PhattSessionz: Recording 1000 Adolescent Speakers in Schools in Germany

Chr. Draxler, A. Steffen

Bavarian Archive for Speech Signals (BAS)
c/o IPSK, Ludwig-Maximilians-Universität München
{draxler|al-x}@phonetik.uni-muenchen.de

Abstract

PhattSessionz is a distributed speech data collection to create a regionally balanced speech database of more than 1000 adolescent speakers in Germany. Recordings are planned in more than 35 secondary schools all over Germany. The recording equipment consists of a headset and a table microphone connected to the PC via a USB audio device with a signal quality of 22.05 KHz 16 bit. All recordings are immediately uploaded via the WWW to the PhattSessionz server at the BAS. This server also provides the prompt sheets which are customized for each recording session. The local organization of the recordings, e.g. speaker recruitment and recording schedule, is delegated to the participating schools.
A number of technical issues arose during the first experimental recordings in Munich schools. These problems have been solved now and production recordings have started.

1. Introduction

To our knowledge there currently do not exist publicly available large speech databases of adolescent speakers for German. Such databases are necessary for the development of speech technology and spoken language research.

Adolescent speech differs in many respects from that of adults, e.g.

- Physiological: accelerated growth, the change of hormone system, and the breaking of male voices affects the voice quality.
- Social: speaking style is used conscientiously to establish peer groups.
- Educational: exposure to foreign languages in school and the use of domain-specific terminology, e.g. natural science education, games, sports, results in a growing and multi-lingual active vocabulary.

Furthermore, adolescent speakers are an interesting target audience for speech technology because young people are considered to be eager adopters of innovative technologies, e.g. mobile phones, games, etc. and applications.

In most European countries, the legal age is 18 or higher. The recording of younger speakers thus requires their parents' consent, the observation of child protection laws, and continued supervision, leading to a very high administrative and procedural overhead. Until now, this overhead has been an unsurmountable obstacle in the collection of large adolescent speaker speech databases.

Within PhattSessionz we have developed a new approach in which the administrative and procedural workload is distributed while the overall coordination is performed centrally. Data transfer is performed automatically by performing recordings over the WWW. The workload is thus reduced to an acceptable level for each participant in the project.

The remainder of the paper is structured as follows: Section 2 describes the PhattSessionz database contents and structure. Section 3 presents the recording equipment and software used, and section 4 outlines the workflow. Section 5 reports on the first experimental recordings and the technical issues that had to be solved, and section 6 gives a summary.

2. Database Contents

For compatibility reasons, the PhattSessionz database is an extension and superset of the RVG database [1] and the German SpeechDat-II database [5], [6].

RVG is a regionally balanced database of application oriented vocabulary with 500 speakers from 9 main dialect regions of Germany, Austria and the German speaking part of Switzerland. The recordings were performed in 22.05 KHz, 16 bit signal quality using a headset and three table microphones. All recordings took place in quiet office rooms at IPSK in Munich. Speakers were recruited locally, which for geographically distant dialects was a considerable challenge. The dialect classification was performed by an expert.

SpeechDat-II is a speech database for the development of voice operated applications and services with up to 5000 speakers. Speakers called a telephone server over the mobile or the fixed telephone network and read prompts from a prompt sheet. In the German SpeechDat project, 4000 speakers called via the fixed network, 1000 via the mobile network. For the dialect classification, speakers were asked in which federal state they entered school, resulting in 18 classes. Speakers were recruited all over Germany [4].

2.1. PhattSessionz contents

PhattSessionz contains all prompts from RVG and SpeechDat-II. Additionally, it contains non-scripted speech, including task-specific and expressive speech. This type of speech was elicited via questions, e.g. "What did you watch on TV last night?" and prompts, e.g. "Please tell us about a funny thing that happened to you", or "Leave a message on the answering machine of your friend."

A recording session comprises 129 utterances (see table 1). Every recording session uses its own distinct prompt sheet. This prompt sheet contains a standard introductory section used to acquaint the speaker with the procedure and to test the recording equipment. The main section consists of prompt items in randomized sequence;

each item has a fixed maximum recording duration. In the last section, the expressive speech items are collected.

