

# On Web-Based Creation of Speech Resources for Less-Resourced Languages

Christoph Draxler

BAS Bavarian Archive for Speech Signals  
 Institut für Phonetik und Sprachverarbeitung  
 Ludwig-Maximilians-Universität München, Munich, Germany  
 draxler@phonetik.uni-muenchen.de

## Abstract

Web-based creation of speech resources is a new paradigm for producing spoken language resources. It is particularly suited for less resourced languages, i.e. languages for which no readily available speech resources exist. This paper maps the speech resource creation tasks to the client-server architecture of the WWW. It presents two tools that have been developed for web-based speech resource creation, and it demonstrates the effectiveness of this approach by three use cases: 1) high bandwidth recordings of new speaker populations in geographically distributed locations, 2) recordings in adverse recording environments, e.g. hospitals, and 3) field recordings of endangered languages. The only infrastructure requirements are electricity for the equipment and an Internet connection.

**Index Terms:** speech database collection, web-based recording, web-based annotation, client-server system.

## 1. Introduction

The history of creating speech resources can be divided into three main phases:

- In the *direct observation phase*, researchers transcribed spoken languages by direct observation of native speakers of a given language; these transcriptions were then compiled as a speech resource.
- The *analog recording phase* added permanence: spoken language could now be stored for later processing, in general phonetic transcription, but also signal processing. Speech resources of this phase typically consist of collections of tapes or records in combination with printed or machine readable transcriptions.
- The *digital phase* introduced unlimited duplication of the recorded signals and powerful signal processing. The speech resources of this phase generally are the speech databases we know today, e.g. TIMIT [9], Switchboard [10], Macrophone [2] and SpeechDat [11], CGN [16], or others.

The major driving force for creating digital speech resources has been the development of speech technology, focusing primarily on languages with a large market potential. Other scientific areas, e.g. linguistic or phonetic research, or ethnological field work, profited enormously from this development, but contributed only marginally. The following quote on the motivation for the CGN corpus illustrates this very clearly [13]: "The fact that to date for Dutch few relevant language resources are available forms a serious complication for the advancement of Dutch language and speech technology. The present project seeks to ameliorate this situation."

It is now time to enter the "*web-based digital phase*" which is characterized by collaborative creation and exploitation of speech resources using industry standard network protocols (and devices). This approach will make the creation of speech resources much more flexible and efficient, and thus decouple the creation of such resources from commercial interests. Following this approach, a speech scientist will be able to create a modern speech corpus all by himself. By sharing this corpus with the speech community, others can contribute to this corpus by extending it with new recordings or enhancing it with additional annotations.

The remainder of the paper is structured as follows: Section 2 provides basic definitions and outlines the architecture for web-based speech resource creation systems. Section 3 describes tools that have been built for web-based speech recording and annotation, and section 4 presents three use cases. Section 5 discusses the approach.

## 2. Web-based resource creation

In this paper, we define a "less-resourced language" to be

- a language spoken by a specific speaker population in a given environment or a particular geographic distribution, for which very few or no speech resources exist.

This definition is intentionally broad. It includes 'classical' endangered languages, large but not yet sufficiently digitally covered languages, and highly specific subsets or varieties of large and otherwise well-covered languages.

### 2.1. Architecture

The WWW is based on the *client-server architecture*. A client, usually a browser, requests web pages or data from a server; the server either retrieves the requested pages from the local file system, a database management system, or computes it on the fly. The http-protocol organizes the flow of control between server and client.

The creation of speech resources comprises the following tasks: specification, recording, annotation, validation, documentation and distribution. The data consists of specification text, speech and sensor signal data, structured annotation and lexicon text, and documentation and legal texts (see [15] for details).

The tasks usually are performed sequentially: specification, then recording and annotation, then validation, and finally distribution; documentation is necessary for every task. Recording and annotation can be intertwined – as soon as the first recordings have been completed annotation may start – and both can be executed in a distributed manner.

