Forschungsberichte des
Instituts für Phonetik und
Sprachliche Kommunikation
der Universität München
(FIPKM) 35 (1997) 121-125

THE BAVARIAN ARCHIVE FOR SPEECH SIGNALS 1995 - 1997

F. Schiel

Institut für Phonetik und Sprachliche Kommunikations,
Ludwig-Maximilians-Universität München

INTRODUCTION

The Bavarian Archive for Speech Signals (Til95) was founded as a non-profit institution in January 1995 and is hosted by the Institut für Phonetik und Sprachliche Kommunikation at the University of Munich (http://www.phonetik.uni-muenchen.de/Bas/).

BAS is dedicated to making resources of spoken German accessible to the speech science community as well as to speech engineering. Furthermore, we seek to promote scientific progress in the new field of Computational Phonetics by applying new techniques of speech processing to large corpora. The outcome of these activities will hopefully influence the performance of ASR systems as well as Speech Synthesis systems.

In the section following this introduction we give a short description of the infrastructure we developed for BAS. In the following sections we briefly describe the two main resource types available at BAS: speech corpora and pronunciation knowledge and how to obtain them from BAS. In conjunction with that we describe a special format designed to act as a universal description of speech data (Partitur Format). The next two sections deal with some of the related scientific activities at BAS: - automatic annotation of large speech corpora (MAUS) and - speech recognition (ASR). Finally, a short overview of ongoing cooperations is given in the last section.

INFRASTRUCTURE

The administration and dissemination of large speech resources require – as in software development – a reliable technical and administrative structure. Speech resources (being empirical data) and their symbolic descriptions tend to have frequent updates and revisions. Therefore mass production of data media like CDROMs is clearly not the best means of dissemination. On the other hand the amount of data is often to big to use standard Internet transfer techniques like FTP. To ensure that each receiver of a resource obtains the latest revision we chose a 'press on demand' approach for our general BAS policy (there are exceptions though!).

Writable CDROMs ('Gold CDs') can store up to 670 MB of data, can be read by every computer platform that mounts Iso 9660 standard CDs and can be written at up to 4 times realtime (realtime is about 1 hour per CD). The following procedures are applied to each resource: Preliminary procedure:

- cut resource data into reasonable chunks of smaller than 650 MB size
- tore these chunks on-line as so called 'masters'

Dissemination procedure:

- press an actual copy of the required masters of the ordered resource
- ship to end user

esource data consist of three different types of data:

- empirical data (digitized speech, multi-channel measurements, etc.)
- symbolic data (annotations, time information, labeling, etc.)
- documentation

The latter two are usually of very small size compared to the first type and tend to be the most frequently updated/revised parts of the resource. On the other hand, the empirical data itself is too big to keep on-line all the time (at least that is true for the present situation). Therefore we decided to physically separate these data, keeping the symbolic data and documentation on-line all the time (and hence easily accessible for revisions), while the !!empirical!! data are stored on slower background storage devices.

In 1995 the Leibniz Rechenzentrum of Munich (LRZ) started a pilot study together with BAS on the usage of very large databases called TERABACK. Within this project BAS was assigned two storage resource accessible over a fast fiberoptic line:

- an on-line Andrew File System (AFS) storage space of 10.000 MB (called 'CACHE')

- a practically unlimited storage space based on IBM's professional backup technique 'ADSM' (called BACKROUND) [1]

Furthermore these two storage devices are connected by a fast FDDI fiberoptic connection, allowing very fast retrievals from the BACKROUND into the CACHE and vice versa (average of 0.5 MB / sec). The CACHE is subdivided into 15 slots for masters that can be kept online. Since these slots are accessible via AFS, they give an elegant way to include symbolic data and documentation by using standard UNIX file links to files that are physically stored on BAS workstations.

Finally, we set up two CDROM writers (CD-R) connected to BAS workstations to copy the combined empirical/symbolic/documentation data to writable CDROMs.

SPEECH CORPORA

Currently a total of 23.0 GB of German (and partly American) speech data is stored at BAS; a subset of 19.8 GB can be ordered on CDROM. The collection includes read speech, dictation speech, spontaneous dialog speech, a small portion of American dialog speech and some special corpora such as a collection of foreign accents. The domains cover Diphone Balanced German, Train Inquiry, Appointment Scheduling, Newspaper Texts and Robot Control. Most of the speech corpora are the outcome of former cooperations with industrial partners and of nationally and internationally funded research projects.

Evaluation:

Before being included in the BAS archive each corpus undergoes a detailed evaluation of the contents and the documentation. If not produced by the University of Munich, a protocol of this evaluation is distributed together with the online documentation of the resource.
The BAS evaluation procedure ensures that:

- Signals are readable.
- Headers (if any) are correct.
- Software (if any) is working.
- 10% of annotations (if any) are checked and errors found are reported.
- The online documentation is reviewed and augmented, if necessary.
- A package of standard software tools is added to the CDROM.

