Three New Corpora at the Bavarian Archive for Speech Signals - and a First Step Towards Distributed Web-Based Recording

Christoph Draxler, Florian Schiel

Bavarian Archive for Speech Signals (BAS)
University of Munich
Germany
{draxler|schiel}@bas.uni-muenchen.de

Abstract

The Bavarian Archive for Speech Signals has released three new speech corpora for both industrial and academic use: a) Hempels Sofa contains recordings of up to 60 seconds of non-scripted telephone speech, b) ZipTel is a corpus with telephone speech covering postal addresses and telephone numbers from a real world application, and c) RVG-J, an extension of the original Regional Variants of German corpus with juvenile speakers. All three corpora were transcribed orthographically according to the SpeechDat annotation guidelines using the WWWTranscribe annotation software. Recently, BAS has begun to investigate performing large-scale audio recordings via the web, and RVG-J has become the testbed for this type of recording.

1. Introduction

BAS (Bavarian Archive of Speech Signals) is one of the major speech database providers for German speech and language resources. It was founded in 1995 and it is hosted by the Institut für Phonetik und Sprachliche Kommunikation at the university of Munich (Tillmann et al. 1995).

BAS has participated in many national and European projects, e.g. Verbmobil, SpeechDat, EAGLES, etc. and currently is the primary data collector and first annotator for the German SmartKom project (Schiel et al. 2002, Türk 2001).

The three corpora presented here are spin-offs of past projects and they extend the corpora collected in these projects by additional material that is of relevance to both technology development and research.

Corpus	#Session	#Item	Type of speech
Hempels Sofa	3920	1	spontaneous speech
ZipTel	1960	4	real world application
RVG-J	185	128	prompted juvenile speech

Table 1: New BAS Corpora

The paper is structured as follows: Section 2 gives an overview of the new corpora. Section 3 presents the corpora in detail, together with a first analysis. A proposed procedure for high-quality recordings via the WWW is given in section 4 and section 5 gives an outlook on future work.

2. Overview of the corpora

2.1. Recordings

Hempels Sofa and ZipTel are telephone recordings in ISDN quality, i.e. 8 KHz sampling rate and 8 bit a-law encoded quantization. The recordings are stored as headerless raw data files. A telephone speech server prompted the speakers either with a prerecorded prompt message or simply a beep.

The prompt for Hempels Sofa was the last item of the German fixed network SpeechDat-II prompt sheet. The speech server was configured to stop recording either upon reaching a given recording length, or via silence detection.

In ZipTel, speakers were prompted auditively only. Speaker could navigate through the welcome and help messages via DTMF tones; once they entered the address mode, they could not change the course of action. The speech server prompted for each address item explicitly, e.g. family name, given name, street, zip code, etc.

RVG-J was recorded in two different environments: a modified office room and a recording studio. The recording stations were first generation Power Macintosh computers.

In the office room, the recording station was positioned on the desk in front of the speaker. In the studio the speaker sat in a sound-proofed StudioBox in front of an LCD display; the recording station was outside the StudioBox and could not be heard. Two microphones were used: a Beyerdynamic MCE10 headset and a Beyerdynamic NEM192 clip microphone attached to the speaker's collar. The recording duration for every item was fixed and speakers could not alter the recording sequence. All prompts were presented in text form only; requests for spontaneous speech differed in font from prompts that the speaker was asked to read. The recordings were not supervised.

The sampling rate was 22,050 KHz with 16 bit linear quantization. The recording format was QuickTime which was then converted to WAV. One second of environment noise was recorded prior to each speech recording.

2.2. Web-based annotation

The recordings are annotated according to the SpeechDat-II guidelines (Senia, van Velden 1996). The annotation is orthographic with four marker symbols for speaker and non-speaker noise and hesitations plus markers for mispronounced words, incomprehensible speech and signal truncation.

The three corpora were annotated using WWWTranscribe (Draxler 1997). WWWTranscribe is a web-based client/server annotation system. The server stores signal file and common resources, e.g. the lexicon and log files, whereas the annotations are performed on the clients using standard web browsers. The annotation is facilitated by editing buttons that automate common tasks, e.g. conversion from digits to strings etc. The annotation text is checked for formal correctness on the client and only sent back to the server when it is correct (Fig. 1).

