The Partitur Format at BAS

Florian Schiel[1] & Susanne Burger[2] & Anja Geumann[2] & Karl Weilhammer[2]

[1]Bavarian Archive for Speech Signals (BAS), Munich, Germany
[2]Department of Phonetics and Speech Communication
University of Munich, Germany
[bas@phonetik.uni-muenchen.de]

Abstract

Most spoken language resources are produced and disseminated together with symbolic information relating to the speech signal. These are for instance orthographic transcripts, labelling and segmentation on the phonologic, phonetic, prosodic, phrasal level. Most of the known formats for these symbolic data are defined in a 'closed form' that is not flexible enough to allow simple and platform-independent processing and easy extensions.

At the *Bavarian Archive for Speech Signals* (BAS) a new format has been developed and used over the last few years that shows some significant advantages over other existing formats. This paper describes the basic principles behind this format, discusses briefly the advantages and gives detailed definitions of the description levels used so far. Furthermore, we will give some examples for easy processing of the format and distributed work on the same data.

In the future all corpora produced and disseminated by BAS will be distributed with the new BAS Partitur Format, if they contain segmental information of any kind. The former used formats will be retained but not further updated.

General Overview

Most file formats containing segmental information on speech signals have the disadvantage that

- they are not easy to extend (without rewriting software that uses the existing format).
- they are not easy to process with UNIX standard tools.
- they mix different description levels (which leads to technical and conceptual problems)
- they were defined as ad-hoc solutions for very specialised problems and are not capable of being re-used in a different setup.

Therefore a new open format based on the SAM Label Format was developed, which circumvents most of the mentioned problems. In this format all levels of description may be annotated independently but are time aligned like the individual tiers of a score. Hence this format was called 'BAS Partitur Format' ('Partitur' = German for 'score').

In the future all BAS corpora will be distributed with the new BAS Partitur Format if they contain segmental information of any kind. The formerly used formats will be retained but not further updated.

A first draft of the BAS Partitur Format was published in [Atmanspacher et al., 1995].

The BAS Partitur Format has the following features:

- SAM compatible header structure
- Easy to extend and to process by simple UNIX commands
- Open format; extensions to the format can be implemented without alterations to the software that reads the older format
- Time-aligned independent description of a virtually unlimited number of different levels of the speech signal (see examples later in this paper).
- Symbolic links between the independent levels allow logical assignments aside from the physical time scale. These links are based on the word units of the utterance.

Definition (Version 1.2)

A Partitur file name has the same prefix as the corresponding signal file (8 Bytes for Iso 9660 compatibility) but the extension *.par*. All contents are in 7-bit ASCII exclusively (to guarantee portability to all platforms). Each line starts with a three-byte label followed by a colon; this label defines synopsis and semantics of the ensuing line. The following units of the line are separated by 'white spaces' (blank, tab).

The Partitur file is structured into a header and a body (like SAM description files). The header stretches from the beginning of the file to the label *LBD:*; the body from the label *LBD:* to the end of file where the last line has to be closed by a 'new line' or a 'CR + LF' (the final SAM label *ELF:* was omitted

for the BAS Partitur Format since it prevents effective processing of the Partitur files).

The header contains SAM-compatible lines of general information. The following entries are compulsory:

```
LHD: Partitur file version
REP: Place of recording
SNB: Number of Bytes per Sample
SAM: Sampling Frequency in Hz
SBF: Byteorder (Intel 01, Motorola 10)
SSB: Bit Resolution
NCH: Number of Channels
SPN: Speaker ID
LBD:
```

Example:

```
LHD: Partitur 1.2
REP: Muenchen
SNB: 2
SAM: 16000
SBF: 01
SSB: 16
NCH: 1
SPN: PS1
LBD:
```

