consonants are classified as being located in the onset or coda of the syllable, or as being ambisyllabic. All single intervocalic consonants are considered ambisyllabic. The next level of distinction was based on manner of articulation: stops, fricatives, nasals, liquids, and glides. Stops are subdivided into a closure and a release phase. In the onset and ambisyllabic locations, obstruents are further classified according to their voicing status. The voicing opposition is not applicable to obstruents in the syllable coda in German; exceptions to this rule (as for the [d] in "Redner") are too small in number to justify a separate leaf in the tree.

4. PARAMETER ESTIMATION

In this analysis, we did not explore the full space of sums-of-products models; for practical reasons, we only fitted the additive and the multiplicative model. Since the multiplicative model had a uniformly better fit, we only report results on the latter. By fitting the multiplicative model, the resulting parameter estimates can be considered as approximations of the marginal means in a hypothetical data base where each factorial combination occurs equally often (a balanced design). For this reason, we call these parameter estimates *corrected means*. Table 1 shows the best estimates of corrected means for the entire database. Table 2 gives correlations and root mean squared deviations of observed and predicted data. Because of differences in the numbers of observations and in the ranges of durations, these statistics are not strictly comparable with each other. The overall correlation between observed and predicted segmental durations for the entire database is .896.

5. SUMMARY

We constructed a quantitative model of segmental duration in German by estimating the parameters of the model based on a segmented speech database. This approach uses statistical techniques that can cope with the problem of confounding factors and factor levels, and with data sparsity. The results show rather homogeneous patterns in that speech sounds within a given segment class gener-

Position	Segment class	Observ.	Corr.	RMS
Nucleus	full	4552	.80	25
	schwa	906	.71	18
	diph	693	.74	30
	vow (bef. [ɐ]	840	.75	22
	[ɐ] (aft. vow)	840	.73	15
Onset	USt-Cl	1123	.61	17
	VSt-Cl	795	.74	15
	USt-Re	868	.86	9
	VSt-Re	820	.61	5
	UFr	1096	.66	20
	VFr	368	.72	16
	Na	451	.46	18
	Li	487	.42	17
	Gl	31	.75	14
Ambisyll	USt-Cl	458	.55	17
	VSt-Cl	699	.64	13
	USt-Re	410	.63	14
	VSt-Re	792	.46	4
	UFr	558	.80	16
	VFr	251	.63	14
	Na	681	.59	15
	Li	293	.36	15
	Gl	29	.95	14
Coda	St-Cl	1187	.67	19
	St-Re	1187	.88	12
	Fr	1369	.86	25
	Na	1917	.66	22
	Li	306	.67	20

Table 2: Results of model parameter estimation: Syllable part, segment type (for legend see Figure 1), number of observations, correlation and root mean squared deviation of observed and predicted data.

ally exhibit similar durational trends under the influence of the same combination of factors. Among the most important factors are: a) syllable stress (for nuclei, and to some extent for stops and fricatives in the onset); b) word class (for nuclei); c) presence of phrase and word boundaries (for coda consonants, and to some extent for nuclei). The analysis yields a comprehensive picture of durational characteristics of one particular speaker. The duration system has been implemented in the German version of the Bell Labs text-to-speech system.

6. REFERENCES

1. J. Allen, S. Hunnicutt, and D.H. Klatt. *From text to speech: The MITalk system*. Cambridge University Press, Cambridge, 1987.
2. K. Bartkova and C. Sorin. A model of segmental duration for speech synthesis in French. *Speech Communication*, 6:245–260, 1987.
3. W.N. Campbell. Syllable-based segmental duration. In G. Bailly, C. Benoit, and T. Sawallis, editors, *Talking machines: Theories, models, and designs*, pages 211–224. Elsevier, 1992.
4. R. Carlson and B. Granström. A search for durational rules in a real-speech database. *Phonetica*, 43:140–154, 1986.
5. Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel. *The Kiel corpus of read speech, vol. 1*, 1994. CDROM.
6. N. Kaiki, K. Takeda, and Y. Sagisaka. Statistical analysis for segmental duration rules in Japanese speech synthesis. In *Proc. of ICSLP-90*, pages 17–20, Kobe, 1990.
7. D.H. Klatt. Interaction between two factors that influence vowel duration. *J. Acoust. Soc. Am.*, 54:1102–1104, 1973.
8. D.H. Klatt. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *J. Acoust. Soc. Am.*, 59:1209–1221, 1976.
9. K.J. Kohler. Zeitstrukturierung in der Sprachsynthese. *ITG-Fachbericht*, 105:165–170, 1988.
10. K.J. Kohler. Lexica of the Kiel PHONDAT corpus, vol. 1/2. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung (AIPUK)*, 27/28, 1994.
11. A. Maghbouleh. An empirical comparison of automatic decision tree and hand-configured linear models for vowel durations. In *Proc. of SIGPHON-96*, Santa Cruz, 1996.
12. B. Möbius, J. Schroeter, J. van Santen, R. Sproat, and J. Olive. Recent advances in multilingual text-to-speech synthesis. In *Fortschritte der Akustik — DAGA '96*. DPG, Bad Honnef, 1996.
13. J.P. Olive and M.Y. Liberman. Text to speech – an overview. *J. Acoust. Soc. Am.*, 78 (Suppl. 1):S6, 1985.
14. J.F. Pitrelli and V.W. Zue. A hierarchical model for phoneme duration in American English. In *Proc. of Eurospeech-89*, pages 324–327, Paris, 1989.
15. M.D. Riley. Tree-based modeling for speech synthesis. In G. Bailly, C. Benoit, and T. Sawallis, editors, *Talking machines: Theories, models, and designs*, pages 265–273. Elsevier, 1992.
16. C. Shih and B. Ao. Duration study for the Bell Laboratories Mandarin text-to-speech system. In J.P.H. van Santen, R.W. Sproat, J. Olive, and J. Hirschberg, editors, *Progress in speech synthesis*. Springer, 1996.
17. A.R.M. Simoes. Predicting sound segment duration in connected speech: An acoustical study of Brazilian Portuguese. In *Workshop on Speech Synthesis*, pages 173–176, Autrans, 1990. ESCA.
18. R. Sproat and J. Olive. Text to speech synthesis. *AT&T Technical Journal*, 74(2):35–44, 1995.
19. J.P.H. van Santen. Contextual effects on vowel durations. *Speech Communication*, 11:513–546, 1992.
20. J.P.H. van Santen. Analyzing N-way tables with sums-of-products models. *J. Math. Psychology*, 37(3):327–371, 1993.
21. J.P.H. van Santen. Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech and Language*, 7:49–100, 1993.
22. J.P.H. van Santen. Timing in text-to-speech systems. In *Proc. of Eurospeech-93*, volume 2, pages 1397–1404, Berlin, 1993.
23. J.P.H. van Santen. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8:95–128, 1994.