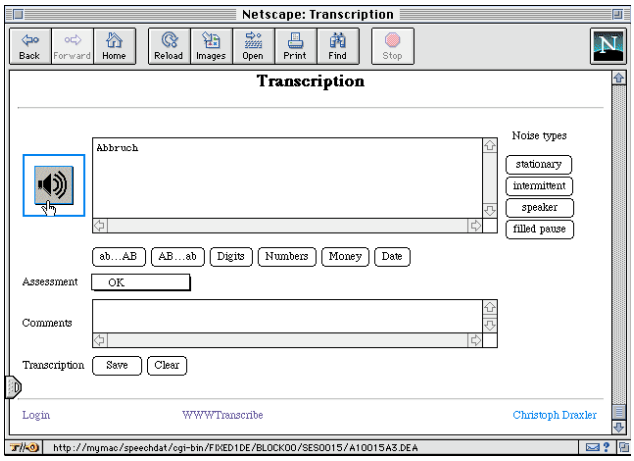

Fig. 1: WWWTranscribe screenshot

The annotations are stored in SAM label files in ISO 8859 format. For every signal file there is a corresponding label file. The label file also contains signal data, e.g. signal duration, sample rate, quantization and administrative data such as recording location, date and time.

2.3. Distribution

All three corpora are formatted following the SpeechDat-II database exchange format (Senia 1997). This means that they contain the speech signal, label files in both SAM and BAS Partitur Format (Schiel et al. 1998), a SAM-PA pronunciation dictionary, signal statistics and log files and documentation.

The three corpora are now available on CD-ROM or DVD via BAS.

3. The Corpora in Detail

3.1. Hempels Sofa

Hempels Sofa consists of up to one minute of spontaneous speech by 3920 speakers.

3.1.1. Demographic distribution

The demographic distribution of speakers meets the SpeechDat-II criteria for age, gender (49.2% female, 50.7% male, 0.1% unknown), speaker accent and the environment from which the speaker called (public phone booth, home, or office). The German federal states were used as a basis for determining speaker accent. Speakers were asked in which federal state they had entered school. The federal states do not match the dialect regions exactly, but speakers could easily provide the required information (the acronyms in Table 2 represent the German federal states).

Accent	#Session	Accent	#Session
BB	97	NW	681
BE	167	OTHER	106
BW	432	RP	195
BY	816	SH	82
HB	22	SL	47
HE	328	SN	223
HH	50	ST	95
MV	70	TH	85
NI	355	UNKNOWN	69

Table 2: Speaker accent distribution

Environment	#Session	%
Booth	85	2,2%
Home	2627	67,0%
Office	554	14,1%
Unknown	654	16,7%

Table 3: Calling environment distribution

3.1.2. Elicitation of colloquial speech

The aim of this recording was to obtain colloquial non-scripted speech with marked regional accent. To achieve this goal, speakers were prompted to speaker about everyday life: "*Erzählen Sie, was Sie in der letzten Stunde gemacht haben. Sie haben dazu eine Minute Zeit*" (Report what you have been doing during the hour preceding the recording. You have one minute.).

This prompt appeared as the last item in the German SpeechDat-II prompt sheet. By the time this item was recorded speakers were familiar with the recording procedure and they knew that the interview would be over soon. It was expected that this knowledge would put them into a relaxed mood in which they were likely to switch to a colloquial speech style.

The transcriptions contain 189951 tokens (including marker symbols), the pronunciation lexicon contains 12215 types. The average word count is 47.9, the average duration of the recordings is 23.4 seconds.

3.2. ZipTel

ZipTel contains data from a real application where speakers were interested in achieving a particular goal, namely to register as a participant for the German SpeechDat-II data collection and to obtain a prompt sheet.

3.2.1. Demographic distribution

Calls for participation in the SpeechDat-II project were published in German daily newspapers, company magazines (SiemensWelt) and in the journal of the German engineering society (VDI-Nachrichten). The daily newspapers were addressed according to their regional distribution and we asked to put the call for participation into the local news section.

As a consequence, the demographic balance by region is not representative for Germany. The gender distribution is approx. 53% female, 47% male.

Speakers were prompted by a speech server to provide their name, address and telephone number.

