The Vienna Prosodic Speech Corpus: Purpose, Content, and Encoding

Friedrich Neubarth, Kai Alter[+], Hannes Pirker, Elisabeth Rieder, Harald Trost
Austrian Research Institute for Artificial Intelligence, Vienna, Austria [*]
[+] Max Planck Institute of Cognitive Neuroscience, Leipzig, Germany

Abstract

This paper presents a corpus of spoken German especially designed for the investigation of prosodic properties of speech. After a short discussion of the content and set-up of the corpus, we describe in detail the additional linguistic information, introduced into the corpus by labelling and annotation. In this project, both qualitative and quantitative methods have been used for the acquisition of data. Our main concern is the development of a well-defined and transparent scheme for the structuring of this heterogeneous information. A second task is to incorporate all these data - generated by different tools with different data-formats - into a single data-base.

1 Motivation

The corpus to be presented here is designed to serve multiple purposes in the research of prosodic properties of spoken language. Most of this research is performed within the SpeeDurCont project (Speech Duration in Context-to-Speech, cf. Alter et al., 1998, Pirker et al. 1996), which is an ongoing investigation of durational variation in German speech. The project goals are twofold: From a theoretical point of view it aims at a better understanding and quantification of the many factors influencing the duration of speech items. As a practical goal durational models are derived which are applied to improve the prosodic quality of synthetic speech.
The general approach of our project is to achieve these goals by applying machine learning techniques to a suitably annotated and labelled speech corpus. The construction of such a corpus not only is a tedious "necessary evil" but raises interesting theoretical and practical questions as well.

A corpus is determined by three dimensions:

- the underlying speech material,
- annotations of various features, and
- the representation of the data.

The remainder of this paper is structured accordingly.

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

2 Corpus

The corpus comprises approx. 70.000 phones (2 hours of speech) of Standard Austrian German. It was recorded of a single speaker. Although a single speaker might have the shortcoming of possibly producing an idiosyncratic model, we feel that there is good reason for this decision. Since the results are to be used for speech synthesis, the "modelling" of a unique speaker seems justifiable. Moreover, the speaker is identical to the speaker of the synthesis inventory. Modelling cross-speaker variance is a much more complex task and can only be begun when the methods for the analysis and modelling of the data of a single speaker are entirely set up. On a par, there is good evidence that an existing model derived from a single speaker can be adapted to various target models for other speakers with much less effort (cf. Shih et al. 1998).
The data was recorded on DAT and transferred to wave-format in 16-bit encoding, 44.1 kHz sampling rate, stereo. One channel was used for the actual acoustic signal, the other for laryngographic data. The corpus comprises quite different types of read material for different purposes of analysis:

- phonetically balanced material,
- material where the information-structure is controlled for the analysis of the relation between focus structure and prominence properties, and
- connected text, which presumably is realised in the most natural way, as a control set.

The diversity of the material is motivated by the general aim of the project. We do not just aim at the development of a durational model by straightforwardly applying available statistical methods to a corpus (which should be much larger then anyway), but the the main concern of our project is to study the interdependence of determin-

able factors like prominence, syllabic and morphological structure and inherent properties of phonemes and phoneme classes.

The relevant parts of the corpus are in detail:

- 300 isolated sentences (Marburg sentences, Berlin sentences and others), parts of which are also found in the PHONDAT corpus (Kohler1994).
- 24 samples of connected text: 2 passages of the PHONDAT corpus and 22 passages of daily news from different Austrian newspapers.
- 250 question-answer pairs (q/a-corpus) The latter was recorded in order to allow for a maximum control of syntactic, phonological, and information structure. The questions do not form part of the corpus itself, they are just means to evoke the intended focus structure. The basic structure of the answer sentences is a control matrix verb and two embedded conjoined infinitival groups. Either the matrix or the first infinitival has to have an object.

