Reducing Segmental Duration Variation by Local Speech Rate Normalization of Large Spoken Language Resources

Hartmut R. Pfitzinger

Department of Phonetics and Speech Communication
University of Munich, Schellingstr. 3, 80799 München, GERMANY
hpt@phonetik.uni-muenchen.de

Abstract

We developed a time-domain normalization procedure which uses a speech signal and its corresponding speech rate contour as an input, and produces the normalized speech signal. Then we normalized the speech rate of a large spoken language resource of German read speech. We compared the resulting segment durations with the original durations using several three-way ANOVAs with phone type and speaker as independent variables, since we assume that segment duration variation is determined by segment type (intrinsic duration), by the speaker (speech rate, sociolect, ideolect, dialect, speech production variation), and by linguistic effects (context, syllable structure, accent, and stress). One important result of the statistical analysis was, that the influence of the speaker on segment duration variation decreased dramatically (factor 0.54 for vowels, factor 0.29 for consonants) when normalizing speech rate, despite the fact that sociolect, ideolect, and dialect remained almost unchanged. Since the interaction between the independent variables speaker and phone type remained constantly, the hypothesis arises, that this interaction contains most of the speaker-specific information.

1. Introduction

Estimating phone durations is not only a central concern in information theoretical approaches to the human speech code. It is also a source of definite improvements on spoken language processing methods. Automatic speech segmentation gained substantially (Stöber & Hess, 1998), and automatic speech recognition is expected to gain (Pols, 1999). The estimation of phone durations is also an important problem for automatic speech generation.

The quality of automatically synthesized speech strongly depends on the underlying timing models, which, at the present time, are mostly constructed on speech segment duration measurements based on large spoken language resources. Each class of speech segments (e.g. phonemes, sub-phonemic variants, phones) shows a significant amount of duration variation even if the underlying speech corpus reaches saturation regarding segmental duration. This means, further enlargement of the speech corpus would not further increase the standard deviations of the segmental durations.

The question arises as to whether it is possible to reduce the amount of unexplained duration variation to discover the intrinsic segmental durations. In this paper the influence of speech rate normalization on phone duration variation is investigated experimentally.

Since speech rate is a continuously varying prosodic feature we use a "momentary" or local approach to acoustic speech rate measurements, which was introduced earlier by Pfitzinger (1998; 1999; 2001). It is based on a linear combination of local syllable rate and local phone rate and, in contrast to just syllable rate or phone rate, it is well-correlated with perceived local speech rate ($r = 0.91$) since it represents the linguistic structure of words more accurately. The output of this method is a time-varying speech rate contour.

1.1. Vowel durations and standard deviations

Measurement and analysis of phone durations was a scientific topic for many years (Menzerath & de Oleza S. J., 1928). Weitkus (1931, p. 11) estimated an overall mean duration of 161.6 ms on a spoken German corpus comprising 7480 phones. He claims that the corpus had an average speaking rate, but in our view his results seem rather large. Weitkus (1931, p. 20) also estimated mean vowel durations between 88.3 ms for /ə/ and 386.2 ms for /ɛ:/.

The standard deviations of German vowel durations reported by Tillmann et al. (1990) are roughly between 20–40 ms depending on the vowel category. The overall mean vowel duration was 88.0 ms. The duration values were estimated from 5761 vowel realizations of the PhonDatI spoken language resource.

The mean duration measurements deviate significantly from vowel data reported by Heid (1998, p. 253) whose measurements are based on 10300 vowel realizations taken from the PhonDatII spoken language resource (see section 4.). Until now, this was to our knowledge the largest survey on German phone durations. The overall mean vowel duration was 75.6 ms. It is worth mentioning that the standard deviations reported in the studies summarized here are in basic correspondence.

1.2. Duration models

Kohler (1992a) conducted duration measurements on stressed vowels taken from a third of the PhonDatII spoken language resource. He found mean durations between 66.9 ms and 168.4 ms depending on the factors ±*last syllable in word*, ±*tense*, and ±*final*.

Möbius & van Santen (1996) analysed the durations of 16499 consonants and 6991 vowels taken from one speaker, 'k61', of the Kiel Corpus of Read Speech (IPDS, 1994). They neither reported measured durations nor standard deviations but predicted segment durations, which all were significantly larger than the measurements by Heid (1998, p. 249). To predict phone durations, they built a category tree as suggested by Riley (1992, p. 267).

Much work in this field was done by Klatt (1973; 1975; 1976; 1987), Kohler (1983; 1986b; 1986a; 1990; 1992a; 1997), Carlson & Granström (1986), Bartkova & Sorin (1987), and van Santen (1990; 1992a; 1992b; 1994; 1997; 1998), who split duration models into four categories (van Santen, 1993): i) the *sequential rule system*, in which intrinsic phone durations were modified by successively applying rules, ii) the *purely additive/multiplicative model*, where actual phone duration is a sum/product of a number of contextual parameters, iii) *category/factor tree*-based systems, where a number of nodes determines the path to a

leave providing the actual duration, and finally iv) *stochastic models* based on neural nets or HMMs.

In the factor tree built by Wang (1997, p. 129) (see also Pols (1999, p. 13)) speech rate as the first of four factors is split into three categories: fast, average, and slow speech. This small number of levels could be a consequence of limitations of the underlying corpus size. Wang first considered 11 factors and later restricts himself to four factors (speech rate, stress, syllable location in a word and in a sentence). Even the hugest spoken language resource provides not enough entities in each leave when there are too many nodes in a factor tree.

To increase the number of entities per leave, it is desirable to eliminate the factor speech rate by normalization of local speech rate changes. For this task a reliable normalization procedure is required.