For privacy reasons, only the street name with number, ZIP code with city name and the telephone number is included in the corpus. ZipTel comprises 7721 utterances from 1960 recording sessions with a total of 43866 tokens. The lexicon contains 3049 types (Table 4).

The recording duration was preset for every item, additionally, silence detection was used.

Code	Item	Recordings	Count
Z0*	first name	2140	
Z1*	family name	2060	
Z2	street name	2018	1947
Z3	ZIP code	2007	1949
Z4	city	1993	1906
Z5	telephone no	1972	1919
Z6*	repeat all	1931	
Z7*	mobile phone yes/no	1696	
Z8*	source	1702	

Table 4: ZipTel Corpus (items marked with * are not included in the corpus)

Because of the nature of the recordings, many real application phenomena such as barge-in, meta-talk and error correction occur in the ZipTel corpus. These phenomena are of particular interest because they are difficult if not impossible to elicit via prompt sheets or in scripted recording settings.

Although speakers were explicitly informed that the speech server would prompt for one item after the other, many speakers provided all information in the first recording.

Fig. 2 shows the number of recorded items in the recording sequence vs. the number of items included in the corpus.

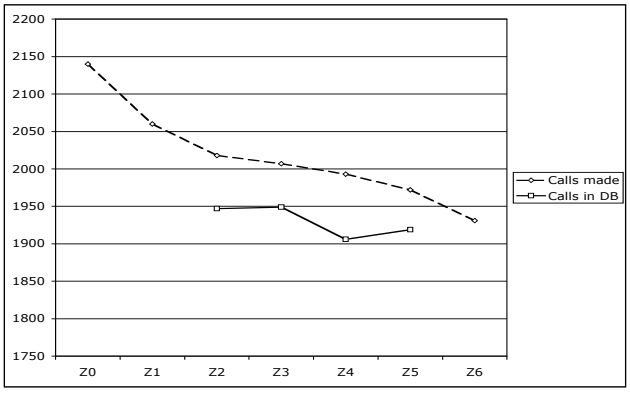

Fig. 2 Count of recorded items

The upper curve shows that only 79.5% of all speakers reached the end of the recording session and that there is a sharp decline from items Z0 and Z1 to the following items (items Z7 and Z8 were not counted because they were recorded only for a subset of all sessions). The reasons for this are that although they were told that each part of the address would be recorded separately they often provided all information at once; when they realized this at the second prompt, they aborted the call and called again immediately. Once they knew the procedure only a few calls were aborted.

The lower curve shows a marked decline for the city name and the telephone number items. In Germany, zip code and city name are tightly coupled. Hence, speakers would give the city name in the zip code item already and subsequently did not speak the city name again at the next item. The lower number of recorded telephone numbers stems from the fact that some speakers were reluctant to reveal their telephone number when they had already given a postal address.

3.3. RVG-J

RVG-J (Regional Variants of German - Junior) is an extension of the original RVG corpus which was collected by BAS in collaboration with AT&T (Burger, Schiel 1998).

The material recorded consists of the original 45 RVG prompts (digits, numbers, phonetically rich words, spontaneous speech), plus 83 extra items which were taken from the SpeechDat-II corpus (names, naturally formatted credit card and telephone numbers, phonetically rich sentences and spontaneous speech). In the added spontaneous speech prompts, speakers were requested to perform common tasks that are of interest for the development of industrial applications and services, such as leaving a message on an answering machine, inquiring for a given name at a directory service, or describing the route to the current location.

3.3.1. Demographic distribution

In RVG-J, juvenile speakers of age 11 to 19 were recorded. 11 was considered to be the minimum age because some prompts contained foreign language words; 19 was the upper limit because this is generally the age at which pupils finish school. They were recruited via calls for participation distributed at public secondary schools (Gymnasium and Realschule) in the Munich area.

RVG-J comprises 185 sessions with 93 female and 92 male speakers. 96.2% of the speakers entered school in Bavaria (BY), most of them in the Munich area. 91.9% of the speakers were 13-16 years old, 5.9% were younger, 2.2% were older. 93 recordings took place in the office room, i.e. with environment noise from the computer, 91 recordings were done in the StudioBox.

RVG-J contains a total of 21852 recordings with 133035 tokens, the lexicon contains 4629 types.

