BAVARIAN ARCHIVE FOR SPEECH SIGNALS (BAS) STATUS REPORT 2000 – 2001

Florian Schiel, Christoph Draxler, Phil Hoole

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

Outline of this Report

The *Bavarian Archive for Speech Signals* (BAS) is a joint initiative of the Bavarian State and the Ludwig Maximilians Universität München. It is located at the host organisation *Institut für Phonetik und Sprachliche Kommunikation* (IPSK) and collects, evaluates, produces and disseminates speech based resources to the scientific community. Our focus is the German language covering a large geographical part of central Europe.
This short report covers BAS related activities in the period of Mar 2000 to Dec 2001 and is strictly confidential. The purpose of the report is to update the members of the international scientific review commitee as well as the members of the steering committee about recent BAS activities.
The report is organised as follows.
The first part deals with general developments that are crucial for the future work of the BAS. Also, rough financial figures are given here.
The second part lists projects and/or cooperations that have been finished, conducted or started during the reporting period. Furthermore, associated projects where BAS is not directly funded but agreed to act as a focal point or distribution/evaluation center are given here.
The third section describes changes about the administration and dissemination methods currently used by BAS. We also will give numbers about how many resources were distributed in the reporting period to how many users as well as details about our cooperation with the European Language Resources Association (ELRA).
Section 4 is reserved for minor BAS activities and contacts.
In appendix A you will find a concise description of all German speech and speech-related resources that are currently available at BAS.

1 General Developments

Demands
The last two years have shown an increasing demand for specialized speech resources. At BAS the following types of data were finished or started to produce within the reporting period

- multi-modal data: synchronized recordings of voice and body/face movenments
- biometrical data: bench-mark data for voice verification synchronized with other biometrical data such as form of hands, signature, fingerprint.
- canonical/most-likely pronunciation of proper names
- data recorded under 'real-life' or 'wizard-of-Oz' conditions
- bi-lingual speech corpora, e.g. a dialog

between two speakers of different L1.

Funding of resources
Main funding for the production of multimodal data stems from the SmartKom project funded by the Ministry of Education and Science. 7-9 members of the permanent staff and 20-28 part-time students are assigned to SmartKom at BAS.
SpeechDat related projects (mostly subcontracts) continued throughout 2000 and the first half of 2001. Some minor speech resources were produced in close connection to these activities for industrial partners in the consortia.
Two major industrial funded projects concern a corpus for bench-marking speaker verification over the telephone network (T-NOVA) and the speech of kids to control electronic devices (PHILIPS). Public parts of these corpora will be distributed via the BAS after the contracted blocking periods.
A total funding of approx. EUR 1.537.054,- was used within the reporting period for the direct or indirect production of speech resources.

Funding of permanent personnel
Florian Schiel is still acting as a executive director of BAS and is currently funded by the University of Munich. Christoph Draxler as well as our part-time secretary Almuth Preugschat and part-time system operator Klaus Jänsch were funded by the royalties for distributed language resources. Currently Christoph is acting as a replacement for a vacant professoral position at Munich University and will therefore be funded by the university for at least 6 months. Phil Hoole is a member of the permanent staff of our hosting institution IPSK and continues to act as a link to the fundamental scientific activities.

Royalties
Incoming royalties and fees for CDROM production and dissemination as well as minor contracts was approx. EUR 57.712,- in 2000 and EUR 152.146,- in 2001.
Spending (including payment of royalties to external copyright holders, personell and infrastructure) was approx. EUR 46.581,- in 2000 and EUR 116.279,- in 2001.

Other activities
For the first time BAS acted as a validation center for an external corpus, the CGN Dutch corpus initiative. The validation took place in Oct 2001 and resulted in a very good overall evaluation of the work in the Netherlands.

