Speech and Speech-Related Resources at BAS

Florian Schiel

Bavarian Archive for Speech Signals (BAS)
University of Munich, Germany
[schiel@phonetik.uni-muenchen.de]

Abstract

The *Bavarian Archive for Speech Signals* (BAS) located at the *Ludwig Maximilians Universität München, Germany* collects, evaluates, produces and disseminates German speech resources to the scientific community. Our focus is the German language covering a large geographical part of central Europe.

Speech and speech-related resources are usually produced for certain tasks or projects. Therefore, it is not easy for scientists or engineers starting a new project or application to decide, whether existing resources may be re-used for their special purpose, or whether it is necessary to finance an new specialised data collection (which is usually very expensive). With this contribution we'll try to facilitate this decision by giving detailed information about existing resources as well as possibilities to produce new resources.

This paper has two major parts. The first part deals with our experiences during the last three years to produce new, highly re-usable speech resources in close cooperation with industrial partners. We will explain our BAS policies that ensures that valuable resources do not only satisfy short-term needs of certain industrial applications but are useful for other research or engineering activities as well. Some example projects will high-light these points. Finally we encourage other speech centres in Europe (for other languages than German) to adopt these policies.

In the second part we will give a concise description of all existing German speech and speech-related resources that are currently available at BAS. The description will not only give details about the properties of the corpus, a quality assessment in regard to possible applications, but also include an outline under what circumstances the original resource was produced. The latter information might be very important to answer special questions that are usually not contained in the standard description of a speech resource.

Introduction to BAS

The *Bavarian Archive for Speech Signals* (BAS) is a public institution hosted by the *University of Munich*. It was founded in January 1995 with the aim of making corpora of current spoken German available to both the basic research and the speech technology communities via a maximally comprehensive digital speech-signal database. The basic speech material should be structured in a manner allowing flexible and precise access, with acoustic-phonetic and linguistic-phonetic evaluation forming an integral part of it. 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.

BAS is mainly funded by the *Bavarian State* (scientific staff) and the *Ludwig Maximilians Universität München* (infra-structure, administrative staff).

The last few years have seen an abrupt increase in the demand for large speech signal data collections, both on the part of academic investigators carrying out basic research as well as on the part of engineers from industry working in the new integrated field of speech and information technology. There are many reasons for this. Primarily, however, the sudden increase in demand must be attributed to the breakneck pace of hardware and software development in speech signal processing. The increasing number of techniques for acoustic-phonetic signal processing, and the increasing amount of speech data that can be efficiently handled and processed together generate an accompanying demand not only for linguistically interesting text material (which of course emerges automatically from the modern printing industry) but also for reliably acquired and phonetically evaluated spoken language material. A number of national and international initiatives (such as BDSON, PHONDAT, LDC, SPEX or COCOSDA) have already resulted in the collection and distribution of large speech corpora. However, they exhibit a variety of formats, corresponding to the variety in the aims pursued. For German, a central institution was clearly lacking that could carry out such tasks within a long-term perspective. (http://www.phonetik.uni-muenchen.de/Bas/).

The BAS Policies Outlined

The production of a new speech resource is always twofold: Usually expensive resources are not produced without a reason. The most common reasons are that the resources will be necessary for a publicly funded research project or that the resources are needed for the development of a new product. In both cases it's in the interest of the funding part of the project not to spend more money as is needed to fulfil the basic needs at hand. On the other hand we all know that resources which were designed and produced with the solely purpose of a single project or product will not be very useful for other tasks in the future. With other words: there always will be the opposition of expenses versus re-usability.

Our work at BAS during the last 3 years has shown that in most cases *it is possible* to combine these antagonistic positions in a way that satisfies both sides perfectly. Let us give two examples:

The *Regional Variants of German 1* (RVG1) corpus that was collected in close cooperation with *AT&T* and *Lucent Technologies* contains speech data for short-term needs (digits, phone numbers, commands) as well as phonetically rich sentences and spontaneous speech for research and future application purposes (see [Burger, 1998] in this proceedings). Furthermore, the data were collected not only with standard low cost recording technique but with two additional high quality channels.

The first EU funded *SpeechDat I* project (1000 speakers per language recorded over public telephone lines, see [Draxler, 1996]) contained phonetically rich sentences and - in some languages - spontaneous speech as well.

