The Bavarian Archive for Speech Signals – Serving the Speech Community

H.G. Tillmann, F. Schiel, Chr. Draxler, Ph. Hoole

Bavarian Archive for Speech Signals (BAS)
University of Munich, Germany

ABSTRACT

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* and it 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 paper gives an concise overview about the BAS activities during the first 5 years of its existence.

1. INTRODUCTION

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 currently 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*.

The second part after the general introduction deals with our experiences during the last five years to produce new, highly re-usable speech resources in close cooperation with industrial partners. We will explain our BAS policies that ensure 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 highlight these points.
The third part lists a selection of projects and/or cooperations that have been conducted or started during 1995 to 2000. Furthermore, associated projects where BAS is not directly funded but agreed to act as a focal point or distribution/evaluation center are given here.

Section 4 gives a brief overview of some special scientific multi-modal data collected at BAS.

2. BAS POLICIES

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 than 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 sole purpose of a single project or product will not be very useful for other tasks in the future. In other words: there always will be the trade-off between expenses and re-usability.

Our work at BAS during the last 5 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. 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 M* 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 in 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 by Speech Engineering for highly specialised speech corpora. Examples are:

- speech under real life noise conditions
- speech in the running car
- speech for specific dictation domains
- speech of children
- speech with special microphones / headsets / built-in-mics
- dialectal speech
- 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 which 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.

3. COOPERATIONS / PROJECTS

RVG1

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. Mode details can be found in appendix A, section RVG 1.
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 will be distributed for the first time in April 2000.
The industrial partners aggreed to use their license royalties from this corpus to process and publish the corpus via BAS.

SpeechDat

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.

SpeechDat(M) defined a common core of approx. 45 items: digits, numbers, digit strings, spellings, command words and phrases, and phonetically rich material; all speech is recorded via the telephone. Eight databases of 1000 speakers each were collected for Danish, English, French, German, Italian, Portuguese, Spanish, and Swiss French; additionally, a database of 300 Italian speakers via the mobile phone was recorded.

In **SpeechDat(II)**, the BAS was a subcontractor to SIEMENS AG for the collection of 4000 German speakers via the fixed telephone network, and to Vocalis Ltd, UK for 1000 speakers via the mobile telephone network. Both the fixed and the mobile network database were annotated by BAS; additionally, 500 German speakers were annotated under a contract by Lernout & Hauspie, Belgium.

In **SpeechDat-Car**, BAS is responsible for the collection of the German database under a contract with Robert Bosch GmbH and BMW AG, Germany. In this project, 600 sessions will be 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%) ([Draxler, 1996]).

Phonolex

Since nowadays SLP application in German are still designed on a word based structure, there is always a general need for a very large pronunciation dictionary of contemporary German. In 1996 BAS started an initiative called PHONOLEX to produce a very large (500.000 and up) reproducable pronunciation dictionary for this purpose. As a first partner the 'Deutsches Forschungszentrum für künstliche Intelligenz' (DFKI) joined the initiative and de-

livered a first fully inflected base list of German (650.000 entries). At BAS a text-to-phoneme system kindly provided by the University of Bonn was ported to C and adapted to the PHONOLEX task. A first version (1.4) of PHONOLEX was published end of 1996. In 1998 the University of Leipzig joined the group and added an empirically based word list of contemporary written German of 1.000.000 entries to the project. The list in its current version comprises over 1.6 Million entries.

SI1000P

In 1996 this corpus was recorded in cooperation with Siemens AG ZT, Munich. The aim of this project was to produce high quality speech samples for PSOLA-like concatenative speech synthesis. Two professional speakers read the sentence list of the SI1000 corpus (newspaper texts) in a echo cancelled environment. The speech signal as well as the laryngographic signal were recorded. The latter was analysed for the location glottal pulses in the time stream. Furthermore the recordings of one speaker were labelled prosodically, that is boundary markers B3, B2 and B9 (GTobi) and sentences stress (primary, secondary and contrastive) were marked in the transcript of the recording.