The average session duration is approx. 25 minutes with a maximum recording duration of 21 minutes; the average speech duration is 7 minutes.

Table 1: PhattSessionz databases

Count	Description
3	question
6	description
2	character sequence spelling
3	person name spelling
5	city name spelling
1	spontaneous speech
3	read date expression
3	read company name
3	read person name
3	read city name
30	read (phonetically rich) sentence
12	read command
3	read PIN code (4 digits)
3	read credit card number (16 digits)
13	read telephone number (6-16 digits)
3	read time of day
11	read isolated digit
3	read digit string (10 digits)
19	read number
129	

2.2. Speaker demographics

The PhattSessionz database will contain 1000 speakers in the age range of 13 to 18 with a balanced sex distribution. Furthermore, it will be balanced by dialect regions. The dialect regions are those defined in [3]. Note that this specification of dialect regions was not available for RVG or SpeechDat, and that the regions used in these projects must be mapped manually to the regions used in PhattSessionz.

The dialect regions will be covered by performing recordings in major cities in all regions. The speaker data collected will give a first indication of the dialect region. This will then be refined during annotation, where a dialect expert will determine the dialect of the given speaker.

3. Distributed recordings

To achieve the regional coverage, recordings will be distributed geographically. In PhattSessionz, schools in major cities will be asked to participate in the project (see fig. 1 for the geographic distribution).

3.1. Recordings in schools

Schools in Germany usually have a reasonably fast connection to the Internet, e.g. via DSL, they have skilled and motivated staff, e.g. teachers, leaders of interest groups, tutors, etc., they have class rooms or offices available during daytime, especially in the afternoon, and they have established means of informing parents of ongoing activities and obtaining their consent for such activities. Because the collection of the speech database is a non-profit effort aimed at supporting basic research and technology development, schools can decide on their own whether they contribute to the project, without requiring permission by the education ministry of the federal state.

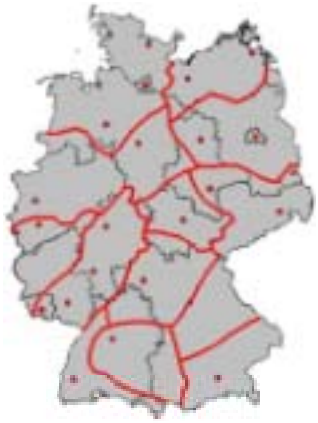

Figure 1: PhattSessionz dialect regions and recording cities (adapted from [3])

3.2. Incentives offered

For their participation, schools are offered 200 € for 30 speakers (see 5.2). The school is free to decide on how to use the money. Furthermore, classes of the participating schools are invited to visit the IPSK when they are on excursion in Munich. This is an opportunity to see how the data the school provided is being processed (as a side-effect, it allows IPSK to inform future students about a subject that is not covered by school curricula). Finally, all participating schools are listed on the project web page.

Schools were also assured that the recordings would not permanently install software on their computers, and that no files would remain on the local disks after recordings. Furthermore, we promised to respect the access policy to Internet connections of the school (see section 5.1 for the problems encountered).

3.3. Privacy issues

To guarantee the privacy of the speakers, all recordings are anonymous. Furthermore, the form with the parents' signature and the name of the pupil remains at the school; once all required recordings are done, the schools sends a form to IPSK. IPSK then checks the recording contents and signal quality. If it finds problems, then the school is asked to re-record the speaker identified via the speaker code; if there are no problems, the school can destroy the parents' forms. To facilitate the procedure, IPSK has set some tolerance limits for the number of speakers and the age and sex distribution.

Great care is taken to inform schools and parents openly about the project. Telephone numbers of staff members and an informative web site have been set up, and

providing quick responses to inquiries of parents or schools is a top priority.

3.4. Finding schools

The first step of getting in contact was to obtain a complete list of all schools in every county of all federal states of Germany. Then the dialect-region and the county were entered into a database so that mailings and phone calls could be prepared. This database includes the address of the school and its administrative head. If the school is participating, then also the person who will be running the recordings, administrative details and finally the payment details are entered into the database. The addresses of the schools were provided by the regional education ministries who at this occasion were informed about the project.