The following entries are optional; apart from these, other entries are tolerated as long as they do not conflict with compulsory and optional entries:

```
FIL: SAM File Type
TYP: Type of SAM Label File
DBN: Corpus Name
VOL: Number of Volume
DIR: Directory in Volume
SRC: Name of speech file
BEG: Begin of labelling sequence
END: End of labelling sequence
RED: Date of Recording
RET: Duration
RCC: Recording Conditions
CMT: Comment
SPI: Speaker Information
PCF: Name of Protocol File
PCN: Protocol Number
EXP: Name of Segmenter
SYS: Labelling System
DAT: Date of Labelling
SPA: SAM-PA Version
```

The body starts after the label *LBD:* and stretches to the end of file. It contains the different tiers of the BAS Partitur Format. Each tier is identified by a unique label. The order of tiers as well as the order of lines within a tier is not significant.

In the following sections the five basic classes of tiers are defined.

Tiers with symbolic relation *(class 1)*

A line of this tier contains:

- the tier label
- a comma-separated list of integers (*symbolic links*)
- a string with the labelling information

The symbolic links refer to a reference tier which numbers the word units beginning with zero. The label string has an internal synopsis which is defined in the tier definition.

Example:

```
TRL: 6,7 mit´m
```

In this example the word events 6 and 7 of an utterance are transliterated.

Tiers with time-consuming events *(class 2)*

A line of this tier contains:

- the tier label
- two integers denoting the begin and duration of the event.
- a string containing the labelling information

The semantics of the integers is defined by the tier definition (possible are samples, millisecs, etc.)

Example:

```
PHN: 13456 3450 aU
```

In this example a phonemic segment labelled */aU/* stretches from sample *13456* for the next *3450* samples.

Tiers with non time-consuming events *(class 3)*

A line of this tier contains:

- the tier label
- an integer denoting the time position of the event
- a string containing the labelling information

Example:

```
PRO: 13456 TON: P*; FUN: PA
```

In this example the prosodic event labelled *TON: P*; FUN: PA* (GTobi, see [Grice et al., 1995]) takes place at sample *13456* of the utterance.

Tiers with time and symbolic relation, time-consuming *(class 4)*

A line of this tier contains:

- the tier label
- two integers denoting the start and duration of the event.
- a comma separated list of integers (*symbolic links*)
- a string containing the labelling information

Example:

```
SAP: 13456 3450 9 aU
```

In addition to the example above this tier not only gives the starting point and the duration of the phonemic segment but also a pointer to the word unit where it belongs (word 9).

Tiers with time and symbolic relation, not time-consuming *(class 5)*

A line of this tier contains:

- the tier label
- an integer denoting the time position of the event.
- a comma separated list of integers (*symbolic links*)
- a string containing the labelling information

Example:

```
PRB: 13456 9 TON: P*; FUN: PA
```

Again, in this example the prosodic event is not only placed in time but also assigned to a word of the utterance (word 9).

Remarks

- If not otherwise noted, durational parameters are given in samples counting from the beginning of the digitised utterance
- An item may be referred to more than one word in the utterance (suprasegmental events, assimilation at word boundaries, phrases, etc.)
- If the symbolic link in a tier is not (or not yet) known, the symbolic link is set to *-1* (e.g. noises from other sources than the recorded speaker).
- The same symbolic relation may occur in different lines of a tier (for example if more than one event can be assigned to the same word of an utterance).

Definition of Tiers

The following sections give an overview of the currently defined tiers in the BAS Partitur Format (version 1.2.2). Please keep in mind that this is an *open list* in the sense that new tiers can be defined whenever there is a need for it. If somebody would like to work with speech resources from BAS and to define a new tier for his or her specific problem, please contact the BAS to get a new tier label assigned. By doing this we can keep up a consistent documentation of the format and avoid conflicts between matching labels. The version of the BAS Partitur Format is incremented by one on the third digit whenever a new tier definition is added to it. In accordance to the basic principle this does not imply that any software has to be changed.