2. Estimation of local rates

In (Pfitzinger, 1996) we introduced a mathematically sound formula to estimate local rates of speech units (e.g. phones, syllables, words, morphemes):

The distances between subsequent speech unit marks S_i falling in a window w of constant length (e.g. 625 ms) were accumulated and then divided by their number. The reciprocal of the quotient is a measure for the local rate of the underlying speech unit:

$$\text{rate}_{LR} = \frac{\frac{S_{l+1}-w_L}{S_{l+1}-S_l} + \frac{w_R-S_r}{S_{r+1}-S_r} + r - l - 1}{S_{l+1} - w_L + w_R - S_r + \sum_{i=l+1}^{r-1} S_{i+1} - S_i},$$

were w_L is the left and w_R is the right window boundary. Since the left (S_l) and the right (S_r) segment most frequently are covered only partially by the accumulation window, they have to be accumulated proportionately to guarantee a constant window length. This procedure leaves slight discontinuities in the resulting curve of the local rate.

A second method which is very time-consuming but removes any discontinuities from the resulting curves is described in brief: The first step is to estimate a new time-domain signal that corresponds to the underlying speech signal. All values of the new signal which fall in a speech unit receive the reciprocal of its duration. The second step is to convolute the signal with a Hanning window of e.g. 625 ms length. The result is a curve representing the local rate of the underlying speech units. It is similar to the result of the first method.

In this paper we applied a third method presented in (Pfitzinger, 2001) which is mathematically equivalent to the second method but as fast as the first method. It is worth mentioning that the application of each of the methods described above requires to exclude speech pauses because they would produce unrealistically slow rates.

3. Perceptive Local Speech Rate (PLSR)

There is no homogeneous opinion of what speech rate actually is. Undoubtedly, a high speech rate is characterized by above-average syllable rates and phone rates, but previous research has shown little correlation between local syllable rate and local phone rate ($r \approx 0.6$, see fig. 1) indicating that the information contents of both differ (Pfitzinger, 1996). The existence of words such as *banana*, showing twice as many phones compared to the syllables, in contrast to the word *stretchmarks*, having five times more phones than syllables, suggests the hypothesis that syllable rate as well as phone rate are involved in speech rate perception.

In earlier studies (Pfitzinger (1998; 1999; 2001)) we conducted a series of four perception experiments to obtain a perceptual reference for local speech rate. On the basis of these results we developed several acoustic models to predict the perceptual judgements. Our results have shown that perceptive local speech rate (PLSR) is predictable by means of an acoustic model with fair accuracy ($r = 0.91$). The simplest model we proposed consisted of a linear combination of local syllable rate and local phone rate.

Another result was that the linear correlation coefficient $r \approx 0.79$ of the syllable rate with PLSR was not significantly different from the linear correlation coefficient of the phone rate with PLSR. Therefore the term *speech rate* should not be used if *syllable rate* or *phone rate* is meant. The main outcome was that our PLSR prediction models seem to be accurate enough to work with in spoken language research.

3.1. Evaluation of speech rate estimation

The Evaluation of the accuracy and generalization properties of our PLSR prediction models requires a test corpus containing spoken language data which was unseen during the development process of our models. Since the development corpus was taken from PhonDatII (see section 4.), we conducted a new perception experiment with 100 stimuli taken from the VerbMobil spoken language resource (Wahlster, 2000). Hence speakers, speaking style, recording equipment, and vocabulary differed.

30 subjects participated in this listening test. The result of this evaluation was that the accuracy on the test corpus was nearly the same as on the training data. This allows us to conclude that our models have good generalization qualities.

3.2. Normalization

How can a local speech rate curve be used to normalize speech rate variation? Stretches of speech with fast local speech rate have to be slowed down to the average speech rate and vice versa. The inverse of the local speech rate curve exactly fulfills this condition, and is the required control input to a conventional time-stretching algorithm enabling it to exactly even out deviations from the average speech rate.

After such a procedure no single stretch of the resulting speech signal should show a significant deviation from the average speech rate. Since at present our local speech rate estimation method has a mean deviation of ca. 10% there is a small residual speech rate variation in the normalized speech signals. But it is reduced to less than 10% of the original speech rate variation.

3.3. Evaluation of speech rate normalization

To evaluate this procedure we conducted a preliminary perception experiment in which 10 subjects were instructed to sort 100 speech stimuli according to the perceived speech rate. All stimuli, each having a duration of 625 ms, were taken from speech-rate-normalized utterances which originally had strong speech rate variations. In contrast to earlier perception experiments based on the original speech signals (Pfitzinger (1998; 1999; 2001)) the subjects are not able to sort the stimuli adequately since the perception results did not exceed chance level. These preliminary results evidence our speech rate estimation method as well as our normalization procedure.

4. The PhonDatII Spoken Language Resource

This investigation is based on the PhonDatII[1] spoken language resource, which, after 10 years of providing the base for various scientific investigations, is worth of being thoroughly described in the following sections. The initial aim of setting up the PhonDatII spoken language resource was the development and evaluation of automatic recognition of continuous speech.

4.1. Material

The speech material consists of 200 sentences in the domain of train inquiry. 100 sentences, the so-called "Erlangen"-sentences, were based on transliteration of real train inquiry dialogues. The other 100 sentences, the so-called "Siemens"-sentences, were theoretically worked out.

In (Thon, 1992) all PhonDatII sentences as well as their canonical transcriptions are shown.

4.2. Subjects

5 subjects from Bonn (2 female, 3 male), 5 subjects from Kiel (2 female, 3 male), and 6 subjects from Munich (3 female, 3 male), giving a total of 16 subjects (10 male, 6 female), participated as speakers in the recordings each reading aloud all 200 sentences giving a total of 3200 sentences. The age of 11 subjects was between 20 and 30, the age of the other 5 subjects was between 30 and 56. The subjects were native German speakers and their dialect was High German slightly biased by the just-mentioned regions.

4.3. Recordings

The speech data was recorded in the spring 1992 at three departments of phonetics in Germany (Bonn, Kiel, and Munich). The subjects were seated in an anechoic chamber (Munich) or in sound-treated rooms (Bonn, Kiel) and were recorded using a *Neumann U-87* professional condenser microphone with cardioid polar pattern, a *John Hardy M-1* microphone pre-amplifier, and a *Sony PCM-2500* DAT recorder. Finally, speech data was resampled at 16 kHz with 16 bit amplitude resolution.