  Der Freund verspricht der Direktorin zu arbeiten und das Büro zu putzen.
  the friend promises the director to work and the office to clean

  This scheme is systematically varied: Syntactic variation is induced by changing the transitivity of the matrix verb (e.g. using '...verspricht **die** Direktorin zu **entlasten**...') which radically changes the syntactic phrasing. Thus the effect of syntactic variation on accentuation and prosodic phrasing can be investigated in detail.
  The focus structure of the utterances is controlled by asking different questions, e.g. "What happened?" (*wide focus*) vs. "To whom does the friend promise to work?" (*narrow focus* on "director"). Parts of sentences in the q/a-corpus also contain focus particles (e.g., "sogar" - *even*) which offer an additional level of control to the focus structure. Last but not least the words in the sentences systematically vary with respect to the number of syllables (e.g., "Udo" vs. "die Direktorin") which allows to check for quantity effects.

3 Labelling

Qualitative as well as quantitative methods have been used in analysing the corpus data. Qualitative methods include perceptive prominence labelling, intonation labelling (GToBI) and phonological phoneme labelling. Quantitative methods deal with phonetic segmentation (S_TOOLS X) and parameterisation of intonation events (Tilt, INTOFIT). Therefore, one of the main tasks is to align qualitative and quantitative data in a way that makes them accessible for cross-correlation in order to provide straightforward explanations for different kinds of interface relations.
Matching operations between theoretical, qualitative assumptions and their quantitative counterparts are explored, leading to further insights for the development of a theoretical basis within a modular framework.

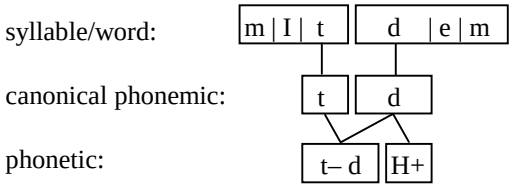
3.1 Canonical transcription

For the coding of **phonetic labels** we use a slightly modified SAMPA-notation, adapted from the PHONDAT corpus (Kohler 1994) by adding the following symbols and conventions: "L" (lateral fricative, occurs only in /tl/-sequences), "H" (aspiration), "#" (begin/end of utterance), "_" (pause), "Q" (glottal stop), as a prefix: "q" (creaky voice), suffixes: "+" (insertion), "–" (deletion), "%" (substitution) and "!n" (concatenation of n segments).
For the **canonical transcription** we assume an Austrian German standard, which deviates from the PHONDAT corpus in certain ways. For example, /r/ is realised as [R] only if it occurs before a vowel. /s/ is never voiced (with very few lexical exceptions). Canonical glottal stops are transcribed as "?", according to standard. All other symbols of the phonetic transcription are missing, whatever information they comprise will be annotated as additional tags. Most, if not all of these deviations can be formulated in terms of contextual rules, so the interconnectivity to other data-bases of German speech is guaranteed. In addition we introduced a three level distinction of lexical stress (main stress, main stress of parts of compounds and minor stress) and also indicated syllabification.
In order to have full control over the canonical transcription, especially the coding of accent and syllabification, we use a specifically designed semi-automatic transcription tool, which gives us an extendable full-form word list. Starting from transcribed text as source, it is an easy task to expand this into a full hierarchy of word, syllable and phoneme labels. The canonical phonemic layer derived from the lexicon, which encodes all hierarchical relations, can then be automatically linked to the actual phonetic segmentation by applying certain matching rules as described in the following section.