In both examples the monetary effort was mainly due to the recruiting and recording techniques. The additional cost for the extra recordings were marginal. But these extra recordings already begin to play a major role in the exploitation of these resources. (This can be seen for instance by the fact that in SpeechDat II the phonetically rich words and sentences got a prominent role.)

The last three years have also shown that there is a growing demand of Speech Engineering for very specialised speech corpora. Examples are:

- speech under certain real life noise conditions
- speech in the running car
- speech for certain dictation domains
- speech of children
- speech with special microphones/headsets/build-in-mics
- speech with special dialects
- speech with non-prompted utterances

At the moment we do not expect this tendency to come to an end. Thus, the production of re-usable resources should be a major issue for all speech resource producing sites in Europe.

Consequently, our main policy issues at BAS are:

1. BAS offers to co-produce specialised corpora together with industrial or scientific partners, but – whenever possible – BAS will make sure that the properties of the resulting resource will be usable for more than one very specialised application.
2. BAS will negotiate a certain time span for the resource, during that the funding partner(s) will have the exclusive rights to use the resource for his/their commercial application. After that 'blocking period' (usually 1 year) the resource will be available to other members of the scientific community from BAS (a possible license fee is negotiated with the industrial partner as well).
3. BAS takes care that the produced resource satisfies a minimum of internationally accepted standards of technique and quality, thus again improving the re-usability of the resource by others than the directly involved partners.

Following these three basic policies we hope that in a few years there will be a vast amount of re-usable speech and speech-related resources available at BAS.

We strongly recommend other speech focus centres in Europe to follow similar policies while producing speech resources in other European languages. We would be very interested to discuss these points with colleagues from other European countries.

Resources at BAS

The following is a short listing of all speech resources currently available at BAS. Please note that in-depth information to each resource can be found on our Web server. In most cases even online access to the original documentation is possible via WWW. Also note that resources that are not available for the next year starting with April 1998 are listed in a separate section at the end of the paper.

Strange Corpora

The 'Strange Corpora' series was motivated to facilitate the investigation of certain well known problems in speech engineering as well as in the speech sciences. Such fields of investigation are:

- Speaker characteristics (speaker adaptation / normalisation)
- Pathological speech
- Speech of children or the elderly
- Speech in real life noise (Lombard effects, robustness)
- Prompted and non-prompted speech (intonation)

- Typical spontaneous speech effects as hesitations, repairs, breaks
- Accents
- Dialects

The SC series is a collection of smaller corpora (compared to nowadays collections like the SpeechDat project!) which give well documented reference data (bench marks) in the above mentioned topics. Researchers as well a speech engineers might use these corpora to verify their algorithms or applications under controlled and reproducible conditions.

Currently available:

SC1 - Accents
Type: read speech
Format: 14 Bit / 16 kHz / PhonDat
Environment: studio
Recording sites: 1
Speakers: 88
Transcript: orthographic
Segmentation: phonemic
Total: 88 stories (111 words each)
Medium: 1 CDROM
Description:
The corpus contains the same text read by 16 native German speakers and 72 speakers from other cultures/countries. The reference speakers (native Germans) were manually segmented and labelled into SAM-PA phonemic segments ([SAM, 1989]).
Original Purpose of Recording:
Scientific investigation of foreign accents; forensic classification of unknown voices.
Other Usabilities:
Automatic accent detection; adaptation to foreign accents; robust ASR; forensic applications; speaker verification.

Available in Jan 1999:

SC2 - Noises
Type: read speech
Format: 16 Bit / 16 kHz / NIST
Environment: car maintenance hall
Recording sites: 1
Speakers: 10
Transcript: orthographic
Segmentation: noise markers
Total: 8000 utterances (average 4.6 words)
Medium: 1 CDROM
Description:
10 speakers from a car diagnosis firm were asked to read 800 'automobil diagnosis phrases' from a corpus of 100 different phrases (each phrase read 8 times). The recording took place in a car maintenence hall with up to 6 active car lines. The speech was prompted via screen; the speech signals were recorded via a DECT phone system directly to a portable IBM compatible PC. Backround noise of all kinds were classified during the validation process.
Original Purpose of Recording:
Speech recognition for car diagnosis.
Other Usabilities:
Robust ASR under heavy noise conditions; Lombard effects; noise cancelling techniques.

