INTRINSIC PHONE DURATIONS ARE SPEAKER-SPECIFIC

Hartmut R. Pfizinger

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

ABSTRACT

This study examines the speaker’s influence on mean phone durations. As long as speech rate variation is present, the result of such a study would be trivial because every speaker has a particular speech rate that naturally modifies phone durations. Therefore, in order to eliminate its influence on phone duration, we developed a normalization procedure which even out the local variability of speech rate, and then applied it to a large database of spoken German. As would be expected, general linear model statistical analysis (GLM) showed that speech rate normalization strongly reduced the variance explained by the factor ‘speaker’. Nevertheless, the variance explained by the interaction between ‘speaker’ and ‘phone type’ remained constant. Consequently, each speaker has individual intrinsic phone durations.

1. INTRODUCTION

In the past, a concept of intrinsic phone duration appeared which subsumes any speaker-dependent modification of phone durations under the category of ‘phonetic variability’, and instead concentrates on contextual effects. Most of the timing models used in speech synthesis explicitly or implicitly follow this concept regardless of the type of model to which they belong. Van Santen (1993 [1]) makes a distinction between four duration model types: i) the sequential rule systems, which adjust intrinsic phone durations to the respective context by successively applying rules, ii) the purely additive or multiplicative models, which estimate actual phone duration by adding or multiplying a number of contextual parameters to intrinsic phone durations, iii) category tree-based systems, which find a path through a number of branches to a leaf providing the actual phone duration, and finally iv) stochastic models based on ANNs or HMMs. While the first model type requires explicit rules formulated by an expert, the other three types extract their ‘knowledge’ from phone duration distributions calculated from large spoken language resources.

Much work in the field of constructing duration models was done by Klatt [2], Kohler [3], and van Santen [4]. The goal of any duration model to generate natural sounding timing “cannot in any practical sense be achieved because durational phenomena are too complex” (van Santen 1993 [1, p. 1398]). Five years later he writes: “there is a sizable amount of durational variability that cannot be predicted from text” [4, p. 139].

This study experimentally investigates whether it is possible to reduce the amount of unexplained duration variability by means of speech rate normalization. Since speech rate is a continuously varying prosodic feature we use a “momentary” or local approach to acoustic speech rate measurements, which we introduced earlier [5, 6]. 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.

The elimination of all local speech rate deviations requires a reliable local rate estimation procedure as well as a reliable normalization procedure.

2. ESTIMATION OF LOCAL RATES

In [7] we presented a mathematically sound formula to estimate local rates of speech units (phones, syllables, etc.): The distances between consecutive speech unit marks $S_i$ falling in a window $w$ of constant length ($625$ ms) are accumulated and then divided by their number. The reciprocal of the quotient is a measure for the local rate of the underlying speech unit:

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

were $w_L$ is the left and $w_R$ is the right window boundary. Since the left ($S_i$) and the right ($S_{i+1}$) segment most frequently are covered only partially by the accumulation window, they have to be added proportionately to ensure a constant window size. 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 samples of the new signal which fall in a speech unit receive the reciprocal of its duration. The second step is to convolve 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.

It is worth mentioning that the application of both methods described above requires exclusion of speech pauses because they would produce unrealistically slow rates.

3. PERCEIVED LOCAL SPEECH RATE (PLSR)

There is no homogeneous opinion as to 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 only moderate correlation between local syllable rate and local phone rate ($r=0.6$, see Fig. 1) indicating that the information contents of both differ [7]. The existence of words such as banana, showing two phones per syllable, in contrast to the word stretchmarks, with five phones per syllable, suggests that syllable rate as well as phone rate are involved in speech rate perception.

In earlier studies [5, 6] 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 perceived 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 middlingly high linear correlation coefficient $r=0.79$ 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 [8], we conducted a new perception experiment with 100 stimuli taken from the VerbMobil spoken language resource [9]. Hence speakers, speaking style, recording equipment, and vocabulary differed.

30 subjects participated in this listening test. The result of the 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 stretches of slow speech have to be accelerated. The inverse of the local speech rate curve exactly fulfills this condition, and is the required control input to a conventional time-stretching algorithm which then exactly evens out the 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 about 10% from the perception results 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 [5, 6] the subjects were not able to sort the stimuli adequately since the perception results did not exceed chance level. These preliminary results support our speech rate estimation method as well as our normalization procedure.

4. EXPERIMENT

This investigation is based on the PhonDatII spoken language source, which consists of 16 speakers each producing 200 sentences of German read speech [8]. 64 sentences were selected for manual segmentation of phones and syllables giving in total 39612 phone tokens and 15083 syllable tokens.

The first step was to estimate local phone rate and local syllable 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.

Then we applied the local speech rate normalization procedure introduced in section 3.2 to the entire PhonDatII corpus. Finally, we estimated local phone rate and local syllable rate based on the speech rate normalized corpus. The result is shown in Fig. 2.

4.1. 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 corpus. Compared with Fig. 1, it is remarkable that the normalization process 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 stretch-marks, having five times more phones than syllables, retains this ratio independent of the direction and amount of change of the speech rate.

4.2. Normalization causes duration changes

Fig. 3 shows a histogram of phone duration changes caused by our speech rate normalization procedure. We plotted a Gaussian curve into the histogram to make the deviations from the normal
distribution clear. A logarithmic scale was chosen because it makes clearer that the obtained distribution slightly skews to the right.

Phone durations reduced by more than half were more numerous than was to be expected from standard distribution (corresponds to the abscissa section below -0.6 in Fig. 3). So 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 represent the well-known utterance-initial speech rate *accelerando*.