Canonical Pronunciation

- *Tier label:* KAN
- *Class:* 1
- *Synopsis:* (symbolic links) (transcript)

This tier is the **reference tier** for all other tiers that use symbolic links. It contains a list of the spoken words within the utterance annotated in extended German SAM-PA (see [SAM, 1989] for a general definition of the SAM-PA and [SAM, 1996] for a special description of the extended German SAM-PA as used in several German projects). Note that these forms are the phonologically expected citation forms, **not** the actually spoken form.

The segmentation of the whole utterance is done into word units, where everything counts as a word that is produced by the articulatory organs of the speaker and can be seen as *speech*. Following this definition hesitations are words, whereas laughing, coughs, etc. are not. This separation isn't always clear, but on the other hand the selection of word units is arbitrary as well. The main point here is a unique reference tier for symbolic relations in other tiers. Another problem is the reduction of words that are annotated in the orthographic form, e.g. "mit'm". In these cases the reduction is restituted (in this example /mIt de:m/). The reason for this lies in the fact that some of these reductions should later be automatically accessible.

Example:

```
KAN: 0  j´a:
KAN: 1  Qalzo:
KAN: 2  QE:m
KAN: 3  h´OYt@
KAN: 4  Qo:d6
KAN: 5  m´O6g@n
```

Orthography

- *Tier label:* ORT
- *Class:* 1
- *Synopsis:* (symbolic links) (lexical orthography)

The tier *orthography* contains the orthographic (lexical) strings corresponding to the units in the tier *canonical form*. Words are not capitalised at the beginning of an utterance or sentence within an utterance (except nouns of course). German 'Umlauts' and other letters not included within 7 Bit ASCII are written in LaTeX notation. This tier is used for easy lexical access; therefore no additional markers except lexical words are allowed. There is no punctuation in this tier. Lexical words include items that are contained in the KAN tier (e.g. hesitations, repairs, word fragments, etc.). This tier can be used to access customised pronunciation dictionaries, to create unique word frequency lists, etc.

Example:

```
ORT: 0  ja
ORT: 1  also
ORT: 2  <"ahm>
ORT: 3  heute
ORT: 4  oder
ORT: 5  morgen
```

Verbmobil Transliteration - VM I

- *Tier label:* TRL
- *Class:* 1
- *Synopsis:* (symbolic links) (transliteration)

The tier *transliteration VMI* contains the orthographic transcript of the utterance according to the VM I conventions 3.0. The transliteration is segmented into the units of the tier *canonical pronunciation*. Therefore multiple references may occur (e.g. if a reduced form of two words is written as one unit in the transliteration). Although especially defined for the German Verbmobil I project, this format has been used in many other resources of spontaneous speech as well. See [Kohler et al., 1994] (German only) or online [Burger, 1995] for a detailed description of the VM I Transliteration format.

Example:

```
TRL: 0  <A>
TRL: 0  ja ,
TRL: 1  also
TRL: 2  <"ahm>
TRL: 3  <:<#Klicken> heute:>
TRL: 4  oder
TRL: 5  morgen .
```

Verbmobil Transliteration - VM II

- *Tier label:* TR2
- *Class:* 1
- *Synopsis:* (symbolic links) (transliteration)

The tier *transliteration VMII* contains the orthographic transcript of the utterance according to the VM II conventions. A detailed definition of this format can be found in [Burger, 1997] (German only). In contrast to the VM I format this new updated definition has the advantage of being fully parsable. Furthermore, with this format multi-party and multilingual dialogs may be transliterated (because compatible definitions for the languages English and Japanese do exist). To denote overlapping speech parts between different speakers in a dialog, a new tier *SUP* was defined (see below).

Superimposed Speech - VM II

- *Tier label:* SUP
- *Class:* 1
- *Synopsis:* (symbolic links) (transliteration)

This is a very specialised tier to denote overlapping speech in multi-party recordings. The synopsis of the turn marker and the transliteration is defined for the VM II transliteration format (see above). The speech annotated in this tier stems from a different speaker who actively superimposes his speech on the speech of this Partitur file. See [Burger, 1997] (German only) for a detailed description of superimposed speech in the VM II format.