4.4. Speaking style

The speaking style of most of the subjects was read aloud speech, while some speakers tried to speak as if they were in a familiar environment talking about a subject they chose for themselves. Consequently, the corpus includes a range of speaking styles from fluently read aloud speech to semi-spontaneous speech.

4.5. Phone labels and inventory

64 of the 200 sentences were selected for manual segmentation of phones and syllable nuclei. Tab. 3 shows a list of modified SAM-PA symbols being the base for labelling the PhonDatII spoken language resource (Kohler, 1992b). We avoid using the term 'phonemes' even though the modified SAM-PA symbol inventory is, regarding its structure and size, not very different from the IPA symbol inventory suggested for German (International Phonetic Association, 1999, p. 86f).

The major difference consists in counting combinations of vowels followed by /6/ as a single 'segment' in order to avoid uncertain segment boundaries (van Dommelen, 1992, p. 201). As you can see in tab. 3 this results in some possible 'segments' to have not appeared in the PhonDatII spoken language resource ([2:6] as in *Gehör* (ear), [96] as in *Wörter* (words)) and others to have counts below 50 ([Y6], [I6], [a6], [E6], and [E:6]).

In this paper we use the term 'phone' though it is dangerous to call a diphthong like [e:6], as in the German word *sehr* (very), a phone because phonologically it consists of the vowel /e:/ and the consonant /r/ which would be vocalized to /6/ in most German dialects. We also say 'phone types' when we would like to emphasize the contrast to 'phone tokens'. It should be clear that each spoken realization of a phone type is a phone token.

Even though the phone labels include information on insertions, substitutions, and deletions of phones, we had no confidence in the substitutions and therefore we referred to the canonical forms if substitutions occurred. Additionally, we corrected all insertions to make certain that they no longer introduce not allowed phone symbols.

Following these considerations, 41 conventional phone types and 13 vowel+vocalized-/r/ phone types gave a total of 54 phone types. The only German phoneme, which did not occur due to the small number of word types, is the voiced post-alveolar fricative [Z] as in the German word *Garage* (garage). Finally, we would like to present some expressive counts:

	types	tokens
phones	54	39612
words	191	9424
sentences	64	1024
syllables	(not counted)	15083

Table 1: Counts of types and tokens in the manually segmented part of the PhonDatII spoken language resource.

5. Evaluation

Our investigation of phone duration variation is based on the PhonDatII spoken language resource described above. The first step was to estimate local phone rate and local syllable rate. Then, on the basis of these data, we evaluated and revised the segmentation marks to protect statistical analysis against destructive segmentation errors.

5.1. Local phone rate and local syllable nucleus rate

Applying the local rate estimation procedure introduced in section 2. to manually labelled phones and syllables every 100 ms step through the entire PhonDatII corpus leads to the data shown in fig. 1. Each point in the scatter plot represents the phone rate (ordinate) and the syllable rate (abscissa) of a 625 ms frame.

5.2. Evaluation of the segmentation data

Very slow as well as very fast local rate values were potential candidates for segmentation errors. Therefore we rechecked the segmentation of sentences responsible for extreme rate values. After correcting segmentations which were wrong in terms of the original labelling instructions (van Dommelen, 1992) we recalculated the rates and inspected again extreme rate values. Conducting this correction procedure cyclically we approached values which at the end were repeatedly confirmed. They are shown in the following table:

[1] *PhonDatII* was funded by the then German Federal Ministry of Research and Technology (BMFT) from 1/1/1991 until 12/31/1992 under contract DLR01IV103.

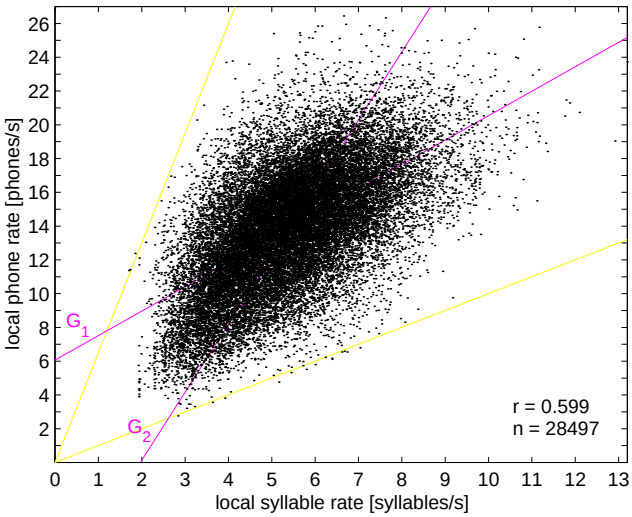

Figure 1: Scatter plot of the local phone rate based on manually segmented phone boundaries versus the local syllable rate based on manually segmented syllable nuclei. Compare to (Pfitzinger, 1998, p. 1087).

	mean	std.dev.	min.	max.
phon duration [ms]	71.9	36.9	11.6	435.9
syllable nucleus distance [ms]	184.5	98.4	28.3	591.8
phone rate [phones/s]	13.90	3.50	2.76	26.45
syllable rate [syllables/s]	5.42	1.45	1.70	13.20

Table 2: Measurements concerning the PhonDatII spoken language resource.

We expect that further evaluation and correction would not change these numbers by more than 0.1% since we evaluated the corpus several times until none of the values varied by more than 0.05%.

While in (Pfitzinger, 1998) roughly 20% of the data was rejected automatically because of glitches in the manual segmentation, we could now incorporate the entire speech data as most of the errors were recovered since 1998. The final speech data base consists of 16 speakers $\times$ 64 sentences $\times$ 2.78 seconds mean duration $=$ 2850 seconds of actual speech. Together with ca. 1000 seconds of silence the total corpus size is 64 minutes.