3.3.2. Extension of the project

RVG-J is an ongoing project. It is our goal to perform similar recordings in all German speaking regions (including Austria and Switzerland) in a distributed effort: BAS provides software, prompt material and logistic support, while the remote partner organizes the recordings locally. Annotations will be produced either at BAS or at the remote site. Royalties will be shared among all contributors to RVG-J.

4. Web-based recording

Speech recordings can either be performed on site, or remotely via the telephone. On site recordings require that either the speaker come to a recording studio, or that an interviewer visit the speaker. In a recording studio, the technical setup allows perfect recordings, but naturalness is bound to suffer. Recordings in the field are often of lower technical quality, but they are highly natural

because the speaker remains in a familiar environment. Both types of recording are very time-consuming and expensive because of the travel of either the speakers or the interviewer and the equipment.

Recordings over the telephone allow the speaker to remain in a comfortable environment and they are cheap because in general only telephone charges apply. However, the technical quality is restricted to ISDN quality at best, i.e. single channel 8 KHz 8 bit with lossy alaw compression.

Web-based recording offers the best of both worlds: the advantages of on site recordings, namely high quality recordings in a familiar environment, are combined with the distributed and cheap recordings via the telephone. The basic idea is to use an Internet-savvy application or a web browser to record speech and then transmit this speech to a server as data packets. In this procedure the quality of the recorded signal is decoupled from the transfer rate of the medium – high quality recordings may be transmitted via ISDN or even GSM phone lines (of course not in real-time).

BAS has begun to investigate the feasibility of performing high-quality speech recordings via the Internet.

For RVG-J, the JSpeechRecorder software was developed. It is implemented in Java using QuickTime, the multi-media architecture for Windows and Macintosh. JSpeechRecorder features not only text-prompts, but can also display pictures, audio, or video for the elicitation of speech.

WebRecorder, the successor software to JspeechRecorder, is now under development, an alpha version is currently undergoing tests. In WebRecorder, access to the WWW is built-in: prompt texts and images are downloaded from a prompt server and speech signals converted to data packets are uploaded to a recording server. The format used is similar to that of mail attachments: a multi-part document contains the recordings, and this document is sent to the server. The individual parts of the multipart document contain administrative information and the signal data proper.

In the final version, the recording will be performed by an applet within a standard web browser. This has the advantage that the client always uses the latest version of the software and that no software needs to be installed on the client.

5. Summary and outlook

Hempels Sofa, ZipTel and RVG-J are the latest additions to the collection of BAS corpora. Both Hempels Sofa and ZipTel are complete.

RVG-J really has just begun. BAS encourages participation in RVG-J and it is willing to 1) provide software for both recording and annotation, and 2) include the data recorded by partners in its catalog of speech resources. RVG-J will also be the testbed for WWW-based recordings. In the first phase, schools all over Germany will be given the audio equipment necessary for high quality recordings and our recording software. In exchange they will be asked to deliver recordings of their pupils. In a later phase, anyone willing to donate speech may register and receive a microphone or some other reward in exchange for the speech recordings.

6. References

Draxler Chr., 1997. WWWTranscribe - A Modular Transcription System Based On The World Wide Web, *Proc. of Eurospeech 1997*, Rhodos, Greece

Burger S., Schiel F., 1998. RVG 1 - A Database for Regional Variants of Contemporary German. *Proc. of the 1st Int. Conf. on Language Resources and Evaluation 1998*, Granada, Spain, pp. 1083-1087

Schiel F., Burger S., Geumann A., Weilhammer K., 1998. The Partitur Format at BAS. *Proceedings of 1st Intl. Conference on Language Resources and Evaluation*, Granada.

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.

Schiel F., Steininger S., Türk U., 2002. The SmartKom Multi-modal Corpus at BAS, *Proceedings of the 3rd Intl. Conference on Language Resources and Evaluation*, Gran Canaria

Senia F., 1997. Specification of Speech Database Interchange Format, SpeechDat report LE2-4001 - SD1.3.1 Version 4.3

Senia F., van Velden J., 1997. Specification of Orthographic Transcription and Lexicon Conventions, SpeechDat report LE2-4001 – SD 1.3.2 Version 2.4

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.

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

Türk U., 2001. The Technical Processing in SmartKom Data Collection: A Case Study, *Proceedings of EUROSPEECH Scandinavia, Aalborg*, 2001, pp.1541-1544.