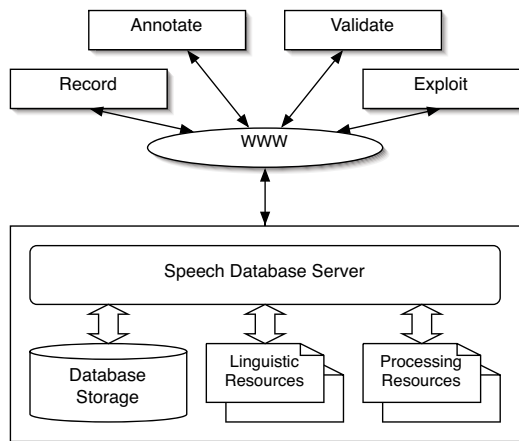


Figure 1: Mapping speech resource creation tasks and data to the client-server architecture of the WWW.

The tasks necessary for the creation of speech resources, the data accessed, and the flow of control can be mapped directly to the client-server model: the tasks are executed on the clients, the server organizes the workflow, provides the processing resources needed to perform the tasks, and stores the signal and annotation data in a central location (see fig. 1).

The charm of the client-server architecture is that client and server are only loosely coupled: they are implemented independently of each other, and their communication is based on a protocol. Hence new clients – providing new services or implemented on new devices – can be added easily if they adhere to the given protocol. Furthermore, a client may connect to more than one server, so that the same machine can participate in many different speech recordings with zero or no configuration effort. Finally, recordings can be performed in full signal quality – a significant advantage over telephone recordings with their limited voice quality.

Note that in the linguistic community, web-based *exploitation* of language corpora has long been available, e.g. for the BNC, CGN and many other corpora [1], [14]. There have also been a number of proposals for web-based *annotation* in corpus linguistics, e.g. [4], and in speech [5], [12]. However, only with recent developments of technology web-based *recording* has become a reality, so that now all speech resource creation tasks can be performed via the WWW.

## 2.2. Implementation

For client-server systems in the WWW, there exist two main implementation technologies: a) AJAX or 'Web 2.0', and b) embedded or standalone applications.

In AJAX, a client accesses the server using a page-internal programming language without having to rebuild the currently displayed page. AJAX has shown to be useful for dynamic web interfaces for *text* and *still image* applications – the technologically advanced Google maps or mail applications such as Zimbra are perfect examples.

For time-aligned data such as speech or video, embedded or standalone applications are more common, e.g. Java applets or Flash animations, and Java Web Start respectively. The programming languages applied here have built-in support for time-aligned data, so that writing signal processing applications is greatly facilitated. Applications of this type are downloaded

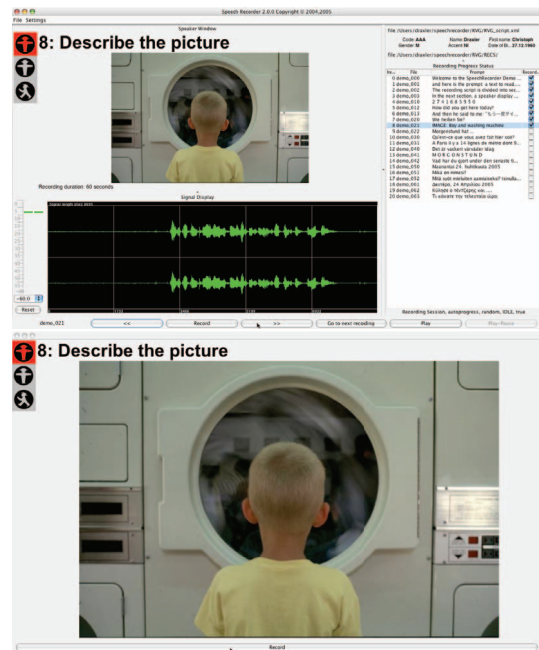


Figure 2: SpeechRecorder experimenter (top) and speaker (bottom) views

from the server and operate within a restricted environment for security reasons.