At IPSK, members of the team designed a guideline for the initial, informal phone call to schools. In this first step it was already checked whether the school has a sufficiently fast Internet connection.

The second step was to send by fax or mail an information package including:

- A letter to the headmaster of the school describing the project and the aims of the study, the possible contribution of the school and the incentives offered.
- Posters to be placed on the public panels to inform the pupils
- A technical information how to install the software and how to run the recording sessions
- Registration forms for interested pupils, including a part to inform the parents and to let them sign the form to officially allow their children to take part in the study.

After finding a contact-person at the school and identifying the responsible partner a time-frame was negotiated for the actual recordings.

4. Recording Setup

The IPSK has set up 7 identical recording kits consisting of an M-Audio Mobile Pre USB A/D interface, and a Beyerdynamic Opus 56 headset and an Acoustic Technology AT 300 tabletop microphone. The Mobile Pre allows the recording level to be adjusted manually. The device is powered by the USB bus, reducing the number of cables on the desk. The Mobile Pre USB requires drivers to be installed on the recording PC (see fig. 2).

The signal quality is set to 22.05 KHz with 16 bit quantization, and two channels are recorded, resulting in a raw data rate of approx. 88.2 Kbyte/s.

The kits are sent to schools participating in the PhattSessionz recordings. The school provides an office or a class room with a standard PC, and the person responsible for the local recordings installs the equipment and performs a test recording.

4.1. Recording software

For the PhattSessionz speech database collection, the SpeechRecorder software is used. SpeechRecorder is a platform independent application for performing scripted speech recordings. Its primary features are a flexible recording script with a precise recording protocol, multi-media prompting and multi-channel input. Furthermore, it can be operated as a standalone application or in a client/server configuration. In the client/server configuration, the client performs the actual recording, and the signal data is transferred to the server via a standard http upload. In this manner, high bandwidth recordings can be distributed geographically [2].

SpeechRecorder is implemented as a Java WebStart application. Thus it does not need to be installed on the local machine, reducing the system administration overhead.

4.2. Test recordings

When the recording equipment is set up, a test recording is performed. This test recording runs the recording script of an entire recording session, but instead of speech a sine wave is recorded. At IPSK these recordings are then analyzed automatically for dropped frames or other signal errors (see section 5.1).

Once the test recording is accepted by IPSK, the school may proceed with the production recordings.

Figure 2:Recording equipment and setup

4.3. Recording procedure

For every recording, the recording supervisor at the school logs into the PhattSessionz server and enters the basic demographic data of the current speaker: age, sex, dialect region, weight, size, smoking habits and whether the speaker has braces or piercings in the mouth.

The PhattSessionz server generates a new prompt sheet which is then downloaded to the PC. The supervisor starts the actual recording and guides the speaker through the introductory section of the session. If the recording equipment works, the supervisor leaves the room, and the speaker performs the remainder of the session in an unsupervised mode.

5. Problems encountered

To test the procedure and equipment, two schools in the Munich area were asked to participate in a field test. This field test was carried out in November 2004 at the Oskar-von-Miller Gymnasium (OvMG) and in January 2005 at the Gymnasium Geretsried (GG).

5.1. Technical issues

At OvMG, the recording equipment was installed by IPSK staff, at GG, the equipment was installed by the teacher responsible for the computer courses.

5.1.1. *Dropped frames*

OvMG provided a 1 GHz PC equipped with 256 MB of RAM with a DSL connection to the Internet. All network traffic was routed through a firewall administered by an external administrator. The OS used was Windows 2000.

Accessing the server for login and downloading the SpeechRecorder software and the prompt sheet worked without problems. The upload data rate available for uploading the signal varied with the time of day: in the early afternoon, about 30 Kbyte/s could be achieved, in the late afternoon 75 Kbyte/s. As a consequence, upload times varied from more than 40 minutes, i.e. twice the recording session duration, to 20 minutes.

Signal analysis at IPSK revealed that frames had been dropped spuriously during recording. The cause of the dropped frames could not be determined exactly: driver updates, installing Windows patches or increasing the buffer size in the Mobile Pre did not eliminate the problem.