Example:

```
TR2: 0 ich
TR2: 1 w"urde
TR2: 2 vorschlagen ,
```

```
TR2: 3 da"s
TR2: 4 wir9@
TR2: 5 dann9@
TR2: 6 <:<#> hinfliegen:> ,
TR2: 7 <:<#> ich:>
TR2: 8 hab´
TR2: 9 jetzt <!1 jetz´>
TR2: 10 aber
TR2: 11 <:<#Rascheln> grade:>
TR2: 12 <:<#Rascheln> keine:>
TR2: 13 Unterlagen
TR2: 14 da . <#>
SUP: 4,5 g002acn2_028_AAK.par @9ja
```

In this example the utterance of another speaker (AAK, utterance "ja") is superimposed on the 4th and 5th word of the Partitur file (utterance "wir dann").

Broad Phonetic Segmentation - PhonDat

- *Tier label:* PHO
- *Class:* 4
- *Synopsis:* (integer) (integer) (list of symbolic links) (label string)

This tier contains a totally time-consuming segmentation into broad phonetic units (extended German SAM-PA). The first number denotes the beginning of the segment in samples counted from the beginning of the speech file; the second number the duration of the segment in samples. The label string contains an additional relation to the canonical pronunciation (aside from the symbolic links to the tier *canonical form*). The '-' sign denotes differences to the expected canonical pronunciation on a segmental level: a leading '-' sign means the following segment was inserted (e.g. /-a:/); a trailing '-' sign means the segment was deleted (e.g. /a:-/); a '-' sign between segment labels means that the canonical expected segment was replaced (e.g. /a:-E:/). This tier also contains prosodic and phrasal labelling and segmentation. The full conventions of labelling and segmentation for German are briefly described in [Pompino, 1992] or online in [PHO, 1995].

Example:

```
PHO:    6637    0       0       #c:
PHO:    6637    916     0       ##%Q
PHO:    7553    820     0       $I
PHO:    8373    919     0       $C+
PHO:    9292    870     1       ##m
PHO:    10162   1424    1       $9
PHO:    11586   724     1       $C
PHO:    12310   966     1       $t
PHO:    13276   689     1       $@+
PHO:    13965   0       2       ##Q-
PHO:    13965   0       2       $-q
PHO:    13965   2024    2       $a:
PHO:    15989   517     2       $b
PHO:    16506   1066    2       $6+
PHO:    17572   539     3       ##Q
PHO:    18111   820     3       $´U
PHO:    18931   1867    3       $n-N
PHO:    20798   0       3       $#g-
PHO:    20798   1111    3       $@
```

Broad Phonetic Segmentation - Verbmobil

- *Tier label:* SAP
- *Class:* 4
- *Synopsis:* (integer) (integer) (list of symbolic links) (label string)

In contrast to the *PHO* tier this segmentation is not stringently time-consuming. That is, there might be pauses in the signal that are not labelled (which happens frequently in spontaneous speech). Furthermore the conventions are different in some points to the *PHO* tier to simplify parsing and processing of the tier. *SAP* is an exclusively phonemic tier; there is no other information encoded here.

Example:

```
SAP: 2541 894 0 m
SAP: 3435 1140 0 aI
SAP: 4575 270 0 n
SAP: 4845 510 1 n
SAP: 5355 1326 1 a:
SAP: 6681 795 1 m
SAP: 7476 277 1 @
SAP: 7753 0 2 Q-
SAP: 7753 614 2 I
SAP: 8367 1457 2 s
SAP: 9824 0 2 t-
SAP: 9824 656 3 t
SAP: 10480 1796 3 s
SAP: 12276 1953 3 E
SAP: 14229 988 3 l
SAP: 15217 535 3 t
SAP: 15752 370 3 -H
SAP: 16122 2097 3 h
SAP: 18219 2608 3 o:
SAP: 20827 1643 3 f
SAP: 22470 4265 3 6q
```

A detailed description of the *SAP* labelling conventions can be found in [Geumann et al., 1997].