6. Experiment

Applying the local speech rate normalization procedure introduced in section 3.2. to the entire PhonDatII corpus leads to the data shown in tab. 3. After that we estimated local phone rate and local syllable rate on the speech rate normalized corpus. The result is shown in fig. 2.

6.1. Phone durations and standard deviations

Tab. 3 shows mean durations and standard deviations of all phone types of the PhonDatII spoken language resource. In most of the cases the standard deviation of the phones taken from the speech rate normalized corpus are significantly smaller than the original standard deviations (examined by means of F-test). In three cases (OY, u:6, y:) the standard deviation was reduced by more that half. On the other hand it was never increased significantly by our normalization procedure. In total standard deviation was reduced by approx. 22%.

		mean duration [ms]			standard deviation [ms]		
phone	N	orig.	norm.		orig.	norm.	
2:	80	92.5	99.3	*	20.9	20.7	n.s.
6	687	75.4	81.8	***	29.7	21.7	***
9	288	87.0	85.4	n.s.	29.5	18.1	***
@	1195	50.8	57.1	***	25.3	21.0	***
C	1245	65.0	68.3	**	28.9	23.0	***
E	385	67.0	75.2	***	20.2	20.3	n.s.
E6	48	121.2	134.0	*	27.7	28.1	n.s.
E:	112	93.3	94.7	n.s.	20.9	16.4	**
E:6	48	104.9	96.2	n.s.	50.8	27.2	***
I	1732	51.1	57.8	***	17.1	17.0	n.s.
I6	16	79.9	81.8	n.s.	15.7	11.2	(*)
N	423	65.0	65.8	n.s.	28.3	21.4	***
O	572	64.5	67.6	(*)	27.8	26.5	n.s.
O6	256	144.1	131.0	***	36.9	25.1	***
OY	272	136.4	124.4	***	42.1	20.6	***
Q	1032	55.4	58.2	**	23.2	20.8	***
S	256	86.7	80.7	**	22.8	18.2	***
U	904	64.3	67.3	*	27.2	23.1	***
U6	496	114.1	97.1	***	41.3	24.6	***
Y	162	66.0	66.4	n.s.	19.0	14.1	***
Y6	16	120.7	118.4	n.s.	19.1	15.8	n.s.
a	1456	74.3	78.1	***	23.3	20.1	***
a6	32	131.6	113.9	*	32.7	20.8	**
a:	1727	110.0	99.6	***	57.9	37.9	***
a:6	96	107.0	102.0	n.s.	34.3	26.0	**
aI	703	117.6	115.6	n.s.	44.1	25.8	***
aU	224	129.3	124.5	n.s.	42.5	28.4	***
b	1241	57.9	61.2	**	25.7	26.8	(*)
d	849	44.7	49.5	***	23.2	23.0	n.s.
e:	606	88.9	89.8	n.s.	36.4	24.6	***
e:6	128	86.3	89.5	n.s.	32.4	20.9	***
f	1599	92.0	88.0	***	24.9	18.5	***
g	878	48.7	49.1	n.s.	25.8	24.7	(*)
h	541	47.7	52.1	***	19.7	20.7	n.s.
i:	270	68.6	76.1	**	31.8	28.3	*
i:6	50	83.6	90.9	(*)	20.7	17.4	n.s.
j	144	71.8	67.7	n.s.	25.8	21.7	*
k	1661	73.3	68.3	***	36.0	27.5	***
l	724	57.5	56.4	n.s.	24.0	19.0	***
m	1995	69.6	71.3	*	26.5	21.2	***
n	5341	68.1	68.4	n.s.	31.0	23.7	***
o:	383	94.0	93.3	n.s.	35.4	29.7	***
o:6	128	93.2	86.5	*	28.8	15.3	***
p	192	78.6	75.7	n.s.	29.1	19.5	***
r	662	51.9	51.2	n.s.	18.2	18.7	n.s.
s	1798	82.7	82.2	n.s.	27.9	22.2	***
t	3159	59.8	59.1	n.s.	30.6	22.2	***
u:	400	66.9	67.8	n.s.	26.8	16.3	***
u:6	160	161.9	132.4	***	68.1	31.0	***
v	558	43.3	46.6	**	18.4	17.8	n.s.
x	1048	63.4	63.2	n.s.	27.1	18.9	***
y:	192	93.0	88.1	n.s.	56.3	25.9	***
y:6	64	57.7	63.7	(*)	18.3	17.3	n.s.
z	378	71.2	69.6	n.s.	23.6	20.5	**
total	39612	71.9	71.9	n.s.	36.9	28.7	***

Table 3: Comparing original vs. speech rate normalized mean phone durations and standard deviations. F-test for homogeneity of two samples and t-test for inhomogeneous variances with Welch-correction of the degrees of freedom.

6.2. Normalization does not change ratios

Fig. 2 shows a scatter plot of local phone rate values vs. local syllable rate values obtained from the speech rate normalized PhonDatII spoken language resource. Compared with fig. 1 it is remarkable that the normalization process

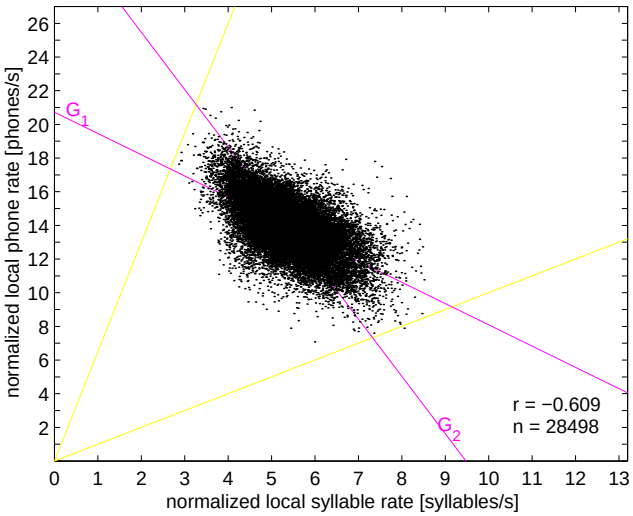

Figure 2: Scatter plot of the local phone rate based on manually segmented phone boundaries versus the local syllable rate based on manually segmented syllable nuclei (compare with fig 1).

causes the correlation coefficient r to only change its sign. This could be due to the fact that speech rate normalization is not able to modify syllable durations and phone durations independently of each other. The *local* relationship between syllable rate and phone rate remains nearly unchanged. E.g. a word like *stretchmarks*, having five times more phones than syllables, remains to have this ratio independent of the direction and amount of speech rate change.