Verbmobil I & II

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 empiric data to the speech science community after the blocking period of one year. Furthermore, BAS re-edits all Verbmobil speech corpora after one year and includes all segmental or other symbolic information available over the time of the project. BAS also acts as a contact point to the European ELRA in Paris with regard to Verbmobil resources.
Verbmobil II started in 1997 with a reduced consortium. BAS aggreed to continue its services to the end of the project in 2000. Currently (Jul 2000) Verbmobil II is in its final phase.
BAS continues to maintain the freely available parts of the collected data. 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) will be included in the final corpus.

SmartKom

The SmartKom project started in Sept 1999 and aims to develop new intelligent human - machine interfaces in the area of personal computing. The IPSK is one of 10 industrial and academic partners of the SmartKom consortium directed by the Deutsches Forschungszentrum für künstliche Intelligenz (DFKI) in Saarbrücken, Germany. All empiric investigations and data collection will be carried out at IPSK in Munich. The BAS agreed to maintain and disseminate the collected multi-modal data to the scientific community after a blocking period of one year.
The multi-modal data consist of audio recordings (10 channels), video streams (3 channels), gesture (pointing) recognition and graphic tableau output. The target interface of SmartKom combines multi-channel speech input under severe conditions (public, car, home environment), hand gestures (2-dimensional) and facial expressions with an intelligent dialogue handling and multi-modal output channels.

VeriDat

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.
Since the database will be SpeechDat compatible (that is, compatible to the SpeechDat specifications regarding speaker verification data bases), the BAS in concerned with the evaluation and validation of the corpus. More detailed, BAS is responsible for the pre-validation test, the transcription and validation of all recordings and for general consulting to ensure SpeechDat quality. The final validation of the corpus will be carried out by SPEX (Netherlands) to ensure an independent objective test set.

4. MULTI-MODAL DATA

Three key categories of articulatory data have been targeted for inclusion in BAS: Measurements of fleshpoint data, NMRI data of the vocal tract, digitized video.

Fleshpoint data

The first set of data available in this category was originally acquired for a project on vowel articulation in German. Movement data was acquired by means of electromagnetic midsagittal articulography for lower lip, jaw and four points on the tongue (see Fig. 1). High quality synchronized audio data was also acquired (16bit, 16kHz). Seven speakers spoke the following corpora:
a) a pseudo-word corpus (in a carrier phrase) of multiple repetitions of the target vowels in three consonant contexts (/p,t,k/). The corpus was recorded at normal and at fast speech rates.
b) a corpus of 105 meaningful sentences (approx. 15 syllables each) containing each target vowel in 15 different contexts. For all corpora a manual segmentation and labeling of the target vowels is available. In addition, for the sentence material a MAUS-based segmentation ([Kipp et al, 1997]) of the complete speech material has been performed. Further details of recording techniques and articulatory pre-processing can be found in [Hoole, 1996]. As an example of analysis results, Fig. 2 shows a speaker-independent articulatory representation of

