Speech Recordings in Public Schools in Germany – the Perfect Show Case for Web-based Recordings and Annotation

Christoph Draxler, Klaus Jänsch

Institut für Phonetik und Sprachliche Kommunikation
Ludwig-Maximilians-Universität München
Schellingstr. 3, 80799 Munich, Germany
{draxler | klausj}@phonetik.uni-muenchen.de

Abstract

In the Ph@ttSessionz project, geographically distributed high-bandwidth recordings of adolescent speakers are performed in public schools all over Germany. To achieve a consistent technical signal quality, a standard configuration of recording equipment is sent to the participating schools. The recordings are made using the SpeechRecorder software for prompted speech recordings via the WWW. During a recording session, prompts are downloaded from a server, and the speech data is uploaded to the server in a background process. This paper focuses on the technical aspects of the distributed Ph@ttSessionz speech recordings and their annotation.

1. Introduction

Ph@ttSessionz is a speech database collection subproject within the BITS (BAS Infrastructure for Technical Speech Processing) project[1]. In Ph@ttSessionz, a database of 1000 adolescent speakers (age 13-18) will be collected.

Adolescent speakers are interesting for both speech technology development and speech research because their voices and speaking styles differ substantially from those of adults. Currently no publicly available database of adolescent speakers exists for German, and the Ph@ttSessionz database will remedy this situation.

The database contents is a superset of the RVG (Regional Variants of German) [Burger/Schiel 1998] and the German SpeechDat databases [Winski et al. 1996], i.e. it contains digits, formatted digit strings, numbers, time and date expressions, phonetically rich words and sentences. Additionally, a number of questions elicit spontaneous speech, e.g. "What did you do last week?". A total of 125 items is recorded per speaker.

The database is balanced by gender and covers all major dialect regions in Germany.

The Ph@ttSessionz recordings take place in public high schools (*Gymnasium*) in Germany [Draxler/Steffen 2005]. The participating schools are responsible for the recruitment of speakers and for the local organisation of the recordings. Each school is asked to record 30 speakers.

The speech data is transcribed according to the SpeechDat annotation schema, i.e. an orthographical transcript with a small set of noise, error and signal quality markers [Winski 1997].

The Ph@ttSessionz database will be publicly available via the Bavarian Archive for Speech Signals, BAS.

2. Technical Infrastructure

The Ph@ttSessionz recordings and annotations are performed in a client-server architecture via the WWW using the SpeechRecorder and the WebTranscribe software developed at BAS [Draxler/Jänsch 2004, Draxler 2005].

For distributed recordings, the server provides the prompts and stores the signal data, and the recordings are performed on the client.

For the annotation, the client requests the signal files to be annotated to the server, and it returns the annotation data to the server.

This architecture allows a distribution of workload, close monitoring of recording progress, immediate access to recorded signals, and sharing of resources, e.g. prompts and administrative data.

2.1. Implementation

The Ph@ttSessionz server is implemented as a web application using Java Server Pages, the Apache Tomcat server software and the PostgreSQL relational database management system on a PC platform running SuSE Linux.

Both SpeechRecorder and WebTranscribe are provided as Java Web Start applications. They are downloaded via a standard web browser and executed by the Java Web Start environment on the client. The applications are signed by a certificate, and the Web Start environment guarantees that the applications do not violate security constraints, e.g. access restrictions.

2.1.1. User classes

The Ph@ttSessionz web site provides in a single consistent interface four different views for different classes of users. These users are visitors, recording supervisors, transcribers and system administrators.

Visitors are allowed to access the public pages on the server, i.e. a project description, media articles, demo videos of a recording, technical information, etc.

Recording supervisors receive a login and password from BAS. They enter the speaker data into the database and start a recording session.

Transcribers access the Ph@ttSessionz server to start the annotation software.

System administrators have full access to the server to monitor recording progress, update web pages, and for software maintenance.

[1] This work has been supported by the German Federal Ministry of Education and Research grant no. 01IVB01 (BITS).

2.1.2. Recording procedure

The recording supervisor at a school enters the login and password via a standard web browser on the client on which the recordings will be performed.

The server opens a new recording session and presents a form to enter the speaker data. This data consists of speaker sex, date of birth, German federal state where he or she entered school, native language of father and mother, smoking habits, presence of dental braces or piercings in the lips or tongue, and size and weight.

Once this data has been entered, the recording supervisor commences the recording session by clicking on a button in the web form. This starts the SpeechRecorder application from the Web Start environment. The software fetches the recording script for the current recording session from the server and displays the initial welcome message and starts the recordings.