In a final step the corpus undergoes an automatic annotation using the MAUS system (see below) and the outcome is added to the volume.

The volumes are stored as master volumes in our archive system and CDROM copies are pressed only on demand to ensure that always the latest version of the corpus is disseminated to the user.

[1] Actually the limit for BACKROUND is presently about 10.000.000 MB and growing!

	GB	NS	Domain	Type	Record
PD1	2.2	201	diphone	read	studio
PD2	0.5	16	train inq.	read	studio
SI100	4.1	101	news	dict.	office
SI1000	4.2	10	news	dict.	office
VM	6.1	925	appointm.	spont.	office
SC1	0.2	88	story	read	studio
ERBA	2.2	106	train inq.	read	office
SPINA	0.3	22	robot	read	office

Table 1: Speech corpora available at *BAS*

Availability:

Table 1 gives an overview of all available speech corpora together with their amount of speech data in GByte (GB), number of speakers (NS), domain (Domain), type of speech (Type) and recording environment (Record). Please refer to the BAS web pages for additional details. BAS speech corpora as well as other BAS resources may be ordered via email from the following address: bas@phonetik.uni-muenchen.de.

Partitur format:

Speech resources, especially speech corpora, generally are accompanied by symbolic information of different levels and contents (so-called tiers). For instance, the Verbmobil corpora come with up to 9 different types of information which are more or less connected to the speech wave. Furthermore, other (symbolic) relations may exist between the different tiers aside from the physical speech signal (consider for instance a segmentation in dialog acts, which is produced on the basis of a transliteration) and hence the update of such embedded tiers causes necessary changes in other levels of description.

Therefore a new concept for the representation of these different tiers - called Partitur (score) format - was developed at BAS.

The symbolic data belonging to each signal (e.g. text, labeling) is stored in a Partitur file with the same prefix as the signal file. Each level of description can be seen as a tier more or less synchronized to the underlying speech signal. The different tiers are independent but linked to each other over a reference tier that is usually the word level. In contrast to other formats the Partitur format is an open definition (that is a new tier can be added without any changes to any kind of user software) and does not need any special software or OS (because it is just a 7-Bit ASCII file). However, the overall structure was designed to be easily processed by standard UNIX line processors such as awk, sed, etc.

Currently the following tiers are defined and used within the BAS corpora:

- lexical orthography
- transliteration (orthography enriched by several markers as used in the Verbmobil projects)

- 'canonic' pronunciation (citation form)
- phonetic segmentation by hand
- automatic phonetic segmentation
- prosodic labeling
- segmentation of dialog acts

Please refer to the BAS web pages for an up-to-date description of the Partitur format.

PRONUNCIATION

Two resources of BAS deal with German pronunciation: the pronunciation dictionary PHONOLEX and the pronunciation rule set PHONRUL.

PHONOLEX is a very large German pronunciation dictionary and is produced and maintained by BAS in close cooperation with the University of Saarbrücken. It comprises more than 600.000 entries and is intended to cover the most commonly used German words. The dictionary is stored in 7-Bit ASCII and therefore independent of any special OS or software. The orthographic strings are given in LaTeX, the pronunciation in extended German SAM-PA (Sam90) as used in several nationally funded speech projects. PHONOLEX is a fully inflected word list, that is every lemma is stored in all possible inflections.

Each entry contains

- orthographic string (in LaTeX)
- syntactic markers (optional)
- 'canonic' pronunciation (citation form)
- list of empirically derived pronunciations together with corpus and number of detections (optional)

The 'canonic' pronunciation is generated by a combined system of exception lists and a rule-based text-to-phoneme system. The latter is a variant of the P-TRA system developed by D. Stock at the University of Bonn. The empirically derived pronunciations are the outcome of the MAUS project (see below). At the moment these cover only a small subset of PHONOLEX (Verbmobil word list), but will be extended to all BAS corpora in the future.

Example:

```
"Uberleben
nou
Qy:b6leb@n
*
"Uberlebens
nou
Qy:b6leb@ns
Qy:b6leb@ns     13      VM      MAUS
y:b6lems        40      VM      MAUS
Qy:b6lebns      23      VM      MAUS
*
"Uberlebensgr"o"se
nou
Qy:b6leb@nsgR2:s@
*
```

The pronunciation rule set PHONRUL was developed mainly to aid the automatic detection of pronunciation variants in the MAUS project (Kip96). This rule set has now been available for a few months as a new BAS resource. The rule set in its present form (version 9.0) consists of approx. 1500 complex rules which expand to 5546 simple re-write rules. The rule set was designed for extended German SAM-PA, but can be translated into other alphabets (e.g. Worldbet, IPA) without much effort. PHONRUL may be used either to provide a priori knowledge about German pronunciation or to support automatic annotation tools like MAUS. It is not intended to model German dialects. As a side effect of the usage of PHONRUL it is possible to provide statistics about the usage of a certain rule in a corpus (see next section for details).