3.2 Phonetic segmentation

The segmentation and annotation of phonetic labels was done manually at the Institute for Acoustic Research of the Austrian Academy of Sciences with S_TOOLS X. This software provides a perfect tool for analysis and graphic display of the acoustic signal and also allows for

a direct annotation of any information to specific segments as a set of tags.
For identifying segment boundaries we used a broad-band spectrogram (window size 1.2 s) and a wave-form zoom (window size: 40 ms), but in many cases additional auditory control was necessary.
For the segmentation certain guidelines were obeyed. For example, boundaries involving voiced segments are set to approximately co-occur with the begin of the period. The pause (occlusion phase) and the release/aspiration of stops are independently marked only if the aspiration phase is longer than 20 ms and clearly recognisable as such. Certain transitions almost exceed the time occupied by the respective segments. In order to maintain the canonical transcription, the boundary is marked approximately in the middle of the transition by auditory control. Phonemes from the canonical transcription missing in the acoustic signal (due to, e.g., schwa-deletion, stop-assimilation - even across word boundaries) were indicated as missing in order to make the information retrievable.
Under a realistic perspective, it turns out to be impossible to hierarchically link a string of phonetic segments (which already includes a high degree of abstraction) to its (canonical) phonemic representation. Splitting a phonemic unit into two or more phonetic events does not pose any problem, but the reverse situation must be dealt with as well. A single phonetic event may correspond to two different phonemic units, either in part (e.g., transition phases) or as a whole (e.g., assimilated sequences of stops). This means that we have to start by establishing two independent layers, a phonetic one with temporal encoding and a phonological one, which comprises the canonical realisation of words and which relates to higher structures. In order to combine these two types of information these layers need to be aligned.
For that purpose, the phonetic encoding should be as close to the canonical transcription as possible, but it need not necessarily rely on a 1-to-1 or 1-to-n relation between phonemic symbols and phonetic segments.
Consider the following schematic representation:



Here we have a phonetic event (the occlusion phase of a stop) which belongs to two (assimilated) phonemes, /t/ and /d/. On the other hand, the second phoneme, /d/, must be aligned to two phonetic events (occlusion and aspiration). This is done in two steps: first the alignment is coded symbolically top-down, where segment borders which do not coincide are marked specifically (for example, the right border of the first phoneme, /t/, has no corresponding segment border on the phonetic level, the same holds for the left border of the following phoneme). In a second step, absolute time values for non-matching borders are calculated.
This is the only way to assign absolute time-values (which are primarily encoded in the phonetic layer and nowhere else) to the phonemic units and also to higher nodes like syllables or feet. The actual format of encoding and its management will be discussed in section 4.

3.3 Intonation

To achieve a better understanding of the interaction of intonation (pitch) and duration is one of the main objectives of the SpeechDurCont project. Thus, special attention is paid to pitch labelling. Intonation is manually labelled according to the G-ToBI labelling scheme (Grice et al. 96).
Intonational events are described as combinations of pitch accents (H ~ high, L ~ low) and edge tones (with additional boundary symbols % and -). Boundary tones indicate the right edge of an intonational phrase (IP) as well as of an intermediate phrase (ip). No clear-cut theoretical distinction seems to be made between these two kinds of phrases in the literature. We decided to label only IPs, indicating both the end of a phrase (L-L%) and the part of phrase after focus bearing nuclear accents.
In addition to this phonologically oriented labelling, also quantitative parameterisations of the intonation events are performed, namely *Tilt* modelling (Taylor 1998) and a calculation of *INTOFIT* parameters (Portele et al.1995) which parameterise the height and the slope steepness of intonation peaks. This will allow for a comparative evaluation of the two parameterisation methods as well as a phonetic specification of the ToBI tones.
Furthermore, we annotated perceived prominence on the syllabic level using a scale from 0 to 4. As this rather pre-theoretic factor draws considerable attention in prosody research, the inclusion of perceptual prominence information hopefully will facilitate the interchangeability of data and result. The basic question we want to address is whether it is possible to isolate the factors intonation, duration, lexical stress and information structure, or if some of them are strictly correlated, at least in German. In order to do so, perceptually determined prominence seems to be an indispensable control measure.
Due to an evaluation after the independent labelling procedures, the correspondence between strong syllables (prominence labelled syllables) and pitch assignment turned out to be straightforward. Pitch accents as H, L events at the intonational level are linked to events of prominence.
Information structure is a major factor influencing prosody. Unfortunately, it can not be reliably deduced for

unrestricted text, but within the q/a-corpus the focus structure of the answers is systematically controlled by the corresponding questions and by focus-attracting particles. Thus, in this part of the corpus prosody can be related to information structure directly.