Automatic Broad Phonetic Segmentation by MAUS

- *Tier label:* MAU

- *Class:* 4
- *Synopsis:* (integer) (integer) (symbolic links) (label string)

This tier contains an automatically generated broad phonetic segmentation in units of German SAM-PA. The segmentation is done fully automatically by the MAUS system ([Kipp et al., 1996]). The segmentation is totally time-consuming and the labelling has no direct relation to the tier *canonical form* as done in the tier *SAP*. (However, there are symbolic links to the words). The units are labelled in extended German SAM-PA as in the definition of the *SAP* tier (see appendix A). Additional labels are <nib> (non-speech event) and <p:> (pause). These labels always get the symbolic link -1 (no link).

Example:

```
MAU: 0 676 -1 <p:>
MAU: 677 7861 -1 <nib>
MAU: 8539 450 0 g
MAU: 8990 2436 0 u:
MAU: 11427 1740 0 t
MAU: 13168 958 1 d
MAU: 14127 1298 1 a
MAU: 15426 3820 1 n
MAU: 19247 303 2 n
MAU: 19551 1785 2 e:
MAU: 21337 624 2 m
MAU: 21962 636 2 n
MAU: 22599 501 3 v
```

Word Segmentation

- *Tier label:* WOR
- *Class:* 4
- *Synopsis:* (integer) (integer) (symbolic links) (word label)

This tier contains a segmentation of the utterance in word or word equivalents. The segmentation need not to be stringent. The label string may contain orthographic or pronunciation information (e.g. in SAM-PA). A '-' at the end of the label string denotes a missing word in the reference of the tier *canonical from* (of course a missing word has zero duration); a leading '-' denotes an inserted word; a '-' between two words (*word1-word2*) denotes a replacement. The symbolic links give the relation to the tier *canonical form*. Note that inserted words have a symbolic link to the previous word in the reference tier.

Example:

```
WOR: 1245 13245 0 <"ahm>
WOR: 14490 10787 1 guten
WOR: 25277 5089 1 -<hm>     # insertion
WOR: 30366 8786 2 Tag
WOR: 39152 3089 3 ich
```

Dialog Act Segmentation

- *Tier label:* DAS
- *Class:* 1
- *Synopsis:* (symbolic links) (marker string)

This tier contains a segmentation in dialog acts according to the ongoing work of the 'Deutsches Forschungszentrums für künstliche Intelligenz' (DFKI), Saarbrücken, Germany. Each marker covers a portion of the speech signal that is denoted by the symbolic links to the reference tier *canonical form*. A description of the format can be found in [Jekat et al., 1995]) or online in [DAS, 1996].

Example:

```
DAS: 0,1,2,3,4,5    @m(REJECT_DATE)
                    @m(GIVE_REASON)
DAS: 6,7,8,9        @(SUGGEST_SUPPORT_DATE)
DAS: 10,11,12,13,14 @(REQUEST_SUGGEST_DATE)
```

Prosodic Segmentation - GTobi

- *Tier label:* PRB
- *Class:* 5
- *Synopsis:* (integer) (symbolic links) (marker string)

This tier contains the prosodic segmentation (by hand) according to GTobi defined by the Technical University of Braunschweig, Germany. A detailed description of the GTobi labelling format can be found in [Grice et al., 1995] or online in [PRB, 1996] (German only).

Example:

```
PRB:    54212    5   TON: H*; FUN: NA
PRB:    63269    7   TON: L+H*; FUN: EK
PRB:    76371    8   BRE: B3; TON: L-L%
PRB:    79967    8   TON: L*+H; FUN: PA
```

Easy Processing and Distributed Work

Since the Bas Partitur Format is strictly line structured, allows only 7-Bit ASCII and the order within a file does have no semantic meaning, it is very easy to use standard UNIX text processing tools like *gawk,*

grep or *sed* to work with data stored in this format. For example the following lines of *GAWK* code will analyse a stream of piped-in BAS Partitur files for a certain phoneme, capture the total length and summarise into a mean value:

```
/^MAU:.*aU$/ {count ++
              totallength += $3
             }
END          {print "Mean Duration for /aU/:"
              print totallength/count }
```

In the same manner single BAS Partitur tiers may be selected, updated or filtered using *grep*, tiers can be easily transformed into format suitable for different kinds of visualising tools (for instance the public domain software package SFS by University College London).