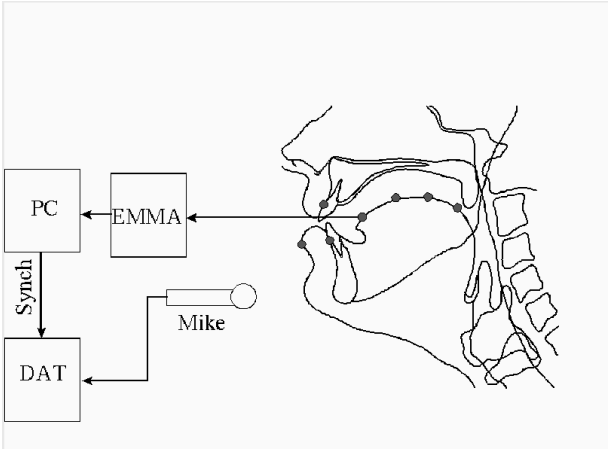

Figure 1: Typical arrangement of sensors for EMMA experiment

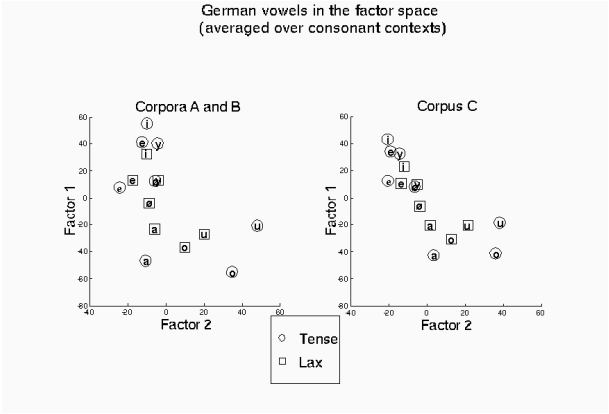

Figure 2: Distribution of German vowels in the factor space determined from PARAFAC analysis of tongue configurations. Corpora A and B are the pseudo-word corpora (normal, fast). Corpus C is the sentence corpus.

the German vowel space based on PARAFAC factor analysis of the recorded tongue configurations (further details in [Hoole, 1998]).

NMRI data

As part of an ongoing study of alveolar consonant production, and as a supplement to the earlier vowel project just mentioned (wherever possible with the same speakers), NMRI scans have been carried out on, to date, 6 speakers producing the consonants /s/, /SH/, /l/, /n/, /t/, and the vowels /a/, /e/, /i/, /o/, /u/, /y/, /oe/. For each sound 23 slices in each of 3 planes (coronal, axial, sagittal) were collected using a T1-weighted FLASH sequence.

References

[Bird & Liberman, 1999] Steven Bird and Mark Liberman (1999). A Formal Framework for Linguistic Annotation Technical Report MS-CIS-99-01 - Department of Computer and Information Science, University of Pennsylvania (expanded from version presented at ICSLP-98, Sydney).

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

[Draxler, 1997] Draxler, Chr. (1997). WWWTranscribe - A Modular Transcription System Based on the World Wide Web; Proc. of Eurospeech 97 Conference, Rhodos, 1997.

[Draxler, 1998] Draxler, Chr. (1998). WWWSigTranscribe - A Java Extension of the WWW Transcribe Toolbox; Proc. of 1st Int'l Conference on Language Resources and Evaluation (LREC), Granada, 1998.

[Hoole, 1996] P. Hoole (1996). Issues in the acquisition, processing, reduction and parameterization of articulographic data. in: FIPKM, 34, 158-173.

[Hoole, 1998] P. Hoole (1998). Modelling tongue configuration in German vowel production. Proc. 5th Int. Conf. Spoken Lang. Processing, 5, pp. 1863-1866.

[Kipp et al, 1997] A. Kipp, M.-B. Wesenick, F. Schiel (1997). Pronunciation Modeling Applied to Automatic Segmentation of Spontaneous Speech. in: Proceedings of the EUROSPEECH, Sept 1997, Rhodos, Greece, pp. 1023-1026. (1997). The Bavarian Archive for Speech Signals: Resources for the Speech Community. Proceedings of the EUROSPEECH 1997, Rhodos, Greece, pp. 1687-1690.

[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.

[Schiel, 1998] F. Schiel (1998). Speech and Speech-Related Resources at BAS; Proceedings of the First International Conference on Language Resources and Evaluation 1998, Granada, Spain, pp. 343-349.

[Schiel et al, 1998] F. Schiel, S. Burger, A. Geumann, K. Weilhammer (1998). The Partitur Format at BAS; Proceedings of the First International Conference on Language Resources and Evaluation 1998, pp. 1295-1301, Granada, Spain.

[Schiel, 1999] Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech Proc. of the ICPhS 1999. San Francisco, August 1999. pp. 607-610.

[Schiel et al, 1999] Schiel, F.; Draxler, Chr.; Hoole, Ph.; Tillmann, H.G. (1999). New Resources at BAS: Acoustic, Multimodal, Linguistic, Proc. of the Eurospeech 1999, Budpest, Hungary, pp. 2271-2274.