6.3. Normalization causes duration changes

In fig. 3 a histogram of phone duration changes caused by our speech rate normalization procedure is shown. We plotted a Gaussian curve into the histogram to make clear, where deviations from the normal distribution are. A logarithmic scale was chosen because it makes clearer that the obtained distribution skews to the left.

There are more phones than expected from standard distribution, whose duration was reduced by more than half (see the abscissa section below -0.6 in fig. 3). Therefore they must have been from stretches of speech having half of the average speech rate. Examination of these stretches revealed that they mostly appear in utterance-final position. Obviously, our normalization procedure compensates for the pre-final speech rate *ritardando*.

The phones belonging to the abscissa section above 0.5 in fig. 3 have a smaller than expected number. They mostly appear in utterance-initial position and represent a speech rate *accelerando*.

6.4. Discussion

It can be assumed that the average speech rate in the PhonDatII corpus caused the peak in the histogram at ca. 0.15 (fig. 3). Speakers deviate from the average speech rate in both directions according to the communicative relevance of the particular speech phrase. Stretches of speech containing mainly function words mostly show an above-average local speech rate while words in sentence focus are produce with a slower speech rate confirming the results of Kohler (1992a).

Excluding pre-final lengthening and utterance-initial *accelerando* from the measurements the communicatively

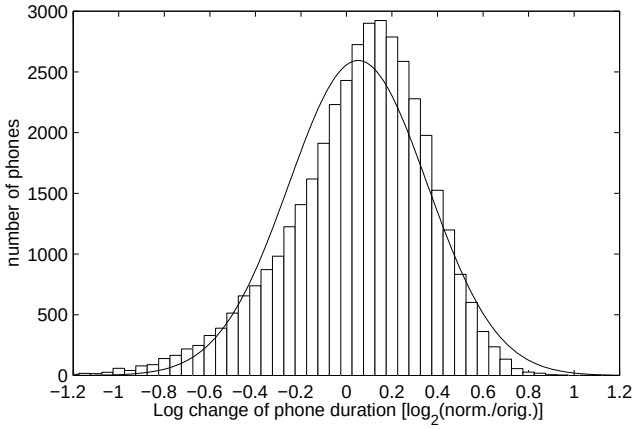

Figure 3: Histogram of logarithmic changes of phone durations through speech rate normalization.

motivated speech rate deviations lie in a factor range of 0.7 to 1.4 from the average speech rate. The linguistic meaningfulness of our speech rate prosody is obvious.

7. Statistical analysis of duration variation

The upper histogram in fig. 4 shows clearly that linear phone durations substantially skew to the right, and therefore statistical analysis by means of general linear model is not appropriate. With regard to the TIMIT speech database Wang (1997, p. 131) concludes that actual phone duration distribution generally has an asymmetrical shape.

Logarithmic phone durations nearly have a Gaussian distribution as shown in the lower histogram in fig. 4. So we apply general linear model statistical analysis to logarithmic duration values in the following sections, although other researchers noticed the skew and decided that it should not have an important effect on the results (Campbell & Isard, 1991, p. 40).

7.1. Speaker effect and phone type effect

We examined the influence of speaker and phone type on original phone durations by tabulating the 10 vowels with the largest frequency. This procedure leads to 10 factor levels for ‘phone-type’. The factor ‘speaker’ had 16 levels giving a total of 160 cells each comprising 31 randomly

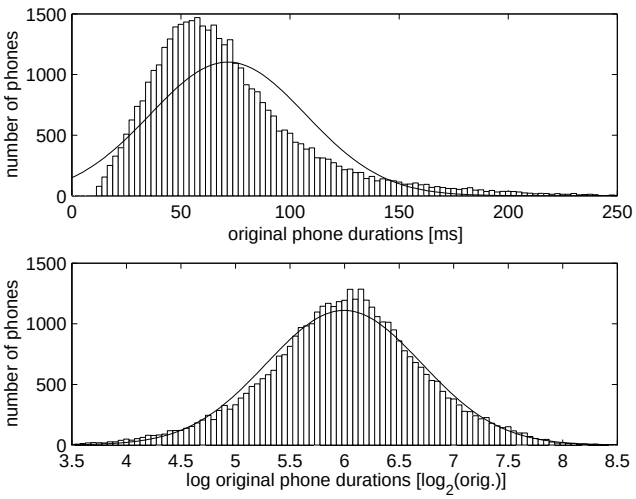

Figure 4: Histograms of original phone durations and of logarithmic original phone durations.

Effect	F	p	%
speaker	F(15,4050)=25.68	p<0.001	4.6
phone-type	F(9,4050)=335.06	p<0.001	35.9
repetitions	F(30,4050)=1.23	p=0.180	0.4
speaker × phone-type	F(135,4050)=1.47	p<0.001	2.4
speaker × repetitions	F(450,4050)=0.96	p=0.704	5.1
phone-type × repetitions	F(270,4050)=1.07	p=0.214	3.4

Table 4: Results of a three-way linear ANOVA on logarithmic original vowel durations.