It can be assumed that the average speech rate in the PhonDatII spoken language resource caused the peak in the histogram at approx. 0.15 in Fig. 3. Speakers deviate from this 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 the sentence focus are produced with a slower speech rate. When excluding pre-final lengthening as well as utterance-initial *accelerando* from the measurements, speech rate varies between 70% and 140% of the average speech rate. The linguistic meaningfulness of our speech rate prosody is obvious.

5. STATISTICAL ANALYSIS OF DURATION VARIATION

The upper histogram in Fig. 4 shows clearly that linear phone durations substantially skew to the left, and therefore statistical analysis by means of general linear model (GLM) is not appropriate. With regard to the TIMIT speech database Wang [10, 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. Therefore, in the following sections, we apply GLM statistical analysis (SPSS) to logarithmic duration values, although other researchers noticed the skew and decided that it should not have an important effect on the results (Campbell and Isard 1991 [11, p. 40]).

<table>
<thead>
<tr>
<th>Effect</th>
<th>F</th>
<th>p</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>speaker</td>
<td>F(15,10879)=31.80</td>
<td>&lt;0.001</td>
<td>2.81</td>
</tr>
<tr>
<td>vowel-type</td>
<td>F(16,10879)=300.94</td>
<td>&lt;0.001</td>
<td>28.40</td>
</tr>
<tr>
<td>speaker×vowel-type</td>
<td>F(240,10879)=1.42</td>
<td>&lt;0.001</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 1. Results of a two-way GLM on logarithmic original vowel durations (’%’ means explained variance).

<table>
<thead>
<tr>
<th>Effect</th>
<th>F</th>
<th>p</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>speaker</td>
<td>F(15,10879)=15.93</td>
<td>&lt;0.001</td>
<td>1.48</td>
</tr>
<tr>
<td>vowel-type</td>
<td>F(16,10879)=270.36</td>
<td>&lt;0.001</td>
<td>26.77</td>
</tr>
<tr>
<td>speaker×vowel-type</td>
<td>F(240,10879)=1.73</td>
<td>&lt;0.001</td>
<td>2.57</td>
</tr>
</tbody>
</table>

Table 2. Results of a two-way GLM on logarithmic speech rate normalized vowel durations (’%’ means explained variance).

5.1. Speaker effect and phone type effect

We examined the influence of speaker and vowel type on the original vowel durations by tabulating the durations of all realizations of the 17 monophongs in the speech database, leading to 17 factor levels for ‘vowel-type’. The factor ‘speaker’ had 16 levels giving a total of 272 cells each comprising all corresponding instances. We submitted these data to GLM (see Table 1). Then we repeated this procedure for all 21 consonants as well as for the vowels and consonants of the speech rate normalized corpus, producing the results presented in Table 2, 3, and 4.

As expected all four GLMs show that both ‘speaker’ (p<0.001) and ‘phone-type’ (p<0.001) have significant influence on phone duration. But the interaction between them is also always significant. 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 intrinsic durations to phone types.

A striking finding is that speech rate normalization reduces the variance explained by the factor ‘speaker’ from 2.81% to 1.48% for vowels and from 1.73% to 0.42% for consonants. By contrast, it slightly increases the explained variance of the speaker×phone-type interaction from 2.01% to 2.57% for vowels and from 2.86% to 2.98% for consonants. We would like to emphasize that these results suggest that speaker characteristics are partly hidden in the individual intrinsic durations, which are not evened out by speech rate normalization. Most of the variance is explained by contextual factors which we focus on in the next section.

<table>
<thead>
<tr>
<th>Effect</th>
<th>F</th>
<th>p</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>speaker</td>
<td>F(15,25388)=38.70</td>
<td>&lt;0.001</td>
<td>1.73</td>
</tr>
<tr>
<td>cons-type</td>
<td>F(20,25388)=289.87</td>
<td>&lt;0.001</td>
<td>17.29</td>
</tr>
<tr>
<td>speaker×cons-type</td>
<td>F(300,25388)=3.20</td>
<td>&lt;0.001</td>
<td>2.86</td>
</tr>
</tbody>
</table>

Table 3. Results of a two-way GLM on logarithmic original consonant durations (’%’ means explained variance).

<table>
<thead>
<tr>
<th>Effect</th>
<th>F</th>
<th>p</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>speaker</td>
<td>F(15,25388)=9.01</td>
<td>&lt;0.001</td>
<td>0.42</td>
</tr>
<tr>
<td>cons-type</td>
<td>F(20,25388)=284.37</td>
<td>&lt;0.001</td>
<td>17.60</td>
</tr>
<tr>
<td>speaker×cons-type</td>
<td>F(300,25388)=3.21</td>
<td>&lt;0.001</td>
<td>2.98</td>
</tr>
</tbody>
</table>

Table 4. Results of a two-way GLM on logarithmic speech rate normalized consonant durations (’%’ means explained variance).
5.2. Stress

To investigate the influence of stress on vowel duration the total number of tense vowels (the phone types i:, y:, u:, e:, ə:, o:, E:) 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 ($F(1,39 > F(0.001;1965, 1723) = 1.16 \times 10^3, f = 29.7 > t(3684; 0.001) = 3.29 \times 10^3$).

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 bimodal character nearly disappeared. Nevertheless the variances of stressed and unstressed vowels remained inhomogeneous ($F = 1.15 > F(0.002; 1965, 1723) = 1.14 \times 10^2$).

Table 5 reveals that 33.0% 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 durations for the same vowels. 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.

5.3. 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 *ritirando* (Cooper and Danly 1981 [12]). 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 they originally appeared in utterance-final position and therefore were modified by pre-final lengthening.

6. CONCLUSION

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. All three factors have highly significant influence. Especially, each speaker realized individual target durations.

It remains to repeat this investigation on other spoken language resources, e.g. VerbMobil [9]. Another important point in our future research will be to construct a model for speaker effects on segment durations.

7. REFERENCES