Coming up next
During 2000 and 2001 we were applying for a major funding of BAS activities at the Ministry of Education and Sciences. Due to the success story of BAS within several projects and due to our close cooperation with industry this project called 'BAS: Infrastructures for Technical Speech processing' (BITS) was finally approved in Dez 2001. The funding will start March 2002 for the follwing 4 years and will extend the staff of BAS by another 5 permanent and 10-15 part-time positions. The aims of this funding are

- the compilation of "cook books" for LR production and evaluation
- the development of Web-based data gathering and annotation tools
- a re-evaluation of all BAS LRs
- a survey of the actual and upcoming needs of the speech and language industry for the next 5, 10 and 15 years
- the production of a freely available, general purpose speech synthesis corpus
- the production of a freely available corpus with kids voices

2 Cooperations / Projects

SMARTKOM – DFKI, Siemens, Philips, Sony, Daimler/Crysler, EML, Media Interface, University of Erlangen, University of Stuttgart

SmartKom is one of the five leading projects within the funding frame work "Man Technique Interaction" (MTI) supervised and funded by the Ministry of Education and Sciences in Germany and a consortium of 8 German industrial partners. SmartKom aims at the development of basic knowledge of how humans may interact with machines in a 'natural way' using not only natural speech input but also gestural input and even input derived from the facial expression of the user. To exemplify the findings in several areas of man-machine communication three prototypes using the same basic processing core will be developed and demonstrated to the public: SmartKom Public, a public accessible information booth, SmartKom Home, an intelligent assistant at home, and SmartKom Mobil, an intelligent guide to be used in the moving vehicle or even as a PDA.

BAS is involved in the empirical investigations of user interaction using Wizard-of-Oz techniques and in the end-to-end evaluation of the prototypes. As a valuable by-product of the former a unique corpus of synchronized audio and video streams is being created that will be disseminated to the scientific community for free after a blocking period of one year. The first release is scheduled for July 2002.

The SmartKom corpus contains the voice signal of the user (7 channels), the system and the backround noise, the video signal of the face, the upper body from the left, the graphical display of the system and the infrared video recording of the display area (to capture hand gestures).

More details can be found in http://smartkom.dfki.de.

RVG

In March 1995 BAS started a long-term collection financed by AT&T Germany (later by Lucent Technologies and AT&T Labs) called 'Regional Variants of German' (RVG). The aim of this project was to collect speech of speakers originating from all German speaking areas in Europe (that is Germany, Austria, Italy and Switzerland). The speech samples were collected 'in-field' with 4 different microphones into a standard PC. The read speech consists of application oriented commands and number items (telephone, banking, dates, etc), phonetically balanced sentences and 1 minute of spontaneous monologue. More details can be found in appendix A, section RVG1.

The first phase of the project ended in 1998 with 500 recorded speakers. The first two microphone channels (low quality) were published after one year of blocking period. The other two channels were recorded on high quality DAT tapes and were processed during 1999. The latter was solely funded by the royalties earned by distributing the RVG corpus. The high quality channels have been distributed for the first time in March 2000.

In Dec 2001 a major update of the total corpus was finished and distributed for all users of the corpus for free. Special thanks go to Ericson Nuernberg for their valuable input to this update.

Beginning of 2001 BAS startet a new extension of the RVG corpus called RVGJ containing the speech samples of children and young persons. This corpus will be probably extended by more recordings within the BITS project in the coming years.

SpeechDat - EU and industry

In the SpeechDat projects, large databases for voice-operated teleservices are created according to common specifications for many European (and now also non-European) languages.

In **SpeechDat-Car**, BAS was responsible for the collection of the German database under a contract with Robert Bosch GmbH and BMW AG, Germany. In this project, 600 sessions have been recorded in nine languages; the recordings are carried out both in a car and synchronously via a GSM phone. In the car, four high-bandwidth channels are recorded; the vocabulary consists of application words and phrases for vehicle control, teleservices, and telecommunication commands (60%), and the standard SpeechDat material (40%) [11]. Finally, the BAS maintains the SpeechDat WWW server: www.speechdat.org from which all SpeechDat projects can be accessed.

SENECA - industry

For their SENECA project, Robert Bosch GmbH asked BAS to record German spellings of geographical names for navigation systems. These recordings have been performed using the SpeechDat-Car recording platform for 30 speakers with approximately 80 recordings each. Presently it is not clear whether this material will be published by BAS.