When all items of a session have been recorded, SpeechRecorder displays a message on the screen. If the transmission of the signal data has not ended yet, the software continues the transmission.

New recording sessions can only be started after all data from the previous session has been uploaded completely.

3. Setup and Test Procedure

To participate in Ph@ttSessionz, a school must meet certain minimum system requirements. These system requirements are listed on the public pages of the web server.

3.1. System requirements

The system requirements for Ph@ttSessionz speech recordings are:

- a fast Internet connection (DSL or better)
- Java Runtime Engine and Java Web Start version 1.4 or newer
- support for USB audio devices
- Microsoft Windows 98 (with USB support added), 2000 or XP, Linux or Mac OS X.

These requirements are met by most German high schools. They usually have a dedicated computer room for computer science courses, or desktop PCs in class rooms.

The PCs were often provided by German Telekom in a nation-wide effort to connect schools to the Internet. These PCs typically are 1 GHz Siemens-Fujitsu PCs with 256 MB of RAM and Windows 2000 or newer.

Most schools are connected to the Internet via a DSL link with a speed of 1024 kbit/s downlink and 128 kbit/s uplink or better. This link usually is shared between all PCs of the school.

3.2. Recording equipment

For consistency reasons with the RVG-1 speech database, it was decided to perform all recordings with a sample rate of 22.05 kHz and 16 bit linear quantization over two channels. The data rate of the signal thus is 705.6 kbit/s.

To achieve a consistent technical signal quality, the same recording equipment was used at all sites. It consists of a Beyerdynamic opus 54 headset microphone, an Audio Technica 3031 desktop microphone, and an M-Audio Mobile Pre USB audio interface.

To allow parallel recordings, 8 recording kits were used. These kits were sent to participating schools, and schools returned them after the recordings.

3.3. Equipment test

At every recording site, the setup must be tested after the installation of the recording equipment. This so-called "sine-test" is a standard full-length recording session, the only difference being that the signal recorded is a sine wave produced by the PC. For this sine test, the audio output jacks of the Mobile Pre device were connected to the input jacks.

For the recording supervisor, the sine test provides an estimate of the duration of the data transmission and hence of an entire recording session. This estimate is necessary to define a realistic recording schedule.

For the system administrators at BAS, the sine test allows a quick visual detection of dropped frames and other interferences. For this, the spectrogram display of Praat was used.

If the sine test was successful, the system administrators at BAS allowed the recording site to proceed to production recordings.

3.4. Level Adjustment

The first five items of every prompt session are used for adjusting the recording level. For these recordings, the supervisor screen with the signal display is shown. The speaker reads the prompts, and the supervisor adjusts the recording level until the signal is clearly visible in the display without clipping. To facilitate this task, one of the first prompts describes what the signal display should look like.

Online monitoring of recordings is also possible. For this, the system administrator at BAS checks the incoming signals. If signals are found to be either too weak or clipped, the remote recording supervisor can be notified of the problems immediately via a message mechanism built into the Ph@ttSessionz web application.

4. Technical Problems

A field test of Ph@ttSessionz recordings was performed in two schools in and near Munich, the Oskar-von-Miller Gymnasium (OvMG) and Gymnasium Geretsried (GG). These field tests revealed a number of technical problems, the most important of which were dropped frames and low data transmission rates.

During the production recordings, further technical problems emerged: USB driver incompatibilities, transmissions blocked by firewalls, and others.

Finally, user interaction with the web site was not entirely robust. Unforeseen user behavior lead to invalid or incomplete data and caused recordings to be lost.

4.1. Corrupted signal problems

In some recorded signal files, frames had been dropped spuriously. These dropped frames occurred at irregular intervals, and not every file was affected.

There were a number of possible explanations for this: insufficient computing power of the client machine, incompatibilities of the operating system and the audio device drivers, incomplete data transfer, or bugs in the Java audio implementation.

To check whether a weak client machine would produce corrupted signal files we set up an old PC at BAS with a modern operating system and a slow network connection (ISDN telephone line). Tests using this configuration showed that the processing power available was fully sufficient for the Ph@ttSessionz recordings.

Next, we tested for incompatibilities between the operating system and the audio device drivers. We tested both the M-Audio USB audio driver and the Windows USB audio drivers, and we experimented with different buffer sizes for the transfer of signal data from the audio device to the PC. The results of the tests were inconclusive.

To test for errors in the data transmission, check sums were computed for the signal files on both the client and the server side; they revealed that data transmission was reliable even over slow connections.