Effect	F	p	%
speaker	F(15,4050)=13.38	p<0.001	2.5
phone-type	F(9,4050)=304.20	p<0.001	34.4
repetitions	F(30,4050)=0.81	p=0.761	0.3
speaker × phone-type	F(135,4050)=1.61	p<0.001	2.7
speaker × repetitions	F(450,4050)=1.05	p=0.235	5.9
phone-type × repetitions	F(270,4050)=0.95	p=0.705	3.2

Table 5: Results of a three-way linear ANOVA on logarithmic speech rate normalized vowel durations.

selected repetitions. We submitted these data to three-way ANOVA (see tab. 4). Then we repeated this procedure for 10 consonants as well as for the vowels and consonants of the speech rate normalized corpus giving the results presented in tab. 5, 6, and 7.

As would be expected, all four ANOVAs showed that both 'speaker' ($p < 0.001$) and 'phone-type' ($p < 0.001$) have significant influence on phone duration. But there is one interaction, which also is always significant: the interaction between these two factors. This means i) that different speakers realize different intrinsic phone durations, ii) that different phone types require different intrinsic durations, and iii) that different speakers use different strategies for the assignment of prototypical durations to phone types.

A striking finding is that speech rate normalization reduces the amount of explained speaker variation from 4.6% to 2.5% for vowels and from 4.2% to 1.2% for consonants. On the other side it increases the explained variation of the speaker × phone-type interaction slightly from 2.4% to 2.7% for vowels and from 3.7% to 3.8% for consonants. We would like to emphasize that these results suggest that speaker characteristics is partly hidden in the individual intrinsic durations, which are not evened out by speech rate normalization.

Effect	F	p	%
speaker	F(15,4050)=18.45	p<0.001	4.2
phone-type	F(9,4050)=142.92	p<0.001	19.7
repetitions	F(30,4050)=0.67	p=0.913	0.3
speaker × phone-type	F(135,4050)=1.78	p<0.001	3.7
speaker × repetitions	F(450,4050)=0.88	p=0.966	6.1
phone-type × repetitions	F(270,4050)=0.91	p=0.847	3.8

Table 6: Results of a three-way linear ANOVA on logarithmic original consonant durations.

Effect	F	p	%
speaker	F(15,4050)=5.03	p<0.001	1.2
phone-type	F(9,4050)=121.77	p<0.001	17.9
repetitions	F(30,4050)=0.55	p=0.978	0.3
speaker × phone-type	F(135,4050)=1.73	p<0.001	3.8
speaker × repetitions	F(450,4050)=0.92	p=0.885	6.7
phone-type × repetitions	F(270,4050)=0.91	p=0.850	4.0

Table 7: Results of a three-way linear ANOVA on logarithmic speech rate normalized consonant durations.

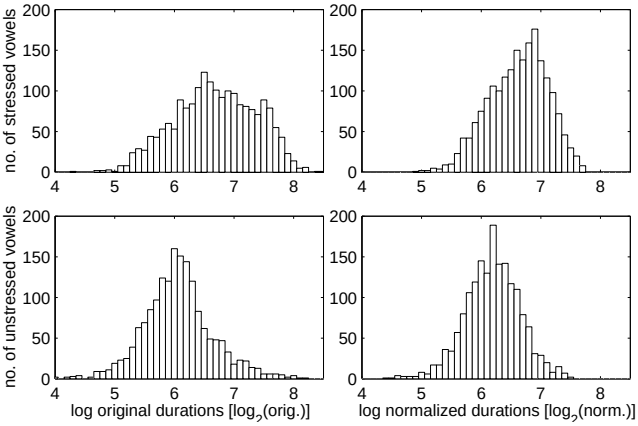

Figure 5: Histograms of original and normalized durations of stressed and unstressed tense vowels.

7.2. Stress

To investigate the influence of stress on vowel duration the total number of tense vowels (the phone types i:, y:, u:, e:, 2:, o:, E:, and a:) was split into two groups according to the factor 'stress'. The first group had a primary stress, the second was unstressed. There were also tense vowels provided with a secondary stress but we omitted them from this examination because of their small number (80). In fig. 5 the duration distribution of these two groups is shown. It is obvious that primary stress leads to longer durations, and statistical analysis confirms these results ($\hat{F} = 1.390 \geq F(0.001; 1965, 1723) = 1.156, \star\star\star$ and $\hat{t} = 29.701 \geq t(3684; 0.001) = 3.293, \star\star\star$).

In addition, fig. 5 shows the effect of normalization on the shape of the duration distributions. The distribution of the stressed vowels became more similar to normal distribution, and the bi-modal character nearly disappeared. Nevertheless the variances of stressed and unstressed vowels remained to be inhomogeneous ($\hat{F} = 1.148 \geq F(0.002; 1965, 1723) = 1.144, \star\star$).

Tab. 8 reveals that 34.6% of the observed variance could be explained by the statistically significant factor 'stress'. The factor 'speaker' has a significant influence on vowel duration, which means that different speakers realize different target vowel durations. The absence of a significant interaction between 'speaker' and any other factor is also an important result. This means that all speakers use the same 'duration rules' e.g. they lengthen the durations in presence of stress and they produce different durations for different tense vowel types. The latter is the reason why the factor 'vowel-type' is also significant.

Effect	F	p	%
stress	F(1,3)=53.25	p=0.005	34.6
speaker	F(15,45)=3.46	p<0.001	3.9
vowel-type	F(3,45)=22.84	p<0.001	5.0
stress × speaker	F(15,45)=0.91	p=0.563	1.1
stress × vowel-type	F(3,45)=8.91	p<0.001	1.9
speaker × vowel-type	F(45,45)=1.02	p=0.435	3.4

Table 8: Results of a three-way linear ANOVA on logarithmic tense vowel durations speech rate normalized with random factor 'vowel-type'.

7.3. Pre-final lengthening

Earlier estimates of mean phone durations suffered fundamentally from pre-final lengthening because it was not obvious how many phones were lengthened at the end of an utterance due to the pre-final *ritardando* (Cooper & Danly, 1981). These phones should have been omitted to avoid distortion of means. Our speech rate normalization procedure solves this problem and allows to include phones even if their durations were originally utterance-final and therefore modified by pre-final lengthening.