Read Speech Corpora

The following speech corpora contain different types of read speech, whole utterances, commands and single words.

PD1
Type: read speech
Format: 16 Bit / 16 kHz / PhonDat
Environment: studio
Recording sites: 4
Speakers: 201
Transcript: orthographic
Segmentation: phonemic
Total: 21681 utterances (average 7.3 words)
Medium: 4 CDROMs
Description:
The corpus contains carefully read speech of 201 speakers recorded in a echo cancelled studio environment. The speech corpus was selected to cover all possible di-phone combinations in the German standard language (without foreign words). The text corpus consists of 450 different sentence equivalents (including alphanumericals and two shorter passages of prose text) and is not domain specific.
Original Purpose of Recording:
Diphone based ASR.
Other Usabilities:
Bootstrapping ASR; concatenative speech synthesis.

PD2
Type: read speech
Format: 16 Bit / 16 kHz / PhonDat
Environment: studio
Recording sites: 4
Speakers: 16
Transcript: orthographic
Segmentation: phonemic,prosodic,words
Total: 3200 utterances (average 12.3 words)
Medium: 1 CDROM
Description:
The corpus contains 16 x 200 sentences from the train inquiry task fluently read by 16 native German speakers. A subcorpus of 64 sentences per speaker was manually segmented and labelled into SAM-PA segments. The whole corpus was automatically segmented by MAUS. The data of 8 speakers (8000 utterances) were annotated and segmented prosodically.
Original Purpose of Recording:
ASR for a train inquiry system.
Other Usabilities:
Bootstrapping ASR; phonetic investigations; prosodic investigations; ASR using prosodic features.

ERBA
Type: read speech
Format: 14 Bit / 16 kHz / RAW
Environment: office
Recording sites: 4
Speakers: 106
Transcript: orthographic
Segmentation: none
Total: 11100 utterances (average 13.1 words)
Medium: 4 CDROMs
Description:
The corpus contains 101 x 100 (training) and 5 x 200 (test) sentences read by native German speakers. Sentences are unique (with some exceptions) and produced by a stochastic sentence generator (grammar). (Therefore, some sentences are somewhat unusual.) Availability is limited to scientific usage.
Original Purpose of Recording:
ASR for a train inquiry system.
Other Usabilities:
General ASR; speaker adaptation; speaker identification.

SPINA
Type: read speech (mostly single words)
Format: 16 Bit / 16 kHz / RAW
Environment: studio
Recording sites: 2
Speakers: 22
Transcript: orthographic
Segmentation: phonemic,word
Total: 10810 utterances (average 1.2 words)
Medium: 1 CDROM
Description:
The corpus contains very specific commands to control an industrial robot. The text corpus consists of 10 robot command sentences and 62 robot command words. Each speaker has read the entire text corpus at least 5 times. Small parts of the corpus are segmented and labelled into SAM-PA and word units.
Original Purpose of Recording:
ASR for robot control.
Other Usabilities:
General ASR.

RVG 1
Type: screen prompted speech
Format: 16 Bit / 22.05 kHz / NIST
Environment: office
Recording sites: 6
Speakers: 500
Transcript: orthographic (with linguistic markers and noise class)
Segmentation: none
Total: 42000 utterances (average 6 words)
Medium: 18 CDROMs
Description:
The *Regional Variants of German* (RVG1) corpus contains 85 screen prompted utterances (digits, phone numbers, computer command phrases, phonetically rich sentences). The 500 speakers were selected according to demographic densities in Germany, Austria, parts of Switzerland and Italy. Speakers were asked to speak informally but not dialectally. See [Burger, 1998] in this proceedings for an in-depth description of this resource.
Original Purpose of Recording:
ASR training material with broad regional coverage.
Other Usabilities:
General ASR; research of pronunciation variants; prosody; phonetic investigations.

Dictation Speech Corpora

The following speech corpora contain read speech recorded in a dictation task. The spoken texts were derived from a German newspaper corpus.