AUTOMATIC ANNOTATION

Huge speech corpora like those archived in BAS cannot be labeled and segmented manually. Although we still produce small reference sets of manual segmentations, more than 95% of the BAS corpora are segmented and labeled with automatic tools. Our phonetic-phonologic segmentation tool MAUS (Munich AUtomatic Segmentation, Wes94,Kip96) is able to reach 95% transcription accuracy compared to the performance of trained phoneticians on spontaneous speech. MAUS is a three-staged system:

In a first step the orthographic string of the utterance is looked up in PHONOLEX and processed into a finite state graph containing all possible alternative pronunciations using PHONRUL. The second stage of MAUS is a standard HMM Viterbi alignment where the search space is constrained by the finite state graph (currently we use the HTK 2.0 as the aligner). The outcome of the alignment is a transcript and a segmentation of 10 msec accuracy, which is quite broad. Therefore in a third stage the segmentation is refined by a rule based system working on the speech wave as well as on other fine-grained features. Details about MAUS are given in [7].

In the future all BAS speech corpora will be processed with MAUS to yield a very large collection of phonetically annotated material. This material

is then used to provide reliable statistics about segmental context, pronunciation (e.g. straightforward in PHONOLEX or by rules in PHONRUL) or discriminative statistical models for speech recognition (see next section).

SPEECH RECOGNITION

This is a rather new activity at BAS although automatic speech recognition has had a long tradition at our host institution, the Institute of Phonetics at the University of Munich. The aim of this project is the consequent usage of the gained statistical knowledge about German pronunciation to improve automatic recognition of spontaneous speech. A number of experiments were conducted by several people within the last years to incorporate a priori knowledge about pronunciation into the search process of standard recognition systems. Although it never hurt the performance, the improvement by using multiple pronunciation was never significant enough to overcome the increased ambiguity of the lexical search space. Furthermore, many of these investigations were carried out using read speech, where the phonetic reductions in German are not very distinct and the underlying acoustic models were usually not changed for the experiment (that is the phonetic models were trained with embedded training to a fixed pronunciation dictionary that does not match the actually spoken utterance).

Our hypothesis is that the acoustic models will gain discriminative power when trained to a transcript that reflects the actual spoken utterance more accurately than a concatenation of citation forms is able to do. For instance we observed that the acoustic model for 'Schwa' in a standard HMM recognizer is able to model a variety of phonetic features (any vowel quality, nasals, laterals, voiced fricatives), even nonverbal sounds like telephone ringing. It does not hurt the recognition results if the acoustic scores of the 'Schwa' model are replaced by a constant factor per speech frame. The same is probably true for other phonemes that occur frequently in German final syllables, which are the subject of many types of reduction and coarticulation. Of course more discriminative acoustic models only make sense, if the modeling of words on the lexical level is done by a fully consistent statistical model, too. (Another possibility would be to skip the embedded forward-backward training and simply bootstrap with the whole training set. If this method shows comparable results, then it would yield a tremendous decrease in computational effort in the training phase.) Using a HTK standard recognizer with 46 mono-phone CDHMMs (9 mixtures per state, 12 MFCC + energy, delta, delta-delta) we set up two different systems: One (the baseline system) was trained in a traditional way using bootstrapping on a sub-corpus of 40 min of hand labeled speech and embedded forward-backward training on 30 h of unlabeled speech with a 'canonic' dictionary.

The second recognizer (the test system) was trained with the same amount of data and the same bootstrap material, but instead of the concatenated 'canonic' pronunciations we used the transcriptions delivered by the MAUS system for the embedded forward-backward training. Test and training material were selected from the Verbmobil corpus of spontaneous speech according to the last Verbmobil evaluation assessment carried out in 1996 by the Technical University of Braunschweig ('Kür', Rei96): the test set is approx. 50 min of speech of CD 14; the training set covers the CDs 1-5, 7 and 12. The language model is a traditional full bigram model derived solely from the training set and has a perplexity of approx. 51. The baseline system using the 'canonic' dictionary yields 65% word accuracy on the 1996 Verbmobil evaluation test set 4. In a second test we used the acoustic models of the test system but the same dictionary as in the baseline system. We expected a decrease of word accuracy, because the lexical model does not match the acoustic modeling. To our surprise we encountered a slight (non-significant) improvement instead.