8. General Discussion

A table of phones together with their mean durations and standard deviations (e.g. tab. 3) is expressively only if it is reliable. In this paper the reliability of concrete duration values depends on how well the underlying spoken language resource represents spoken German. 10 years after the record of PhonDatII it is time to summarize and to look forward.

8.1. The future of the PhonDatII spoken language resource

Apparently, producers and users of spoken language resources have to accept the question as to whether the term 'large' is adequate. Particularly the PhonDatII spoken language resource comprising only 191 word types seems to be anything but 'large' from a superficial view. Undoubtedly, 39612 manually labelled phone tokens and 15083 syllable tokens can only be described as 'large'. And if they were not used for general investigations on word-level[2] or phrase-level they remain to be very valuable, especially for phonetic research even in the year 2002 and for years to come.

How can we refine the PhonDatII data base for the future? The expenditure of manual labelling of the remaining 136 sentences is considered to be excessive regarding the fact that the number of word types/tokens would increase to only 367/34192 even though the number of phone tokens would increase to approx. 143000 and the number of syllable tokens to approx. 54000. Perhaps the manual labelling of a smaller subset of sentences, which was condensed by means of greedy-methods, could effectively raise the value of the data base.

The number of speakers is comparatively small. Additional recordings of new speakers are certainly possible under the same circumstances and using the same recording equipment, which is, even 10 years later, nearly state of the art recording technology. It is even possible to increase the sample rate up to 44.1 kHz since the entire corpus was recorded to DAT tapes. But this would require a lot of work.

Taking into consideration this discourse the attempt to refine PhonDatII seems questionable. And another question arises as to whether it is necessary or even helpful to begin these attempts. The comparability of 'historical' and future results on the PhonDatII spoken language resource could possibly suffer from the refinement. However, the results of this study, which is based on a corrected and extended version of the PhonDatII, are in correspondence with earlier research. Consequently, the actual size of this spoken language resource seems to be sufficient for generalization on the segmental level. No single spoken language resource can meet every researcher's needs.

[2]There are exceptions: PhonDatII is also suited for analysing inter- and intra-speaker variability of e.g. the words *Zug* (*train*) vs. *Zugverbindung* (*train connection*) since there are 16 speakers each producing 11 and 10 repetitions, respectively.

9. Conclusions

Two steps were done making phone duration measurements accessible to statistical analysis: i) we normalized the speech rate in the entire underlying spoken language resource and ii) we used logarithmic durations because they have a probability density function very like that of a normal distribution, whereas distributions of linear durations are generally skewed (see fig. 4). These two procedures permitted us to investigate statistically the influence of speaker and phone type on phone duration and of stress on tense vowel duration. The result was that these three factors have highly significant influence.

It remains to repeat this investigation on other spoken language resources, e.g. VerbMobil (Wahlster, 2000). Another important point in our future research will be to construct a model for speaker effects on segment durations.

10. References

Bartkova, K. and Sorin, C. 1987. A model of segmental duration for speech synthesis in French. *Speech Communication*, 6(3):245–260.

Campbell, W. N. and Isard, S. D. 1991. Segment durations in a syllable frame. *J. of Phonetics*, 19:37–47.

Campbell, W. N. 1992. Syllable-based segmental duration. In Bailly, G., Benoît, C., and Sawallis, T. R., eds., *Talking Machines: Theories, Models, and Designs*, pp. 211–224. North-Holland, Amsterdam.

Carlson, R. and Granström, B. 1986. A search for durational rules in real-speech data base. *Phonetica*, 43:140–154.

Chen, M. 1970. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica*, 22:129–159.

Cooper, W. E. and Danly, M. 1981. Segmental and temporal aspects of utterance-final lengthening. *Phonetica*, 38:106–115.

Heid, S. J. G. G. 1998. Phonetische Variation: Untersuchungen anhand des PhonDat2-Korpus. Forschungsberichte (FIPKM) 36, Institut für Phonetik und Sprachliche Kommunikation der Universität München, pp. 193–368.

House, A. S. 1961. On vowel duration in English. *J. of the Acoustical Society of America*, 33(9):1174–1178.

International Phonetic Association, ed. 1999. *Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge.

IPDS. 1994. *The Kiel Corpus of Read Speech. CDROM no. 1*. Institut für Phonetik und digitale Sprachverarbeitung der Universität Kiel.

Klatt, D. H. 1973. Interaction between two factors that influence vowel duration. *J. of the Acoustical Society of America*, 54:1102–1104.

Klatt, D. H. 1975. Vowel lengthening is syntactically determined in a connected discourse. *J. of Phonetics*, 3:129–140.

Klatt, D. H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *J. of the Acoustical Society of America*, 59:1209–1221.

Klatt, D. H. 1987. Review of text-to-speech conversion for English. *J. of the Acoustical Society of America*, 82(3):737–793.

Kohler, K. J. 1983. Stress-timing and speech rate in German. A production model. Arbeitsberichte (AIPUK) 20, Institut für Phonetik und digitale Sprachverarbeitung der Universität Kiel, pp. 5–53.

Kohler, K. J. 1986a. Invariance and variability in speech timing: from utterance to segment in German. In Perkell, J. S. and Klatt, D. H., eds., *Invariance and variability in speech processes*, chapter 13, pp. 268–289. Lawrence Erlbaum Associates, Hillsdale.

Kohler, K. J. 1986b. Parameters of speech rate perception in German words and sentences: Duration, F_0 movement, and F_0 level. *Language & Speech*, 29:115–139.

Kohler, K. J. 1990. Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In Hardcastle, W. J. and Marchal, A., eds., *Speech Production and Speech Modelling*, vol. 55 of *NATO ASI Series D: Behavioural and Social Sciences*, pp. 69–92. Kluwer Academic Publishers, Dordrecht, Boston, London.

Kohler, K. J. 1992a. Dauerstrukturen in der Lesesprache: Erste Untersuchungen am PhonDat-Korpus. Arbeitsberichte (AIPUK) 26, Institut für Phonetik und digitale Sprachverarbeitung der Universität Kiel, pp. 225–252.