## 3. Tools

The following tools are web-based applications for the recording and annotation of speech. They have been developed at BAS.<sup>1</sup> Their common features are platform independence and making use of the web to store and retrieve speech and annotation data. Both tools are available from the BAS web site.

### 3.1. SpeechRecorder

SpeechRecorder features script-driven recording sessions, multi-channel recordings, multimedia prompts, support for multiple displays and different views, and recordings to the local disk or a remote server.

Speaker and experimenter see different views: the speaker view contains only instructions, the prompt proper and an indicator when to speak, the experimenter view displays a level meter, the waveforms of the current recording, and the list of all items of the session (fig. 2).

The recording of each item is written into a separate audio file. This file may be stored on the local machine, or transferred to a remote server. SpeechRecorder supports audio devices either through the operating system's audio interface, or via ASIO drivers. Hence, almost any audio hardware can be used.

### 3.2. WebTranscribe

WebTranscribe is an extensible framework for speech annotations. It implements the 'select – annotate – save' workflow in a Java Web Start application. The editing functionality is provided via editor plug-ins: the standard editor implements the SpeechDat orthographic transcription conventions; local-

<sup>1</sup>www.bas.uni-muenchen.de

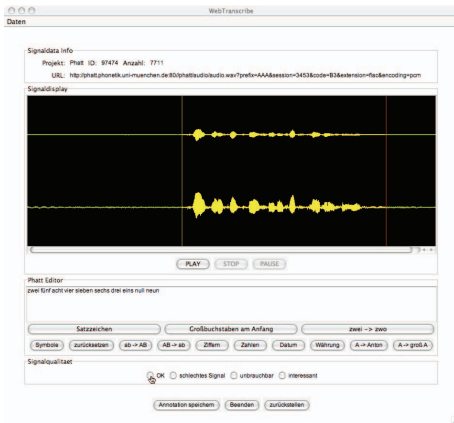


Figure 3: WebTranscribe annotation screen

ized versions for German, English, French, Russian, Hindi and Arabic have been implemented (see fig. 3).

### 3.3. Server

The servers for SpeechRecorder and WebTranscribe are implemented in Java and run by the Apache Tomcat web server. Signal data is stored in the server's local file system, annotation data in a relational database management system, e.g. PostgreSQL.

## 4. Use cases

The following use cases illustrate three very different aspects of web-based corpus creation, namely geographically distributed high bandwidth recordings, aphasic speech recordings in a clinical environment, and speech recordings for endangered languages. The first two use cases are actual projects, the third is (still) hypothetical.

### 4.1. Ph@ttSessionz and VOYS

The Ph@ttSessionz<sup>2</sup> database contains SpeechDat [17] and RVG-1 [3] compatible application oriented speech by adolescent speakers. The recordings were performed in public high schools in more than 40 cities in all dialect regions of Germany.

To achieve a consistent signal quality, the schools received a recording kit consisting of an M-audio mobile pre USB audio interface, a Beyerdynamic opus 54 headset, and an Audio Technica AT 3031 desktop microphone. The signal quality is 22.050 kHz sample rate, 16 bit linear quantization, and stereo.

At each site, an automated test was run to check the recording setup. For the production recordings, the experimenter, usually a teacher or a senior pupil, opened a recording session and entered the demographic speaker data into a web form which was then submitted to the server. The server then uploaded the recording script to the client and started the recording session. One item after the other was presented to the speaker automatically, and the recorded signal was transferred to the server immediately. The data upload to the server ran in a background process so that recordings could proceed without having to wait for the end of the data transmission.

Each school was asked to recruit 30 speakers, and the school received 200 Euro for these recordings. By December

<sup>2</sup>[www.phonetik.uni-muenchen.de/phattt](http://www.phonetik.uni-muenchen.de/phattt)



Figure 4: Phonlab MVP listening window for aphasic speech

2006, 865 speakers with a total of 110.000 utterances have been recorded.