In the next step word statistics of the MAUS transcriptions were calculated and incorporated into the dictionary and the search process of the test system. This includes a straightforward pruning technique to blend out the 'noise' (very rare observtions) from the statistical model: For each lexical item j (of all words in the dictionary $j = 1...J$) all observed pronunciations $o(j)$ are counted to $n(j)$. If $n(j)$ is smaller than the pruning threshold N, the observations are replace by the 'canonic' pronunciation with a posteriori probability 1.0. If $n(j)$ is bigger than N, the single observations $o(j)$ are ordered into K groups of identical transcripts and counted yielding $m(j,k)$. All groups where $m(j,k)$ is less than M percent of $n(j)$ are discarded. The remaining groups are summed up to $n'(j)$ and the a posteriori probabilities are calculated as $p(k'|j) = m(j,k')/n'(j)$. Finally the a posteriori probabilities are normalized to $max(p(k'|j))$ to avoid 'penalties for lexical items with plenty of observations. The pruning parameters N and M influence the size of dictionary extension. For example, using $N = 50$ and $M = 10\%$ yields to an increase of 47.1% (840 to 1236 words). Furthermore, insecure observations caused by bad acoustic conditions or errors of MAUS should be eliminated.

The following is an excerpt of the resulting resource:

```
verbessern    1.000000   f _6 b E s _6 n
verbleiben    0.269231   f _6 b aI b m
verbleiben    0.269231   f _6 b l aI b @ n
verbleiben    1.000000   f _6 b l aI b m
verbleiben    0.230769   f _6 b l aI b n
vereinbaren   0.280992   f _6 aI n b a: _6 n
vereinbaren   1.000000   f _6 aI n b a: n
```

Future work in this particularly field involves

- determining the optimal pruning values for statistical dictionaries.
- verify whether the use of this resource causes any improvements over standard techniques.

COOPERATIONS

BAS was and is involved in a couple of projects to produce new or improve existing speech resources (re-usability). Some of these projects – particularly the more scientifically motivated, like MAUS and ASR – are carried out by BAS itself. The majority of the remaining activities are done in close cooperation with industrial partners and scientific institutions, for instance Lucent Technologies, AT&T, Siemens, University of Bonn, University of Saarbrücken, DFKI, LIMSI, SpeechDat Consortium, Verbmobil Consortium. The following is a short overview of some of those activities that might interest other institutions.

German Dialects:

This speech collection is carried out in cooperation with Lucent Technologies and AT&T. The aim is a field collection that covers all German-speaking regions of Europe (including the 'new states' of Germany, Austria and Switzerland). The speech is recorded via two standard low cost microphones into an IBM compatible PC or Laptop (16 bit, 22.05 kHz) and via two high quality microphones (desk top and headset) on a DAT recorder (16 bit, 48 kHz). The prompted speech includes digits, connected digits, command phrases for computer related work places, phonetically balanced sentences, European telephone numbers and 1 minute of spontaneous speech. The prompted speech is validated (noise markers) and transcribed orthographically; the spontaneous speech is transliterated according to Verbmobil standards. The recruitment of the 500 speakers is done according to the demographic density; speakers are briefed about their dialectal behavior and classified into a German dialect chart.

Verbmobil: This is not a main BAS activity, but nevertheless we accepted to incorporate all Verbmobil speech resources into the BAS archive one year after the first release. The speech corpora are validated and extended by additional data of Verbmobil partners other than the University of Munich.

SpeechDat:

The German collection of SpeechDat I (1000 speakers) was carried out at our site and we are now in the process of collecting telephone speech of another 4000 German speakers (SpeechDat II). For this purpose we have set up several telephone servers based on ISDN technology to collect data in parallel sessions. Data are validated and transcribed using a HTML-based technique that allows us to run validation independently of hardware or location.
The outcome of these cooperations is available to the speech community via the BAS (or ELRA) a certain period after its first release (usually one year).

References

[1] Technical Report SAM, ESPRIT Project 2589, 1990.

[2] M.B. Wesenick, F. Schiel (1994): Applying Speech Verification to a Large Data Base of German to obtain a Statistical Survey about Rules of Pronunciation, Proceedings of ICSLP 1994, pp. 279 - 282, Yokohama.

[3] H.G. Tillmann, Chr. Draxler, K. Kotten, F. Schiel (1995): The Phonetic Goals of the new Bavarian Archive for Speech Signals, Proceedings of the ICPhS 1995, pp. 4:550-553, Stockholm Sweden.

[4] 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.

[5] J. Reinecke: Evaluierung der signalnahen Spracherkennung im Verbundprojekt VERBMOBIL (Herbst 1996), Verbmobil Memo 113, TU Braunschweig, Nov 1996.

[6] M.-B. Wesenick (1996): Automatic Generation of German Pronunciation Variants; in: Proceedings of the ICSLP 1996. Philadelphia, pp. 125 - 128, Oct 1996.

[7] A. Kipp, M.-B. Wesenick, F. Schiel (1997): Pronunciation Modeling Applied to Automatic Segmentation of Spontaneous Speech; Proceedings of the Eurospeech 1997. Rhode, Greece, to appear.