The German *Verbmobil* project gives a good example for the benefits of using the BAS Partitur format at different sites on the same data. For instance the tiers *DAS* and *PRB* were defined by partners at DFKI, Saarbrücken and University of Braunschweig respectively. Since such an extension does not require any basic software to be re-written, these cooperations using the same physical data went very smoothly.

References

[Atmanspacher et al., 1995] S. Atmanspacher, S. Burger, Chr. Draxler, A. Kipp, Chr. Scheer, F. Schiel, M.-B. Wesenick (1995). Partiturformat für die Darstellung unterschiedlicher Repräsentationsebenen von gesprochener Sprache (Verbmobil Memo 90-95). University of Munich, September 1995.

[Burger, 1995] Susanne Burger (1995). Transliterationslexikon (Verbmobil-TechDok 36-95). University of Munich, October 1995. (Online version in English: http://www.phonetik.uni-muenchen.de/VMTraLexeng.html)

[Burger, 1997] Susanne Burger (1997). Transliteration spontansprachlicher Daten - Lexikon der Transliterationskonventionen - Verbmobil II (Verbmobil-TechDok 56-97), University of Munich, April 1997. (Online version: http://www.phonetik.uni-muenchen.de/VMtrlex2d.html)

[Geumann et al., 1997] Anja Geumann, Daniela Oppermann, Felix Schaeffler (1997). The Conventions for Phonetic Transcription and Segmentation of German Used for the Munich Verbmobil Corpus (Verbmobil Memo 129-96). University of Munich, December 1997.

[Grice et al., 1995] Grice, Martine and Ralf Benzmueller (1995). Transcription of German Intonation using ToBI tones; The Saarbruecken System. Paper presented at Tutorial Workshop on Discourse and Dialogue Prosody, Stuttgart, February 1995, modified version also in Phonus 1, University of the Saarland, pp33-51.

[Jekat et al., 1995] Susanne Jekat, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, Joachim Quantz (1995). Dialogue Acts in Verbmobil (Verbmobil-Report 65). Universität Hamburg, DFKI GmbH, Universität Erlangen, TU Berlin. April 1995.

[Kipp et al., 1996] A. Kipp, M.-B. Wesenick, F. Schiel (1996). Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora; in: Proceedings of the ICSLP 1996. Philadelphia, pp. 106-109, Oct 1996.

[Kohler et al., 1994] Kohler, Lex, Pätzold, Scheffers, Simpson, Thon (1994). Handbuch zur Datenaufnahme und Transliteration in TP14 von VERBMOBIL - 3.0 (Verbmobil-TechDok 11-94). IPDS, University of Kiel, 1994.

[Pompino, 1992] Pompino-Marschall, B. (1992). PhonDat - Verbundvorhaben zum Aufbau einer Sprachsignaldatenbank für gesprochenes Deutsch. FIPKM 30/1992, pp. 99-128.

[DAS, 1996] http://www.phonetik.uni-muenchen.de/Bas/BasDialogaktDok/vm-report-for-partitur_1.html

[PHO, 1995] http://www.phonetik.uni-muenchen.de/Bas/BasFormatsPHOdeu.html

[PRB, 1996] http://www.phonetik.uni-muenchen.de/Bas/BasProsodie.html

[SAM, 1989] http://www.phon.ucl.ac.uk/home/sampa/home.htm

[SAM, 1996] http://www.phonetik.uni-muenchen.de/Bas/BasSAMPA