Dropped frames seem to depend on a particular hardware configuration, and hence every school now has to perform a full recording session with a test recording of a sine wave. This allows detecting dropped frames.

5.1.2. *Duplicate transfer of data*

GG provided a PC with Windows 98 with a DSL connection to the Internet via a physical firewall device (Sbox). For the Mobile Pre, USB drivers had to be installed because Windows 98 does not automatically support USB.

The recording application could be downloaded and started without problems. However, the data transfer rate was extremely low (< 7 Kbyte/s). Closer analysis revealed that all data was being transferred twice, each time with a different authentication key. Two certificates were used because the login procedure and the recording program were considered to be two distinct applications by the firewall. A solution to this problem was to integrate starting the application into the login procedure so that only one certificate was issued.

Additionally, all signal data was compressed using the free lossless audio compression package FLAC. For the type of signal recorded in PhattSessionz (short utterances with relatively long initial and final silence), FLAC reduces the amount of data to be transferred to about 50%.

With these changes, recordings can now be performed almost in real time, even via firewalls or shared access to the physical Internet connection.

5.2. Speaker Recruitment Problems

At OvMG, speakers were recruited by distributing information sheets with a response form in the classes. The return rate of forms was very low, and during the first two days only very few pupils could be recorded. In a second try, a member of the IPSK visited several classes, accompanied by the headmaster. The recording procedure and project aims were described briefly, and pupils were asked to enter their names directly into a recording timetable. These pupils were then given the information sheet and were requested to bring the form signed by their parents to the recording. In this manner, a large number of pupils could be recruited quickly. Even more, they all turned up for the recordings, and all could provide a signed form.

At GG, the Internet interest group was charged with recruiting speakers. This group was willing to be recorded, but the technical problems encountered lead to many potential speakers losing their interest. In fact, only a few speakers were effectively recorded at GG. However, the readiness of the interest group to participate and to recruit further speakers was real, and we expect that such a speaker recruitment will be the most effective and efficient.

Note that the only incentive for speakers were the prospect of participating in a technologically ambitious project and some candy after a recording session.

5.3. Status of the Recordings

As of June 10th, we have contacted more than 50 schools in Germany and have received positive responses from 14. Recordings are under way in five cities; the schools are asked to perform the recordings within three weeks.

In order to reduce the recruitment effort, schools are now requested to record only a minimum of 30 speakers (15 female, 15 male) instead of the original 50. As a consequence, more schools have to participate to reach the target of 1000 speakers.

We are now actively trying to get media coverage: an interview with a news agency journalist at DAGA 2005 led to interviews with newspapers and magazines. These will now be used to persuade additional schools to participate.

6. Conclusion

The technical problems have severely delayed the PhattSessionz recordings. However, since they were discovered during field tests and not during production recordings, they did no damage. Hence we are optimistic that once a few schools have participated successfully, further schools will join the database collection.

7. Acknowledgments

We thank Klaus Jänsch for his continued improvement of the recording software implementation, and Meral Akyol and Angela Baumann for contacting schools. Parts of this work have been supported by the German Federal Ministry of Education and Research grant no. 01IVB01 (BITS).

8. References

[1] Burger, S. Schiel, F. RVG 1 - A Database for Regional Variants of Contemporary German. Proc. LREC 1998, Granada, Spain

[2] Draxler, Chr., Jänsch, K. *SpeechRecorder – A Universal Platform Independent Multi-Channel Audio Recording Software*, Proc. LREC 2004, Lisbon, Portugal

[3] Hollmach, U. *Untersuchungen zur Kodifizierung der Standardaussprache in Deutschland*, Habilitationsschrift, Halle (Saale), 2003

[4] Lindberg, B., Comeyne, R., Draxler, Chr., Senia, F. *Speaker Recruitment Methods And Speaker Coverage – Experiences From A Large Multilingual Speech Database Collection*, Proc. ICSLP 1998, Sydney

[5] Velden, J.G. van, Langmann, D., Pawlewski, M. *Specification of speech data collection over mobile telephone networks*. SpeechDat Technical Report SD1.1.2/1.2.2, 1996

[6] Winski, R. *Definition of corpus, scripts and standards for Fixed Networks*. SpeechDat Technical Report SD1.1.1, 1997