BAS RESOURCES FOR PHONETIC LEARNER CORPORA

Christoph Draxler

Bavarian Archive for Speech Signals, Institute of Phonetics and Speech Processing,
Ludwig Maximilian University Munich, Germany
draxler@phonetik.uni-muenchen.de

Keywords: CLARIN, workflow, SpeechRecorder, WebMAUS, percy

1. CLARIN

CLARIN (Common Language Resource Infrastructure) is a European initiative to establish a stable research infrastructure for language and social sciences. In Germany, CLARIN-D consists of nine centres, three of which offer resources and services for speech: Institut für Deutsche Sprache (IDS) in Mannheim, Hamburger Zentrum für Sprachkorpora (HZSK), and Bayerisches Archiv für Sprachsignale (BAS) in Munich. The resources provided by these centres can be used free of charge by academics in Europe.

2. SPEECH DATABASES

A *speech database* is a well-structured collection of digital speech related data on three levels: primary data comprises audio, video, and sensor signal data; this data is immutable (except for format conversion). Secondary data are annotations and derived signal data; they always relates to some signal data, and they can be extended and modified, e.g. by adding annotation levels or correcting existing annotations. Finally, meta-data describe the database contents, ownership, license contracts, processing logs, validation reports, etc.

BAS [1] makes available its speech databases via its CLARIN repository, which may be searched using the virtual language observatory.

2.1. Phonetic learner speech databases

A phonetic learner speech database contains speech signals of non-L1 speakers with at least an orthographic transcription. It may also comprise derived signals such as f0 and formant data, which can be computed automatically, and a phonetic segmentation. This segmentation can be computed automatically, or be performed manually. The speech material in general consists of prompted items which were designed for learners of a language, e.g. minimal pairs, phrases with specific intonation patterns, exercises, etc.

2.2. Workflow

To establish a phonetic learner speech database, the workflow consists of the specification of the database structure and contents, recording audio data, annotating these recordings, and, finally, exploiting the database for phonetic experiments and statistical analyses (fig. 1).

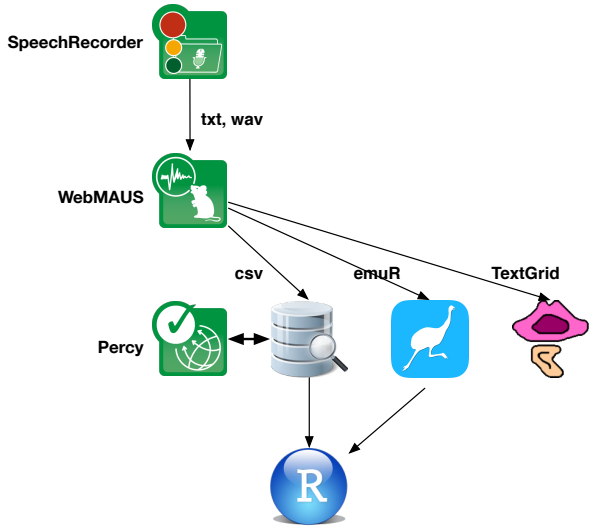

Figure 1: Schematic workflow with BAS and other tools

3. RECORDING: SPEECHRECORDER

SpeechRecorder is a multiplatform application for scripted audio recordings [4]. The contents and structure of a recording session are defined in an XML-formatted recording script. This script is divided into sequential sections, which in turn contain prompt items which may be randomized. Prompt items are either text, images or audio files – the latter two are useful for recording speech without text prompts, e.g. in dialectology, speak-after-me-exercises or with speakers who can't read, e.g. children.

SpeechRecorder supports multiple displays for

different views for speakers and recording supervisors. This facilitates performing recordings in non-standard environments because the speaker interaction can be separated from the controling device, e.g. in a car, class room, or a classical recording studio.

Each recording session consists of a separate directory on disk. Each utterance is written to a separate file; this audio file may be accompanied with a parallel text file containing the prompt item. The contents of the text file are defined in the recording script to allow a normalized form suitable for further processing, e.g. automatic segmentation.

SpeechRecorder is being ported to mobile devices, and it is embedded into the WikiSpeech system for online recordings [5].

4. ANNOTATION: WEBMAUS

WebMAUS (Munich Automatic Segmentation) is a suite of three tools provided as web services to the community. A web service is accessed either via forms or drag&drop interaction with a browser, via direct procedure calls in a terminal shell, or embedded into other applications. For the user, web services are very attractive because they do not need any software installation, all computing is performed on the server side, and the newest version of the software is being used by default.