SI1000
Type: dictated speech
Format: 16 Bit / 16 kHz / PhonDat
Environment: studio
Recording sites: 1
Speakers: 10
Transcript: orthographic
Segmentation: prosodic (in text corpus)
Total: 10000 utterances (average 25.1 words)
Medium: 5 CDROMs (compressed)
Description:
The corpus contains 1000 sentences from a newspaper corpus read by 10 native German speakers in a dictation task (puntuations are spoken). The text corpus is segmented into phrase boundaries B2, B3 and B9 (GTobi, [Grice et al., 1995]) and words are marked with accent labels PA, NA and EK.
Original Purpose of Recording:
ASR for dictation.
Other Usabilities:
Prosodic segmentation; speaker adaptation; speaker verification.

SI100
Type: dictated speech
Format: 16 Bit / 16 kHz / PhonDat
Environment: studio
Recording sites: 1
Speakers: 101
Transcript: orthographic
Segmentation: none
Total: 10100 utterances (average 23.4 words)
Medium: 7 CDROMs
Description:
The corpus contains 101 x 100 sentences selected from two different newspaper text corpora (544 + 483 sentences) read by 101 native German speakers in a dictation task (puntuations are spoken).
Original Purpose of Recording:
ASR for dictation.
Other Usabilities:
General ASR; speaker adaptation; speaker identification.

Spontaneous Speech Corpora

The following gives an overview about speech corpora at BAS containing or consisting entirely of spontaneous elicited speech. The term *spontaneous* as it is used in this paper does not imply a totally unaware recording. The correct terminus technicus would be *un-scripted speech*. However, since the term *spontaneous speech* is used in many publications and documentations, we'll stick to it.

German VM I
Type: dialogues
Format: 16 Bit / 16 kHz / PhonDat
Environment: studio/office
Recording sites: 4
Speakers: 779
Transcript: VM I + VM II transliteration ([Burger, 1995], [Burger, 1997])
Segmentation: phonemic,word,prosodic,dialogact
Total: 13910 utterances (average 22.8 words)
Medium: 9 CDROMs
Description:
The German Verbmobil I corpus contains 1956 dialog recordings of 779 different speakers. In each dialog both speakers had to negotiate up to 4 business appointments. The whole corpus was segmented and labelled by MAUS into SAM-PA segments; parts of the corpus were segmented manually with regard to phonology, prosody and dialogacts.
Original Purpose of Recording:
ASR for online translation German to English / Japanese.
Other Usabilities:
General ASR; dialog systems; research of elicited spontaneous speech; prosody; phonetic investigations.

RVG 1
Type: monologues
Format: 16 Bit / 22.05 kHz / NIST
Environment: office
Recording sites: 6
Speakers: 500
Transcript: VM II transliteration ([Burger, 1997])
Segmentation: none
Total: 500 x 1 minute monologue
Medium: (18 CDROMs)
Description:
The *Regional Variants of German* (RVG1) corpus contains – besides the prompted speech (see above) – 1 minute of free monologue of each speaker. The speakers were asked to tell about their activities of the last week. The data are transcribed according to the Verbmobil transliteration standard ([Burger, 1997]).
Original Purpose of Recording:
Empiric investigations of dialectal variation within Standard German.
Other Usabilities:
General ASR; dialog systems; research of elicited spontaneous speech; prosody; phonetic investigations.

Non-German Corpora

Although BAS is dedicated to the spoken German language there exist a few speech corpora of other languages mostly linked to some German BAS resources.

American VM I
Type: dialogues (American English and 'Denglish')
Format: 16 Bit / 16 kHz / PhonDat
Environment: office
Recording sites: 3
Speakers: 256
Transcript: VM I + VM II transliteration
Segmentation: none
Total: 4029 utterances (average 27.5 words)
Medium: 3 CDROMs
Description:
This corpus contains Verbmobil I style recordings done at Carnegie Mellon University, USA, University of Karlsruhe and Bonn University, Germany. The vast majority of the recordings are done with native American speakers; a small subcorpus was spoken by native Germans with average knowledge of the English language ('Denglish'). *Original Purpose of Recording:*
ASR for online translation German to English / Japanese.
Other Usabilities:
General ASR; dialog systems; research of elicited spontaneous speech; prosody; foreign accents; phonetic investigations.

Japanese VM I
Type: dialogues
Format: 16 Bit / 16 kHz / RAW (compressed)
Environment: office
Recording sites: 1
Transcript: adapted VM I transliteration
Segmentation: none
Total: 800 dialogues
Medium: 4 CDROMs
Description:
This corpus contains Verbmobil I style recordings done at Tokyo University, Japan. This corpus has not been validated by BAS and is distributed 'as is'.
Original Purpose of Recording:
ASR for online translation German to English / Japanese.
Other Usabilities:
General ASR; dialog systems; research of elicited spontaneous speech; prosody; phonetic investigations.