Verbmobil I & II - Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMB&F)

Immediately after its foundation in 1995 BAS agreed to act as an informal partner to the **Verbmobil I** consortium in respect to speech resources resulting from this project. The main task for BAS was (and still is) to ensure a proper publication and dissemination of empirical data to the speech science community after the blocking period of one year.
Verbmobil II started in 1997 with a reduced consortium. BAS agreed to continue its services to the end of the project in 2000.
Throughout the last 3 months of 2000 and the first 6 months of 2001 BAS re-structured all deliveries of the Verbmobil partners and created a final edition of the combined Verbmobil corpus (60 CDROMs) which was edited in Jun 2001 (see details in appendix A, section VM).

VeriDat - German Telekom, T-Nova

The VeriDat project started in Nov 1999 and aims to produce the first very large speaker verification database via public phone lines for German. 150 speakers are recorded in 20 different sessions with 40 read items each. The environment and type of phone line is controlled as well as a concise speaker profile.
The corpus consists of two parts:

- VERIF1DE, a SpeechDat-II compliant subset of 21 items per session for 150 speakers, and
- VERIDAT, the full corpus.

Work on both corpora has terminated in October 2001. VERIF1DE has gone through the same external validation procedure as any other SpeechDat-II corpus. The validation was performed by SPEX in Nijmegen, NL, and the validation results have been excellent.
VERIF1DE will be published by T-Nova, with the actual distribution taken care of by BAS. For VERIDAT, the date of publication is not yet known, but T-Nova has declared its willingness to publish the corpus in the near future.

TAXI - DFKI

In May 2001 the TAXI speech collection containing a multi-lingual field trial was recorded and validated in close cooperation with DFKI in Saarbrücken.
TAXI contains 94 recorded dialogues between a cab dispatcher and a client recorded over public phone lines (network and GSM). The dispatcher always spoke German, while the clients always spoke English. The corpus will be published by BAS in June 2002. More details can be found in appendiy A, section TAXI.

3 Archive and Dissemination

Basic Archive – Changes

The former storage cooperation with the "Leibniz Rechenzentrum" (LRZ) ended beginning of 2000, because storage was getting affordable for BAS in greater quantities. Therefore the whole storage mangement of the master volumes moved locally to the BAS premises.
To cover the huge multi-modal corpora we added another 300 GB cache in 2000 and 1600 GB cache end of 2001. All storage facilities of BAS are still backuped to the very large Tivoli system at LRZ.
These changes enable us for the first time to keep all BAS corpora online for the benefit of researchers at IPSK and neighbor institutions as well as for an easier version management and media production. Furthermore, BAS also offers now corpora on DVD-5 media instead of on multiple CDROMs. For this

purpose three DVD-5 burners were installed at BAS (thanks go to DFKI who generously lent BAS one of their burners). However, the acceptance of DVD-5 is still very low compared to traditional CDROMs.
For security reasons the BAS servers are not visible in the Internet and only accessible from local and authorized workstations. We do not intend to start the dissemination of signal files via the Internet in the near future.

Distribution in Numbers

Within the reporting period approx. 536 CDROMs containing 66 different resources were distributed to 47 users in the scientific community. Most of the users ordered more than one resource from BAS.
This results in 268 CDROMs per annum compared to the last reporting period (1995 - 1999) where BAS distributed only 40 CDROMs per annum. Also, the number of interested parties per annum continues to increase (from 17 to 23 per annum).
These numbers are caused by two facts: the size of resources is constantly growing and the number of companies working on speech is increasing as well.

European Language Resources Association (ELRA)

The EU funded European Language Resources Association together with its distribution agency ELDA have signed a contract with BAS that regulates cooperation and brokership procedures between BAS and ELDA. BAS entitles ELDA to list all of its resources in the official ELRA catalogue and acts as a broker between BAS and the end user. Until July 1998 the commission for ELDA was 10% of the published basic BAS CDROM fee and 0% on additional license fees (claimed by the copyright holder).
Since 1st of Aug 1998 the commission was raised to 30% for the basic CDROM fee and to 20% on the additional licence fees. At the same time the basic CDROM fee was raised to EUR 237,- to compensate the growing production costs at BAS.
In the reporting period 40.1% of all BAS deliveries were mediated by ELRA.

