10 YEARS OF PHONDAT-II: A REASSESSMENT

Hartmut R. Pfiztinger

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

ABSTRACT

In this paper we conduct an evaluation as well as a reassessment of the PhonDatII spoken language resource. 10 years after the record of PhonDatII it is time to summarize and to look into its future. At present, the corpus comprises 39612 manually labelled phone tokens and 15083 syllable tokens of read German utterances. We describe the corpus in detail, and then we present a new method to evaluate segmentation boundaries. Finally, we ask the question as to how we can refine the PhonDatII database for the future. The mean phone duration results of this study, which are based on a corrected and extended version of the PhonDatII corpus, are in correspondence with earlier research. Consequently, the actual size of this spoken language resource seems to be sufficient for generalization of results on the segmental level.

1. INTRODUCTION

This reassessment focuses on the PhonDatII1 spoken language resource, which, after 10 years of providing the base for a great number of various scientific investigations as well as for commercial development, is worth of being thoroughly described in the following sections. The initial aim of setting up the PhonDatII corpus was the development and evaluation of automatic recognition of continuous speech.

1.1. Some Investigations based on the PhonDatII Corpus

In 1992 Kohler [1] conducted duration measurements on stressed vowels taken from one third of the PhonDatII corpus. He found their mean durations to be between 66.9 ms and 168.4 ms depending on the factors ±last syllable in word, ±tense, and ±final.

Reyelt 1993 [2] reported on the development and training of an automatic classifier for the prosodic categories phase accent and secondary accent which demonstrated to be mainly dependent on F0 rather than loudness.

Stöber&Hess 1998 [3] developed an automatic phoneme alignment method that uses a genetic algorithm to incorporate phoneme duration information. The PhonDatII corpus was used to evaluate the accuracy of the phoneme boundary placements since it provides reliable manual segmentation marks. 80% of the automatically determined placements deviated 20 ms or less from the reference positions.

Heid 1998 [4, p. 253] presented a set of phonetic and phonological investigations on segmental variation, on vowel quality variation, and on prosodic variation contours. F0, formant, and phone duration measurements were based on 10300 vowel realizations taken from the PhonDatII spoken language resource. Until now, this was to our knowledge the largest survey on German phone durations. The overall mean phone duration was 72.45 ms.

2. ACOUSTIC DATA COLLECTION

2.1. Material

The speech material of the PhonDatII corpus 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 [5] the orthography of all PhonDatII sentences as well as their canonical transcriptions are shown.

2.2. Subjects

5 subjects from Bonn (2 female, 3 male), 5 subjects from Kiel (2 female, 3 male), and 6 subjects from Munich (2 female, 4 male), giving a total of 16 subjects (6 female, 10 male), 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.

2.3. 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.

2.4. 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.

1PhonDatII was funded by the then German Federal Ministry of Research and Technology (BMFT) from 1/1/1991 until 12/31/1992 under contract DLR01IV103.
3. MANUAL SEGMENTATION AND LABELLING

3.1. Phone Labels and Inventory

64 of the 200 sentences were selected for manual segmentation of phones. The segmentation conventions were presented in detail in van Dommelen 1992 [6]. The Tables 1, 2, 3, and 4 show a list of the 57 modified SAM-PA symbols being the base for labelling the PhonDatII spoken language resource (Kohler 1992 [7]). 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 (IPA 1999 [8, 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 [6, p. 201]). Two of these possible ‘segments’ did not appear in the PhonDatII corpus: [2:6] as in Gehör (ear) and [96] as in Wörter (words). And the only phones having counts below 50 are some of the /r/-diphthongs: [Y6], [I6], [a6], [E6], and [E:6] (see Table 7).

3.2. Syllable Segmentation

Those sentences providing manual phone segmentations were used in a study on automatic syllable detection. For that purpose a new reference segmentation was needed which was produced manually by three segmenters which also re-labelled 96 sentences to test inter-segmenter consistency. They agreed in 95.8% of all placements (Pfitzinger et al. 1996 [9]).

4. RE-EVALUATION

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.

Following these considerations, 42 conventional phone types (Table 1, 2, 3) and 15 vowel-/B7 vocalized-/r/ phone types (Table 4) gave a theoretical total of 57 phone types. The only German phoneme, which did not occur in the PhonDatII corpus due to the small number of word types, is the voiced post-alveolar fricative [Z] as in the German word Garag(ge) (garage). Together with the two /r/-diphthongs [2:6] and [96] which also did not occur, the actual number of phone types was 54 (see Table 7). Finally, we would like to present some expressive counts:

<table>
<thead>
<tr>
<th>types</th>
<th>tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>phones</td>
<td>54</td>
</tr>
<tr>
<td>words</td>
<td>191</td>
</tr>
<tr>
<td>sentences</td>
<td>64</td>
</tr>
<tr>
<td>syllables</td>
<td>(not counted)</td>
</tr>
</tbody>
</table>

Table 5. Counts of types and tokens in the manually segmented part of the PhonDatII spoken language resource.

4.1. Method

Since manual evaluation and correction of phone boundaries is a very expensive task we used a new method to discover segmentation errors. The first step was to estimate local phone rate and local syllable rate. Then, on the basis of these data, we identified stretches of speech producing extreme rate values and/or extreme
rate relations (see Fig. 1). Finally, we manually revised the segmentation marks of the corresponding sentences.

4.2. Local Phone Rate and Local Syllable Rate

In 1996 [10] 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 (e.g. 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:

\[
rate_{LR} = \frac{S_{i+1} - w_L + w_R - S_i + \sum_{j=1}^{i} (S_{j+1} - S_j)}{S_{i+1} - S_i + r - l - 1}
\]

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 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.

Applying the local rate estimation procedure 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.

4.3. Evaluation of the Segmentation Data

Very slow as well as very fast local rate values were potential candidates for segmentation errors since they turned out to often arise from wrongly inserted or deleted segment boundaries. We assume that this criterion more than averagely identifies sentences segmented by less experienced segmenters. Therefore we completely rechecked the segmentation of sentences yielding extremely rate values and/or rate relations. After correcting boundaries which were wrong in terms of the original labelling instructions [6] we recalculated the rates and inspected again extreme rate values. While conducting this correction procedure cyclically we approached rate values which, at the end, were repeatedly confirmed.

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

We expect that further correction would not change these numbers significantly since i) we cyclically evaluated the corpus until none of the values varied by more than 0.1%, and ii) since remaining wrong insertions could be assumed to even out wrong deletions and therefore would neither change means nor standard deviations.

While in Pfitzinger 1998 [11] roughly 20% of the data was rejected from the rate estimation 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 database consists of 16 speakers \( \times \) 64 sentences \( \times \) 2.78 seconds of mean sentence duration = 2850 seconds of actual speech. Together with approx. 1000 seconds of silence and speech pauses the total corpus size is 64 minutes.

5. DISCUSSION

10 years after the record of PhonDatII it is time to summarize and to look into its future.

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 database 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 set of sentences, which was condensed by means of greedy-methods, could effectively raise the value of the database.

The number of speakers is comparatively small. Additional recordings of new speakers are certainly possible under the same

\(^2\)There are exceptions: PhonDatII is also suited for analyzing 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.
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 mean phone duration results of this study (see Table 7), which are based on a corrected and extended version of the PhonDatII corpus, are in correspondence with earlier research. Consequently, the actual size of this spoken language resource seems to be sufficient for generalization of results on the segmental level.

6. REFERENCES


Table 7. Mean phone durations and standard deviations of all 54 PhonDatII phone types (in ms). Right column: Graphical representation of phone durations and standard deviations.