Kohler, K. J. 1992b. Sprachverarbeitung im Kieler PhonDat-Projekt: Phonetische Grundlagen für ASL-Anwendungen. Arbeitsberichte (AIPUK) 26, Institut für Phonetik und digitale Sprachverarbeitung der Universität Kiel, pp. 81–95.

Kohler, K. J. 1997. Parametric control of prosodic variables by symbolic input in TTS synthesis. In van Santen, J. P. H., Sproat, R. W., Olive, J. P., and Hirschberg, J., eds., *Progress in Speech Synthesis*, chapter 37, pp. 459–475. Springer-Verlag, New York, Berlin, Heidelberg.

Lindblom, B. E. F. 1990. Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle, W. J. and Marchal, A., eds., *Speech production and speech modelling*, no. 55 in Nato ASI series D: Behavioural and social sciences, pp. 403–439. Kluwer Academic Publishers, Dordrecht, Boston, London.

Luce, P. A. and Charles-Luce, J. 1985. Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *J. of the Acoustical Society of America*, 78(6):1949–1957.

Menzerath, P. and de Oleza S. J., J. M. 1928. *Spanische Lautdauer. Eine experimentelle Untersuchung*. Walter de Gruyter, Berlin, Leipzig.

Möbius, B. and van Santen, J. P. H. 1996. Modeling segmental duration in German text-to-speech synthesis. In *Proc. of ICSLP '96*, vol. 4, pp. 2395–2398, Philadelphia.

Ohno, S., Fujisaki, H., and Taguchi, H. 1998. Analysis of effects of lexical accent, syntax, and global speech rate upon the local speech rate. In *Proc. of ICSLP '98*, vol. 3, pp. 655–658, Sydney; Australia.

Pfitzinger, H. R. 1996. Two approaches to speech rate estimation. In *Proc. of the sixth Australian Int. Conf. on Speech Science and Technology (SST '96)*, pp. 421–426, Adelaide.

Pfitzinger, H. R. 1998. Local speech rate as a combination of syllable and phone rate. In *Proc. of ICSLP '98*, vol. 3, pp. 1087–1090, Sydney; Australia.

Pfitzinger, H. R. 1999. Local speech rate perception in German speech. In *Proc. of the XIVth Int. Congress of Phonetic Sciences*, vol. 2, pp. 893–896, San Francisco.

Pfitzinger, H. R. 2001. Phonetische Analyse der Sprechgeschwindigkeit. Forschungsberichte (FIPKM) 37, Institut für Phonetik und Sprachliche Kommunikation der Universität München, pp. 3–107.

Pols, L. C. W. 1999. Flexible, robust, and efficient human speech processing versus present-day speech technology. In *Proc. of the XIVth Int. Congress of Phonetic Sciences*, vol. 1, pp. 9–16, San Francisco.

Riley, M. D. 1992. Tree-based modelling of segmental durations. In Bailly, G., Benoît, C., and Sawallis, T. R., eds., *Talking Machines: Theories, Models, and Designs*, pp. 265–273. North-Holland, Amsterdam.

Stöber, K. and Hess, W. 1998. Additional use of phoneme duration hypotheses in automatic speech segmentation. In *Proc. of ICSLP '98*, vol. 4, pp. 1595–1598, Sydney; Australia.

Thon, W. 1992. Struktur eines Datenverarbeitungssystems für das Kieler PhonDat-Projekt: von der Aufnahme ASL-PhonDat 92 zur Datenanalyse. Arbeitsberichte (AIPUK) 26, Institut für Phonetik und digitale Sprachverarbeitung der Universität Kiel, pp. 111–173.

Tillmann, H. G., Hadersbeck, M., Piroth, H.-G., and Eisen, B. 1990. Development and experimental use of PHONWORK: A new phonetic workbench. In *Proc. of ICSLP '90*, vol. 2, pp. 1009–1012, Kobe; Japan.

van Dommelen, W. A. 1992. Segmentieren und Etikettieren im Kieler PhonDat-Projekt. Arbeitsberichte (AIPUK) 26, Institut für Phonetik und digitale Sprachverarbeitung der Universität Kiel, pp. 197–223.

van Santen, J. P. H. and Olive, J. P. 1990. The analysis of contextual effects on segment duration. *Computer Speech and Language*, 4(4):359–390.

van Santen, J. P. H. 1992a. Contextual effects on vowel duration. *Speech Communication*, 11(6):513–546.

van Santen, J. P. H. 1992b. Deriving text-to-speech durations from natural speech. In Bailly, G., Benoît, C., and Sawallis, T. R., eds., *Talking Machines: Theories, Models, and Designs*, pp. 275–285. North-Holland, Amsterdam.

van Santen, J. P. H. 1993. Timing in text-to-speech systems. In *Proc. of EUROSPEECH '93*, vol. 2, pp. 1397–1404, Technische Universität Berlin.

van Santen, J. P. H. 1994. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8(2):95–128.

van Santen, J. P. H. 1997. Segmental duration and speech timing. In Sagisaka, Y., Campbell, W. N., and Higuchi, N., eds., *Computing Prosody. Computational models for processing spontaneous speech*, chapter 15, pp. 225–249. Springer-Verlag, New York.

van Santen, J. P. H. 1998. Timing. In Sproat, R. W., ed., *Multilingual text-to-speech synthesis: The Bell Labs approach*. Kluwer Academic Publishers, Dordrecht.

Wahlster, W., ed. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer Verlag, Berlin, Heidelberg, New York.

Wang, X. 1997. *Incorporating knowledge on segmental duration in HMM-based continuous speech recognition*, vol. 29 of *Studies in Language and Language Use*. IFOTT, Amsterdam.

Weitkus, K. 1931. *Experimentelle Untersuchung der Laut- und Silbendauer im deutschen Satz*. Ph.D. thesis, Universität Bonn.