4 Miscellaneous

Re-validation

BAS is continuously re-validating the existing corpora in the archive. The long-term aim of these activities is to achieve a common technical format of signals, annotations and meta information throughout the different resources in BAS. So far, a number of existing resources have been adapted to the NIST SPHERE technical format and augmented by BAS Partitur Files. We hope to intensify these activities within the BITS project (see above).

BAS Web Services

The BAS maintains several Web servers for documentation and data retrieval. The main server address is: www.bas.uni-muenchen.de/Bas where you will find information about the currently available speech resources, on-going projects as well as documentation of file formats, etc.
Currently, approx. 200 documents are maintained on the primary server alone. Other Web servers are used for the offline annotation of speech corpora using the software tool WWWTranscribe developed by Chr. Draxler.
In Dec 2001 all financial information on the Web server was converted to Euro.

Special event at EUROSPEECH 2001

BAS was responsible for organizing the Eurospeech Special Event on *Existing and Future Corpora - Acoustic, Linguistic, and Multi-Modal Requirements* at Eurospeech 2001 in Aalborg, DK.
At this workshop, 8 plenary presentations were given which gave an overview of ongoing projects and novel approaches to making accessible large collections of language resources. The sessions were well attended and the discussion was lively.
The best student paper award of the conference went to a presentation of this ESE

(Shawn Chang and Mirjam Wester from Steve Greenberg's group at ICSI).
It is planned to organize similar events at future speech related conferences.

For other publications and conference papers within the reporting period please refer to the reference list in this report. Most papers may be downloaded from the following URL:
www.phonetik.uni-muenchen.de/Publications/
If you prefer a hard copy, please send a request to bas@bas.uni-muenchen.de together with your postal address or fax number.

References

[Beringer N., Schiel F. , 2000] The Quality of Multilingual Automatic Segmentation Using German MAUS. Proc. of the International Conference on Spoken Language Processing, Beijing, China.

[Burger, S. et al, 2000] Verbmobil Data Collection and Annotation. in Verbmobil: Foundations of Speech-to-Speech Translation (Ed. Wahlster, W.); Springer; Berlin/Heidelberg.

[Draxler, Chr. et al, 2001] Speaking While Driving - Preliminary Results on Spellings in the German SpeechDat-Car Database. Proc. of the Eurospeech 2001 Conference. Aalborg, Danmark.

[Draxler, Chr. 2001] Sprachdatenbanken. Computerlinguistik und Sprachtechnologie. Hrsg. Carstensen, K.-U. et al. Spektrum Verlag. Heidelberg, pp. 402-409.

[Kurematsu, A. et al, 2000] : VERBMOBIL Dialogues: Multifaced Analysis. Proc. of the International Conference on Spoken Language Processing, Beijing, China.

[Oppermann, D. et al, 2000] What are transcription errors and why are they made? Proc. of the Second International Conference on Language Resources and Evaluation. Athens, Greece.

[Steininger, S. 2000] Transliteration of Language and Labeling of Emotion and Gestures in SMARTKOM. Workshop Proc. of the Second International Conference on Language Resources and Evaluation: Meta-Descriptions and Annotation Schemes for Multimodal/Multimedia Language Resources. Athens, Greece, pp. 49-51.

[Weilhammer, K. et al, 2000]
The influence of scenario constraints on the spontaneity of speech. A comparison of dialogue corpora. Proc. of the Second International Conference on Language Resources and Evaluation. Athens, Greece.

[Beringer N. , 2001]
Evoking Gestures in SmartKom - Design of the Graphical User Interface. Gesture Workshop 2001, London, UK.

[Beringer N. et al, 2001] Possible Lexical Indicators for Barge-In / Barge-Before in a multimodal Man-Machine-Communication. International Workshop on Information Presentation and Natural Multimodal Dialogue,Verona, Italy, 14-15 December 2001.

[Oppermann, D. et al, 2001] Off-Talk - A Problem for Human-Machine-Interaction. Proc. of EUROSPEECH 2001, Scandinavia, Aalborg.

[Siepmann, R. et al, 2001]
Using Prosodic Features to Characterize Off-Talk in Human-Computer Interaction. Proceedings of the ISCA Workshop on Prosody in Speech Recognition and Understanding, Red Bank NJ, October 22-24.