VOYS<sup>3</sup> is a follow-up project. Its major aim is to prove that web-based recordings work across borders – the recordings will be performed in 10 recording locations in Scotland, the server will be at BAS in Munich.

The annotation of both Ph@ttSessionz and VOYS data is again performed via the WWW using WebTranscribe. The annotators work in the lab or at home, and their annotations are added to the server immediately.

### 4.2. Phonlab

Phonlab<sup>4</sup> is a diagnosis application for aphasic speakers. It implements the 'Munich Intelligibility Profile' [18] in a web application. Hospitalized aphasic speakers are asked to read words prompted by the Phonlab server, and the speech signal is transferred directly to the server. For diagnosis, speech therapists annotate the signal by listening to the speech signals and assigning a word from a list of up to 12 candidate words to the signal (fig. 4).

This annotation must meet strict constraints: listeners may listen to each signal only once, they may see the candidate words only after having heard the speech signal, and they may not modify their choice once it has been made. Furthermore, they must complete a session without interruption.

Phonlab is now in its pilot phase. 11 clinics in Germany are actively using the system.

The project also aims at creating a database of aphasic speech. Patients were asked if they would voluntarily contribute speech samples to this database. Quite a few patients did actually provide speech samples, so that eventually a database of German aphasic speech with a consistent set of demographic and meta-data will be available for clinical research.

### 4.3. Endangered languages

The classical example for less-resourced languages are endangered languages. One characteristic feature for these languages is that the speakers of such a language have no scientific knowledge of their language – the knowledge is with the experts in the scientific community. For some languages, there may be only a handful of experts worldwide.

Raw data, be it speech or video recordings, for these languages is recorded in field work. Such an expedition is a long-

<sup>3</sup>[www.phonetik.uni-muenchen.de/VOYS](http://www.phonetik.uni-muenchen.de/VOYS)

<sup>4</sup>[www.phonlab.de](http://www.phonlab.de)

term enterprise: the researcher has to become acquainted with the language of interest, earn the trust of the native speakers of the language, and set up the equipment according to the local infrastructure constraints. Today, this equipment consists of a power generator or solar panels, laptop computers, digital video cameras and audio recorders.<sup>5</sup>

Two very critical aspects of such field work are travel and communication: it may take days to reach a remote location, and once the destination has been reached, communication with the home lab is often difficult and slow, prohibitively expensive, or downright impossible.

Web-based corpus creation is an interesting option in regions with some form of digital communication infrastructure, e.g. community centers with Internet access, settlements with mobile phone networks, or within the reach of long-range wireless networks, e.g. WiMAX or satellites. In such regions, recordings can be uploaded to servers in the home lab and be immediately shared with the other experts of the language, e.g. for transcription and analysis. These results can then be used by the scientist in the field, e.g. to motivate the collection of additional material or discuss issues with the native speakers.

In areas without Internet connection, wireless networks can be set up with little effort, and thus web-based recording and annotation – on a local scale – is feasible.

## 5. Discussion

Web-based resource creation can dramatically cut down the cost of recording and annotation because it reduces the amount of travel and allows parallelizing the most time-consuming and expensive tasks. These distributed recordings can be made in full signal quality via multiple recording channels – something not feasible with telephone recordings.

Web-based resource creation supports the automation of the workflow; as soon as data is available it can be fed into subsequent processing steps without manual interaction. Downloading data and processing resources from a central server ensures that every client is working with the latest version, reducing the risk of data incompatibilities. The administration of data collection projects is greatly simplified because all data is held in one location, and can be monitored on-line and without delay. Finally, and perhaps most importantly, web-based resource creation easily adapts to technological progress – new devices or network technologies can be used once they are available without affecting the rest of the system.