4 Representation

The diverse information collected in a corpus like ours poses a number of practical problems for the representation of this data. The lack of standardised formats and tools for the specific purpose to be carried out here multiplies the necessary efforts and handicaps the interchangeability of data and results.

Typically, the data collected in a speech corpus includes sequential as well as hierarchical information which may be underspecified. A representation should facilitate the maintainability of these interconnected labelling data structures. This is complicated by the fact that different annotation types usually are produced and edited with different tools, which use their distinct encoding and file formats. Last but not least a representation should facilitate possibly complex queries on different levels, such as "extract all closed vowels preceded by an unvoiced plosive in strong syllables".

Recently some interesting proposals on developing a general annotation scheme have been published (Bird & Liberman 1999). These annotation graphs up to now are conceived basically as a formal framework, which does not codify yet the technical realisation in terms of file formats and querying tools.

Other proposals such as the TEI (Text Encoding Initiative, cf. TEI P3 1994) or projects like MATE (Multilevel Annotation, Tools Engineering, cf. Isard et al. 1998) provide standards and working tools for large text corpora. However, the main focus of these implementations lies on text, not on annotation of phonetic/phonemic structure, intonation, and related hierarchies.

These SGML- or XML-based approaches will clearly facilitate the interchangeability of all kinds of language-related data in the future. However, since these approaches are in part work in progress, and not specifically designed for our purposes, we are taking steps of representing the whole set of data collected in our corpus in its own format. Here we attempt to combine the benefits of both, relational and non-relational data-bases. This facilitates the manipulation and parsing of files with simple standard tools and allows for the straightforward encoding and retrieval of relational data.

The actual format looks as follows: Each label is a record which consists of 9 fields. The first is a key which encodes the type of segment, an index and relates it to the sentence (utterance) it occurs in and the soundfile. The second and third encode start and duration in absolute time (unit 10^{-5} sec). The fourth and fifth are used for SAMPA-transcription and text, respectively. The sixth is reserved for all kinds of tags in the form "key=value". The last three fields encode hierarchical information, first and last segment of the lower tier and the path of dominating nodes upwards. For an example for a typical representation on the syllable level (encoded by the prefix 'y' on the key) consider the following two labels representing the sequence "...und die...":

```
y011.s020.fa1|371.75383|0.17896|
     ?'Unt|und_1|pr=0|c032|c035|w007

y012.s020.fa1|371.93279|0.07800|
     d'i|die_1|pr=0|c036|c037|w008
```

Note that the absolute times in field 2 and 3 are not the result of actual segmentation, they are rather transmitted upwards from the canonical phonemes c032-c035, c036-c037 which in turn receive their time values from the phonetic segments, matching rules applied. Accent can be read off from the transcription (secondary accent (and shortened vowel) due to their status as a function word) and could in principle be additionally encoded as a tag in field 6, where we already find the tag for prominence (pr=0), but no tag for tone. The last field relates the syllable to the corresponding word label (in the same sentence/file.)

Now consider the labels corresponding to the boundaries between the two syllables on the canonical-phonemic level (suffix 'c') and the phonetic level (suffix 'p')

```
c035.s020.fa1|371.89478|0.03801|
     t||om=-|p034|p034<|y011.w007

c036.s020.fa1|371.93279|0.03801|
     d|||p034>|p034|y012.w008

p034.s020.fa1|371.89478|0.07601|t-d
```

The fact that there is only one segment on the phonetic level, which corresponds to two phonemes is coded in p034 by its transcription: 't-d' means that the first 't' is has no direct acoustic correlate (except for duration of the segment as a whole.) On the canonical-phonemic level the same thing is coded by giving the first segment the attribute 'om = –', which means by definition "missing on the phonetic level" and by the arrow brackets on the relational fields, e.g., p034<. The absolute time of that marker is then set in the middle of the duration of p034, but could in principle be calculated in any way. For statistic purposes, such sequences will either be filtered out or treated as a separate class. In order to recognise them one only has to look at fields 7 and 8 or to create a special tag in field 6.