[Steininger, S. et al, 2001] Labeling of Gestures in SmartKom - The Coding System. Gesture Workshop 2001, London, UK.

[Steininger, S. et al, 2001] Gestures
During Overlapping Speech in multimodal Human-Machine Dialogues. International Workshop on Information Presentation and Natural Multimodal Dialogue 2001, Verona, Italy.

[Steininger, S. , 2001]
Human-Computer Communication - Wizard Of Oz-Experiments In SmartKom. 5. Fachtagung der Deutschen Gesellschaft fr Kognitionswissenschaft 2001, Leipzig.

A 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.

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.
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.

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.

SC10 - Accents II

Type: read speech
Format: 16 Bit / 16 kHz / NIST
Environment: studio
Recording sites: 1
Speakers: 70
Transcript: orthographic
Segmentation: phonemic
Total: read speech: 4474, monologue: 1168, dialog: 1303
Medium: 2 CDROMs
Description:
The corpus contains read and non-prompted German and mother tongue speech of 70 different speakers from 17 mother languages (L1) in a variety of speaking styles. *Original Purpose of Recording:*
Scientific investigation of foreign accents. *Other Usabilities:*
Automatic accent detection; adaptation to foreign accents; robust ASR; forensic applications; speaker verification.

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 / NIST
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: 498
Transcript: orthographic (with linguistic markers and noise class)
Segmentation: none
Total: 42000 utterances (average 6 words)
Medium: 30 CDROMs / 5 DVD
Description:
The *Regional Variants of German* (RVG1) corpus contains 85 screen prompted utterances (digits, phone numbers, computer command phrases, phonetically rich sentences) randomly selected from a group of larger text corpora. The 498 speakers were selected according to demographic densities in Germany, Austria, parts of Switzerland and Italy. Speakers were asked to speak informally but not dialectally. Recordings were done with 4 different microphones (from high quality studio mi to low cost desk top mic). *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) 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 / NIST
Environment: studio
Recording sites: 1
Speakers: 101
Transcript: orthographic
Segmentation: none
Total: 10100 utterances (average 23.4 words)
Medium: 7 CDROMs / 1 DVD
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 *unscripted 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 / NIST
Environment: studio/office
Recording sites: 4
Speakers: 779
Transcript: VM I + VM II transliteration
Segmentation:
phonemic,word,prosodic,dialogact
Total: 13910 turns (average 22.8 words)
Medium: 9 CDROMs / 2 DVDs
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.

German VM II
Type: dialogues
Format: 16 Bit / 16 kHz / NIST
Environment: office
Recording sites: 2
Speakers: 448
Transcript: VM II transliteration
Segmentation:
phonemic,word,prosodic,dialogact
Total: 58961 turns
Medium: 45 CDROMs / 4 DVDs
Description:
In contrast to Verbmobil I the scenarios of the recorded situations were extended and the format was re-defined for a better parseability as well as a better handling of the English and Japanese parts of the corpus. Furthermore, the speaker and recording database was redefined and standardized for all data. Pronunciation lexica for the three languages, phonemic segmentations of the German part and other linguistic resources (such as dialog act labeling, prosodic labeling, tree banks, parts of speech tagging) have been included into the final corpus. Also, emotional data, the VM lexicon database and other bonus material is now available as part of the corpus.
Original Purpose of Recording:
ASR for online translation German to English / Japanese.
Other Usabilities:
General ASR; dialog systems; research of elicited spontaneous speech; prosody; emotion in speech; phonetic investigations.

RVG 1
Type: monologues
Format: 16 Bit / 22.05 kHz / NIST
Environment: office
Recording sites: 6
Speakers: 500
Transcript: VM II transliteration
Segmentation: none
Total: 500 x 1 minute monologue
Medium: 1 CDROM
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 talk about their activities of the last week. The data are transcribed according to the Verbmobil transliteration standard.
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.

Bi-lingual Corpora

These corpora mostly aim at the development of automatic speech translation systems like Verbmobil.