The distinguishing feature of WebMAUS is that it not only uses acoustical models for the segmentation, but also applies rules that model the coarticulation patterns of a given language [8]. For example, the standard pronunciation of 'haben' (*have*) denoted in a lexicon would be /h a b @ n/, but in general will be produced as /h a b n/, /h a b m/ or even /h a m/. WebMAUS will attempt to align the signal with these alternatives and return the best match. The overall performance of WebMAUS comes close to that of human labellers, at a fraction of time.

WebMAUS takes as input pairs of audio file and its associated orthographic transcript text file. It currently supports 14 languages plus language-independent SAM-PA, and new languages are added once suitable speech databases become available.

The three WebMAUS services differ in their interface: the basic service simply requires a pair of audio and text file, plus the selection of a language; it returns a Praat TextGrid [2] or an Emu database file [10]. The general service features a large variety of input options, whereas the multiple version offers a drag&drop interface for entire directories of audio and transcript files, and returns an archive file with a segmentation for every input pair of files.

WebMAUS may also called from within the ELAN annotation editor [9, 6], and it is integrated into an experimental system to provided feedback to L2 learners in language learning applications.

The grapheme-to-phoneme converter G2P, on which WebMAUS is based, is now available as a web service of its own [7]. It takes as input an orthographic transcript and returns not only its canonic pronunciation, but also POS tags, syllabification, morphology and prosodic information.

5. EXPLOITATION: PERCY

In the context of phonetic learner databases, research questions often concern language interference, learning progress, pronunciation and intonation, and many other. These are often tested in perception experiments.

Percy is a framework for media-rich online experiments [3]. It is designed to run on any device supporting a modern browser, including mobile phones, tablets, computers, game consoles, and TV sets. In percy, the experimenter sets up an online experiment by specifying the experiment items, the associated audio files, and the allowed user input. The standard input is simply an empty text field, which is appropriate e.g. for items like 'Type the word you heard', or a series of buttons labelled with input options. Any other interaction element provided by HTML5 may be used, e.g. popup menus, sliders, etc., but this must be programmed by the percy administrator.

Once the experiment has been tested and approved, the experimenter sends a link to potential candidates, who then log in and perform the experiment. Large audiences can be reached very quickly, e.g. via mailing lists, social media, or other. Any user input is immediately saved to the server. This allows a real-time monitoring of progress and access to the data in the running experiment, e.g. for intermediate analyses, testing analysis procedures and scripts, etc. All data is held in a relational database which can be accessed directly from statistics applications such as R or SPSS.

6. SUMMARY

The three tools presented here are available free of charge. They are tailored to the needs of phoneticians and researchers working with audio data. To improve our tools we apply them in our daily work, integrate them into our curricula, and encourage users to provide feedback. In particular, web services will change the way researchers work, because they offer increased funtionality but at the same time reduce or even eliminate the need for software installation and maintenance.

7. REFERENCES

[1] BAS CLARIN-D online repository. http://hdl.handle.net/11858/00-1779-0000-000C-DAAF-B.

[2] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10), 341–345.

[3] Draxler, C. 2011. Percy – an HTML5 framework for media rich web experiments on mobile devices. *Proc. Interspeech* Florence, Italy. 3339–3340.

[4] Draxler, C., Jänsch, K. 2004. SpeechRecorder – a universal platform independent multi-channel audio recording software. *Proc. LREC* Lisbon, Portugal. 559–562.

[5] Draxler, C., Jänsch, K. 2008. Wikispeech – a content management system for speech databases. *Proc. Interspeech* Brisbane. 1646–1649.

[6] Kisler, T., Schiel, F., Sloetjes, H. 2012. Signal processing via web services: the use case WebMAUS. *Proc. of Digital Humanities* Hamburg. pp. 30–34.

[7] Reichel, U. D., Kisler, T. 2014. Language-independent grapheme-phoneme conversion and word stress assignment as a web service. In: Hoffmann, R., (ed), *Elektronische Sprachverarbeitung 2014* volume 71 of *Studientexte zur Sprachkommunikation*. Dresden, Germany: TUDpress 42–49.

[8] Schiel, F. 1999. Automatic phonetic transcription of non-prompted speech. *Proc. ICPhS* San Francisco. 607–610.

[9] Sloetjes, H., Russel, A., Klassmann, A. 2007. ELAN: a free and open-source multimedia annotation tool. *Proc. Interspeech* Antwerp. 4015–4016.

[10] Winkelmann, R., Raess, G. 2014. Introducing a web application for labeling, visualizing speech and correcting derived speech signals. Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* Reykjavik, Iceland. European Language Resources Association (ELRA).