Other Resources

EMA1
Type: screen prompted speech
Format: Audio: 16 Bit / 16 kHz. EMA: 6+2 sensors, X,Y + velocity + tilt, 250 Hz / 16 Bit
Environment: studio
Recording sites: 1
Speakers: 7 (6 male, 1 female)
Transcript: Orthographic

Segmentation: phonemic, vowel onsets, context consonants, etc.
Total: 3906 utterances
Medium: 1 CDROM
Description:
The corpus contains recording of the movement of the main articulators in the mid-sagittal plane together with the speech signal. The data include the speech signal, X/Y position, X/Y velocity and tilt factor (reliability) of 6 sensors and 2 reference sensors. Four fifth of the text corpus consists of carrier phrases with all German vowels embedded into changing consonantal context spoken with normal and fast speed (2 x 225); one fifth (108) consists of real sentences with the same vowels contained.
Original Purpose of Recording:
Investigation of the vowel production in German.
Other Usabilities:
Modelling of the vocal tract; improving ASR with articulatory parameters; phonetic investigations.

PHONOLEX
Type: Pronunciation Dictionary for German
Format: ASCII
Total: approx. 650.000 entries
Medium: 1 CDROM or FTP
Description:
The PHONOLEX dictionary contains a fully inflected list of the most common German words together with their canonical pronunciation in SAM-PA.
Original Purpose of Recording:
Lexicon lookup for automatic phonemic segmentation with MAUS.
Other Usabilities:
ASR; Speech Synthesis.

PHONRUL
Type: Pronunciation Rule Set
Format: ASCII
Total: approx. 5.000 rules
Medium: Floppy Disk or FTP
Description:
PHONRUL is a collection of simple re-write rules for German pronunciation. Starting with a canonical representation of the utterance in SAM-PA the rule set can be used to create the most likely pronunciation variants expected in Standard German (no dialectal variation).
Original Purpose of Recording:
Calculating pronunciation hypothesis for automatic phonemic segmentation with MAUS.
Other Usabilities:
ASR; Speech Synthesis.

Future Resources at BAS

The following resources are currently produced at BAS and won't be available for at least one year starting with April 1998.

SpeechDat
German SpeechDat(M) (1000 speakers) is now available
at the *European Language Resources Agency* (ELRA) (http://www.icp.grenet.fr/ELRA/home.html). The German part of SpeechDat(II) won't be available for others than members of the consortium until May 2000. The data will be distributed via the ELRA.

Speech in the running car
Currently three corpora with recordings in the running car are or have been produced at the Department of Phonetics, University of Munich:
CSDC2: 238 speakers, 50 utterances, available not before June 1999.
CSDC1: 155 speakers, 92 utterances, available not before June 2000.
CSDC4: 105 speakers, 154 utterances, available not before May 2003.

For more details about BAS resources and ordering information please refer to our WWW documentation:
www.phonetik.uni-muenchen.de/Bas

References

[Burger, 1998] Susanne Burger (1998). RVG1 - A Database for Regional Variants of Contemporary German. Proceedings of this Conference, Granada 1998.

[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)

[Draxler, 1996] Draxler, Chr. (1996). The German SpeechDat Telephone Speech Corpus - Overview and Experiences. Speech Science and Technology 1996 Conference, Adelaide, Australia.

[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.

[Kipp et al., 1996] A. Kipp, M.-B. Wesenick, F. Schiel (1996). Automatic Detection and Segmentation of

Pronunciation Variants in German Speech Corpora. Proceedings of the ICSLP 1996. Philadelphia, pp. 106-109, Oct 1996.

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

[Schiel et al, 1997] F. Schiel, Ch. Draxler, H.G. Tillmann (1997). The Bavarian Archive for Speech Signals: Resources for the Speech Community. Proceedings of the EUROSPEECH 1997, Rhodos, Greece, pp. 1687-1690.

[Schiel, 1997] F. Schiel (1997). Probabilistic analysis of pronunciation with MAUS. The ELRA Newsletter, December 1997, pp. 6-9.

[Tillmann et al, 1995] 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 Stockholm, pp. 4:550-553.