VMII German - American
Type: multi-lingual dialogues (American English and German)
Format: 16 Bit / 16 kHz / NIST
Environment: office
Recording sites: 1
Speakers: 256
Transcript: VM II transliteration
Segmentation: none
Total: utterances
Medium: 6 CDROMs
Description:
Multi-lingual dialogue recording where each partner spoke in his L1 language with(1)/without(5) interpreter.
Original Purpose of Recording:
ASR for online translation German to English
Other Usabilities:
General ASR; dialog systems; research of elicited spontaneous speech; prosody; foreign accents; phonetic investigations.

VM II German - Japanese
Type: dialogues (Japanese and German)
Format: 16 Bit / 16 kHz / NIST
Environment: office
Recording sites: 1
Speakers:
Transcript: VM II transliteration
Segmentation: none
Total: utterances
Medium: 4 CDROMs
Description:
Multi-lingual dialogue recording where each partner spoke in his L1 language with(3)/without(1) two interpreters.
Original Purpose of Recording:
ASR for online translation German to Japanese
Other Usabilities:
General ASR; dialog systems; research of elicited spontaneous speech; prosody; foreign accents; 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 / NIST
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.

American VM II

Type: dialogues (American English)
Format: 16 Bit / 16 kHz / NIST
Environment: office
Recording sites: 1
Speakers: not known yet
Transcript: VM II transliteration
Segmentation: none
Total: not known yet
Medium: 6 CDROMs
Description:
This corpus contains Verbmobil II style recordings done at Carnegie Mellon University, USA.
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 / NIST
Environment: office
Recording sites: 1
Transcript: VM II transliteration
Segmentation: none
Total: 800 dialogues
Medium: 8 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.

Japanese VM II

Type: dialogues
Format: 16 Bit / 16 kHz / NIST
Environment: office
Recording sites: 1
Transcript: VM II transliteration
Segmentation: none
Total: not known yet
Medium: 11 CDROMs
Description:
This corpus contains Verbmobil II style recordings done at Tokyo University, Japan.
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

VM Emotion

Type: dialogue with WOZ
Format: Audio: 16 Bit / 16 kHz
Environment: studio
Recording sites: 1
Speakers: 61
Transcript: Orthographic
Segmentation: emotions
Medium: 3 CDROM
Description:
The corpus contains recording of WOZ experiments to elicite emotions in users of a speech dialogue system. *Original Purpose of Recording:*
Recognition of certain emotions from speech.
Other Usabilities:
Investigations of emotional speech.

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 midsagittal 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. 1.600.000 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 editions at BAS

The following resources are currently produced at BAS and will be available in the near future.

SC2 - noises

The corpus contains read speech of 10 different speakers with screen prompted 'automobil diagnosis phrases' recorded under real conditions in two different car maintenance halls. The language is German. All speakers are male native Germans and have never participated in such a task before. They are all experts in the field of car diagnosis. Each speaker has spoken 800 3-7 word utterances derived from 100 different sentences resulting in a total of 8000 utterances. Release mid of 2002.

SC3 - phone noises

The corpus contains spontaneous speech of 76 different speakers recorded by an automated phone service system under real life conditions. The language is German. The utterances were elicited by questions that the speaker had to answer unprepared (e.g. 'What is your home address?'). Each speaker has spoken 26 utterances answering 26 questions by the automated server resulting in a total of 1998 utterances. Release mid of 2002.

ZipTel

Part of the German SpeechDat(M) corpus (1000 speakers) that is not included into the official SpeechDat data set. Contains phone number, ZIP codes and German street names read via telephone line by approx. 1000 speakers. Release end of 2002.

FormTask
Part of the German SpeechDat(II) corpus (4000 speakers) that is not included into the official SpeechDat data set. Contains the answers of each speaker to 7 questions about a form in the data sheet. Release end of 2002.

Hempel's Sofa
Part of the German SpeechDat(II) corpus (4000 speakers) that is not included into the official SpeechDat data set. Contains max. 60 sec of spontaneous monolue by each speaker Release end of 2002.

VeriDat
Probably in 2002 or 2003 (still ongoing negotiations about the release date).

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 2002.
CSDC1: 155 speakers, 92 utterances, available not before June 2002.
CSDC4: 105 speakers, 154 utterances, available not before May 2003.

TAXI
Will be released in Jun 2002 (see section 2).

For more details about BAS resources and ordering information please refer to our WWW documentation:

www.phonetik.uni-muenchen.de/Bas