Finally, we tested for incompatibilities between Java versions and the operating system. We found (and reported) a bug in the implementation of the Java sound library, and observed that in some cases reverting to Java 1.4 instead of 1.5 resolved the problem.

In the end, no generally applicable and reliable configuration was found that would prevent dropped frames. As a consequence, the sine test was developed to at least detect dropped frames and other signal related problems early.

4.2. Transmission problems

We encountered the following types of transmission problems: slow data transfer, and signal files corrupted during transfer.

4.2.1. Slow transmission

In most schools, the data rate of the recordings was substantially higher than the available uplink speed of the Internet connection. The transmission of the recorded signal data thus took longer than the recording of the signal.

In the first version of the software, a new item could be recorded only after the previous recording had been uploaded completely. This lead to long delays between the recording of the individual prompt items.

To alleviate this problem, SpeechRecorder was modified to write all recorded data into an internal cache. Upload to the server would take place in the background, and recordings could proceed at their normal pace. Only when the cache was full did recordings have to wait. After an entire session had been recorded, the upload would continue until all data was transmitted. New sessions could only be started after the previous upload was complete. We had assumed that the pause between subsequent recording sessions would be long enough to allow all data to be uploaded before the next speaker would be available. However, this was not the case as most schools had a tight schedule for the recordings.

A second measure was to use the lossless audio compression FLAC (Free Lossless Audio Compression) to compress the audio files before transmission. For our speech recordings, FLAC achieved a compression rate of >40 %, which sped up transmission considerably and added only a very small computational overhead.

However, even both methods combined were not sufficient to allow a smooth progress of recordings in real time at all sites.

To solve this problem, SpeechRecorder was modified to write the cache to a temporary file on the client machine. The contents of this file could be uploaded at the end of the day when all recording sessions were completed.

We were aware of the fact that this modification could lead to the loss of data if the client machine was shut down before all data had been transferred and thus we allowed this option only at sites where data transmission was particularly slow and only after having instructed the recording supervisor to verify that all data was transferred after the very last recording session at his or her site.

4.2.2. Firewalls and proxies

Most schools access the Internet via a firewall. In some cases, this firewall was administered not by the school, but by an external agency, which made modifications to the firewall configuration very cumbersome or even impossible.

If the firewall prevented the data transmission, the recordings at this site were aborted.

In early versions of Java 1.5, data transfer via a proxy was very slow due to a bug in this Java version. Reverting to Java 1.4.2 was a suitable workaround until the problem was fixed in Java 1.5.6.

4.3. User interaction problems

The Ph@ttSessionz web site described the procedure for recordings in a step-by-step manner. Despite these instructions, problems occurred due to unforeseen user interactions with the web site.

4.3.1. Use of browser navigation buttons

The prescribed sequence of actions was to enter the speaker data, check this data and then start the recording session.

For every speaker, this procedure should be started from the index menu on the left side of the web page. However, recording supervisors sometimes used the "back"-button of the browser to return to the speaker data form to enter data for a new speaker. This resulted in overwriting all data of the previous session.

To remedy this problem, recording supervisors were warned by the software that the session they were about to open already contained recordings, and that they should open a new session from the menu.

4.3.2. Multiple browser windows and instances

The Ph@ttSessionz web site allowed new sessions only after the previous session had finished transferring its data. Because data transmission was slow, this resulted in long delays between sessions. To avoid these delays, recording supervisors thus sometimes opened additional browser windows, entered the speaker data and started new recording sessions.

For the recording supervisor, the recordings started from the new browser window would proceed as usual. However, the internal buffer of the SpeechRecorder software did not distinguish between data from different recording sessions. Hence, data from different recording

sessions interfered and inconsistent signal files were stored on the server.

To avoid such interferences, SpeechRecorder was modified so that each instance of the software wrote its buffer to its own, separate temporary file on the hard disk. Java has a built-in mechanism that ensures that temporary files are deleted once they are no longer needed. Thus we could be sure that after recording sessions had ended, no data would remain on the client machine.

5. Annotations

For the annotation of the Ph@ttSessionz recordings, the web-based annotation framework WebTranscribe is used. WebTranscribe implements a "segment - label - save" annotation cycle, with annotation editor plug-ins implementing the support for specific annotation schemes.

Both the signal and the annotation data are stored on the server, together with shared resources such as lexicons and meta-data (early implementations of distributed processing are described in [Huckvale et al. 1987], web-based annotation in [Draxler1999, Maeda et al. 2002]).