A potential disadvantage of web-based resource creation is the limited control of the client. On the technology side, this is becoming less of a problem because modern USB devices allow retrieving information about device vendor and type, and even setting parameters. Other information, e.g. demographic data on the speaker, or the suitability of a particular recording environment, etc. may be less reliable.

The quality of the recording will still depend on the skill of the people performing it, and on the equipment used. Hence a lot of effort must go into instructing and motivating speakers and into devising easy to follow and fail-proof procedures.

The experience gained from both Ph@ttSessionz and Phonlab shows that with only minimal training pupils and hospital staff can perform complex recording tasks and produce high quality audio signals. The projects also show that the web-based production of speech resources allows the creation of resources that would not have been feasible with traditional approaches

<sup>5</sup>see [www.mpi.nl/LAN/](http://www.mpi.nl/LAN/) for details

under the given time or budget constraints.

## 6. Acknowledgments

Ph@ttSessionz was funded by BMBF grant no. 01IVB01.

## 7. References

- [1] Aston G., Burnard L., "The BNCHandbook: Exploring the British National Corpus with SARA", Edinburgh University Press, 1998
- [2] Bernstein J., Taussig K., Godfrey J., Macrophone: An American Speech Corpus for the PolyPhone project, ICASSP, 1994.
- [3] Burger S., Schiel F., "RVG-1 – A Database for Regional Variants of Contemporary German", LREC 1998, Granada.
- [4] Cunningham H., Tablan V., Bontcheva K., Dimitrov M., "Language Engineering tools for collaborative corpus annotation", Proc. of Corpus Linguistics, Lancaster, 2003
- [5] Draxler Chr., "WWWSigTranscribe – A Java Extension of the WWWTranscribe toolbox", LREC 1998, Granada
- [6] Draxler Chr., Jänsch K., Speech Recordings in Public Schools in Germany - the Perfect Show Case for Web-based Recordings and Annotation, LREC 2006, Genova.
- [7] Draxler Chr., Jänsch K., "SpeechRecorder – A Universal Platform Independent Multi-Channel Audio Recording Software", LREC 2004, Lisbon.
- [8] Draxler Chr., "WebTranscribe – An Extensible Web-Based Speech Annotation Framework", TSD 2005, Karlsbad
- [9] Garofolo J., Lamel L., Fisher W., Fiscus J., Pallett D., Dahlgren N., "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, NIST, 1986
- [10] Godfrey J., Holliman E., McDaniel J., "Switchboard: Telephone Speech Corpus for Research and Development, ICASSP 1992, San Francisco
- [11] Höge H., Draxler Chr., van den Heuvel H., Johansen F., Tropf H., "SpeechDatMultilingual Speech Databases for Teleservices: Across the Finish Line", Eurospeech 1999, Budapest
- [12] Ma X., Lee H., Bird S., Maeda K., "Models and Tools for Collaborative Annotation", LREC 2002, Gran Canaria
- [13] Oostdijk N., "The Spoken Dutch Corpus. Overview and first Evaluation.", LREC 2000, Athens
- [14] Oostdijk N., Broeder D., "The Spoken Dutch Corpus and its Exploitation Environment", 2003
- [15] Schiel F., Draxler Chr., "The Production of Speech Corpora", Bavarian Archive for Speech Signals, 2003
- [16] Schuurman I., Schoupe M., Hoekstra H., van der Wouden T., "CGN, an annotated Corpus of Spoken Dutch", Int'l Workshop on Linguistically Interpreted Corpora (LINC-03), 2003
- [17] Winski R., "Definition of Corpus, Scripts, and Standards for Fixed Networks", SpeechDat Report LE2-4001-SD1.1.1, 1997.
- [18] Ziegler W., Hartmann E., "Das Münchner Verständlichkeitsprofil (MVP) – Untersuchungen zur Reliabilität und Validität, Nervenarzt 64, pp. 653-658, 1993