The advantage of this way of encoding is that all information related to one record, be it hierarchical or relational (e.g. tone, prominence) is found within that very record, while the coding format is defined as a fixed set of

fields. Query algorithms can be stated very simply while they are able to convey quite complex tasks. Converting the data and its structure, or parts of it, into other formats (e.g., Prolog, or XML) can be done by using simple tools.

5 Conclusion

The Vienna Prosodic Speech Corpus has been created to serve several purposes at the same time. The ultimate goal, to link lexical, syntactic and focus-structure information onto prosody, and in particular phoneme duration, can only be achieved if one investigates the relevant components in isolation.

In order to do so, the corpus must be designed in a way that the specific factors can be maximally controlled. Annotations, whatever tools they are created with, should include all the information in a maximally transparent way.

In order to make the information retrievable for different kinds of analyses, much attention has been paid to the representation of information. We have developed a coding system specific for the purposes of the corpus. Every item of the corpus, be it phonetic labels, hierarchical labels like syllables, feet or phrases, intonational markers or prominence labels, has its own record with a fixed format. But in addition to that, hierarchical and relational information, as well as features either derived from relational dependencies or indicated specifically for that item, are encoded within that record. This should make the corpus maximally useful for all kinds of investigations into prosody.

6 References

[1] Alter K., Buchberger E., Matiasek J., Niklfeld G., Trost H.: VIECTOS: The Vienna Concept-to-Speech System, in Gibbon D. (ed.), Natural Language Processing and Speech Technology, Mouton de Gruyter, Berlin, 1996.

[2] Bird S., Liberman M.: A Formal Framework for Linguistic Annotation, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Tech Report MS-CIS-99-01, 1999.

[3] Grice M., Reyelt M., Benzmüller R., Mayer J., Batliner A.: Consistency in Transcription and Labelling of German Intonation with GToBI, Proc. of ICSLP'96, Philadelphia, pp. 1716-19, 1996.

[4] Isard A., McKelvie D., Cappelli B., Dybkjær L., Evert S., Fitschen A., Heid U., Kipp M., Klein M., Mengel A., Møller M.B., Reithinger N.: Specification of workbench architecture. MATE Deliverable D3.1, August 1998.

[5] Kohler K.: Lexica of the Kiel PHONDAT Corpus, Read Speech, Institut für Phonetik und Digitale Sprachverarbeitung, Universität Kiel, Vol.I, Arbeitsberichte Nr.27, 1994.

[6] Pirker H., Niklfeld G., Matiasek J., Trost H.: From Information Structure to Intonation: A Phonological Interface for Concept-to-Speech, in Proc. of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Université de Montreal, Canada, pp.1041-1045, 1998.

[7] Portele T., Krämer J., Heuft B., Sonntag G.: Parametrisierung von Grundfrequenzkonturen, DAGA-95, Saarbrücken, 1995.

[8] Shih Ch., Gu W., van Santen J.P.H.: Efficient Adaption of TTS Duration Model to New Speakers, in Proc. of ICSLP'98, Sydney Australia, 1998.

[9] Taylor P.: The Tilt Intonation Model, in Proc. of ICSLP'98, Sydney, Australia, 1998.

[10] TEI P3, Sperberg-McQueen C. M., Burnard Lou (eds.), Guidelines for Electronic Text Encoding and Interchange, The Association for Computers and the Humanities (ACH), The Association for Computational Linguistics (ACL), The Association for Literary and Linguistic Computing (ALLC), Chicago, Oxford, 1994.