The Ph@ttSessionz editor plug-in consists of a graphical signal display in which the part of the signal that is annotated is marked, and of editing buttons. These buttons perform often needed tasks such as converting digits or numbers to their string representation, or placing noise markers into the editing field. Before the annotation is saved to the server, it is checked for formal consistency; only if the annotation is syntactically correct it is uploaded to the server.

Due to WebTranscribe's client-server architecture, annotations can be performed in a highly parallel manner, on any available machine, locally at the BAS or even at remote sites, e.g. schools participating in Ph@ttSessionz.

Since all annotations are stored on the server, they are available for further processing immediately. This reduces the risk of inconsistencies, eliminates the need for costly merging operations, and greatly facilitates the creation of derived resources, e.g. wordlists, lexicons, etc.

6. Current status

Production recordings started in April 2005. By the end of February 2006, 651 speakers from 31 schools (excluding the field tests at OvMG and GG) had been recorded, with more than 81.000 recorded items. This is an average of 20.7 speakers per school.

At the current rate of recordings of approx. 80 sessions per month, 1000 speakers will be reached by the end of July 2006.

Fig. 1 displays the duration of recording sessions. The rather high number of recording sessions with a duration of more than 60 minutes can be explained by a slow data transfer; session durations of more than two hours are the result of interrupted data transfer. With the very long recording sessions removed, the average session duration was 44:47 minutes.

Fig. 2 summarizes how many weeks the schools needed to either record the requested 30 speakers or cancel the recordings due to organizational or technical problems. 4 schools needed more than 10 weeks to complete the recordings; in all cases this was due to the long summer holidays. On average (and excluding the schools with 10 or more weeks) a school took 3.4 weeks for the recordings, and 17 (54.8 %) of the 31 schools took three weeks or less.

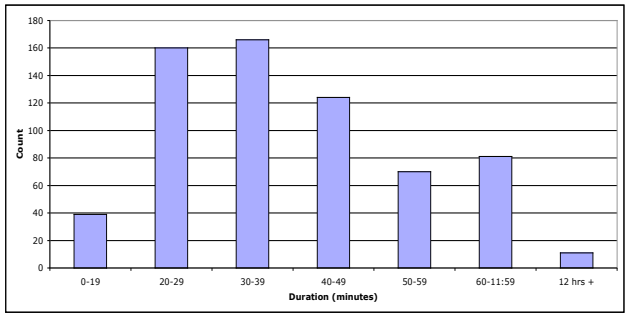

Fig. 1: Duration of recording sessions in minutes

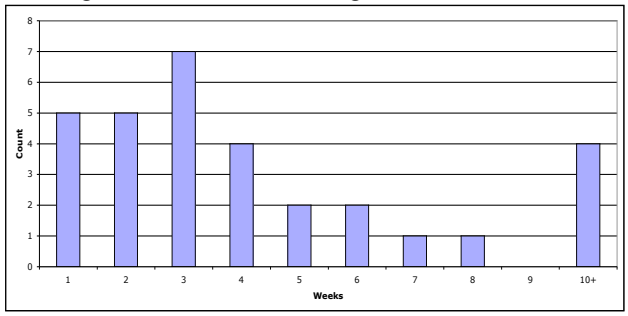

Fig. 2: Duration of entire recording periods in weeks

7. Conclusions

The Ph@ttSessionz project shows that distributed speech data collections and annotations are feasible for large-scale data collections. Most technical problems could be solved or workarounds could be found. The setup of equipment at the sites often required technical support from BAS. A close monitoring of recording progress by BAS for each school ensured that all recording could be achieved within an acceptable time frame.

8. References

Burger, S.; Schiel, F. (1998). RVG 1 – A Database for Regional Variants of Contemporary German. In *Proceedings of the 1st LREC*. Granada.

Draxler, Chr., Jänsch, K. (2004). SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software. In *Proceedings of the 4th LREC*. Lisbon.

Draxler, Chr., Steffen, A. (2005). Ph@ttSessionz: Recording 1000 Adolescent Speakers in Schools in Germany. In *Proceedings of Interspeech*. Lisbon.

Huckvale, M. et al. (1987). The SPAR Speech Filing System, In *Proceedings of European Conference on Speech Technology*.

Maeda et al. (2002). The Annotation Graph Toolkit: Software Components for Building Linguistic Annotation Tools. In *Proceedings of the 3rd LREC*. Gran Canaria.

Winski, R. et al. (1996). Specification of Telephone Speech Data Collection. LRE-63314 SpeechDat(M) Report D1.4.1

Winski, R. (1997). Definition of Corpus, Scripts and Standards for Fixed Networks. LRE-63314 SpeechDat(M) Technical Report SD1.1.1