Creating, Managing and Analysing Speech Databases using BAS Services and Emu: A Hands-On Tutorial

Christoph Draxler, Florian Schiel
Bavarian Archive for Speech Signals
Ludwig Maximilian University Munich, Germany
draxler|schiel@bas.uni-muenchen.de

November 14, 2017

Abstract

The creation of speech databases for spoken language research and development is a time-consuming and largely manual task. In this tutorial we present an optimized and tested workflow comprising the specification, recording, (automatic) transcription, (automatic) segmentation, and effective analysis of a corpus of spoken language. We will demonstrate how to use a) automatic tools, b) an effective integrated management system (EMU-SDMS) to organize and analyse speech data, and c) crowdsourcing wherever possible to speed up the process. We will also show how to apply established tools to under-resourced languages, thus facilitating access to these languages.

Description of the Tutorial

The tutorial consists mainly of live demonstrations and is divided into the following parts:

1. Scripted recording of speech with SpeechRecorder
2. Automatic transcription using ASR
3. Verification of transcriptions using crowdsourcing
4. Automatic speech chunking of very long recordings
5. Phonetic segmentation of speech using WebMAUS
6. The Emu-SDMS system to manage/analyse the corpus

SpeechRecorder: Scripted Recording of Speech

SpeechRecorder is a platform independent multi-channel audio software for the scripted recording of speech [DJ04]. *SpeechRecorder* features text, image, audio and video prompts, and it supports different views using multiple displays. Each recording is automatically saved to a separate file, thus eliminating the need for signal editing.

Recent additions to the software include the creation of annotation template files that facilitate the (automatic) transcription of the recorded speech, and continuous recording, e.g. for linguistic field work.

`www.bas.uni-muenchen.de/forschung/Bas/software/speechrecorder/`

ASR: Automatic Transcription

We demonstrate how to obtain orthographic transcripts of non-prompted speech automatically using the web service *runASR*. Currently supplied as a beta system, *runASR* features fully automatic transcriptions in over 130 languages and language variants. The service also allows automatic speaker diarization for some major languages.

`clarin.phonetik.uni-muenchen.de/BASWebServices/\#!/services/ASR`

OCTRA, EMU-webapp: Verification of Transcriptions

In general, recordings must be checked for quality and contents, especially when they were unsupervised (e.g. web-based recordings). For read material, this check may be as simple checking whether the utterance was read correctly, for spontaneous speech the automatic recognition of the utterance has to be checked. While checking the contents of the signal, noise markers may be inserted, and the overall quality can be assessed.

Orthographic transcriptions/corrections require only basic writing skills, they do not need highly trained specialists. Furthermore, such a task may be performed via the web. On the one hand, this facilitates massive parallel transcriptions using a large workforce of transcribers, on the other hand it allows accessing expert knowledge in less-resourced languages.

Recent new tools include *OCTRA* (orthographic editor) or *Emu webApp* for labelling speech [WR14], which allow task-directed work with minimal data management overhead.

`ips-lmu.github.io/EMU-webApp/`

Chunker: Automatic Pre-Chunking of long Speech Signals

Nowadays very long recordings – e. g. hour-long ethnographical interviews with informants – have become common-place. The processing time required

by the automatic segmentation WebMAUS (see below) grows quadratically with signal duration. Automatically breaking down long signal files using the *Chunker* web service into short fragments greatly reduces the total processing time by a) allowing parallel segmentations of the fragments, and b) limiting the effects of the quadratic cost of computation.

`clarin.phonetik.uni-muenchen.de/BASWebServices/\#!/services/Chunker`

WebMAUS: Phonetic Segmentation of Speech

Based on the orthographic transcript, the public webservice *MAUS* (Munich AUtomatic Segmentation) can be used to derive a word, syllable or phone segmentation and labelling [KSS12].

MAUS combines an acoustical model based on HTK and a statistical predictor for pronunciation variation based on an Markov process to calculate the most likely segmentation in word and phone units [Sch99, Sch15].

MAUS is currently available for 32 languages or language variants as well as a language-independent mode and can be accessed via a REST API or an interactive web interface.

In the tutorial we will demonstrate how a complete hierarchical EMU database can be created automatically based on the recordings and orthographic transcripts; we will also demonstrate how several of our webservices can be chained into a pipeline to avoid unneccessary upload times.

```
clarin.phonetik.uni-muenchen.de/BASWebServices/\#!/
  services/WebMAUSGeneral
```

Managing/Analysing a Speech Database: Emu-SDMS

In this final part of the tutorial we will briefly introduce the Emu speech database management system (EMU-SDMS). We will demonstrate how to load and analyse an Emu DB, how to create a EMU DB from scratch, how to call webservices from within the Emu-SDMS system to create new annotation layers and how to apply speech signal analysis tools to the stored signals (such as fundamental frequency, formants etc.).

`ips-lmu.github.io/EMU.html`

Relevance of the Tutorial

At the BAS in Munich, we are constantly working on optimising the workflow of speech database creation. In this tutorial we will go through the entire process from recording via (automatic) transcription, segmentation, and ingesting the data into a state-of-the-art database management system,

presenting and using automatic and crowdsourcing tools for a consistent and high-quality result.

We feel that such a tutorial is extremely useful for researchers who start to work with empirical speech data, especially from fields such as engineering, linguistics, sociology, dialectology, and ethnology, and for speech technology researchers and developers working on under-resourced languages.

At LREC 2018 we will for the first time demonstrate the new integration of BAS web services, including the web-based ASR services and the Emu Speech Database System.

Tutorial Logistics

Registered participants will be able to download a small-size (< 10 MB) collection of audio files from our web site during the tutorial and apply the demonstrated operations in parallel on their laptops. Since we exclusively apply webservices and web APIs, no software has to be installed in participants' laptops.

Presenters

Christoph Draxler

Christoph Draxler is a computer scientist plus linguist by education and has been working in the field of speech databases at the Institute of Phonetics and Speech Communication in Munich since 1991. He currently is the project manager of the Munich CLARIN center, in collaboration with Florian Schiel. Christoph Draxler and Klaus Jänsch have developed the speech recording tool SpeechRecorder and its web-based extension WikiSpeech. Christoph now focuses on crowdsourcing for speech transcription and annotation, and has developed the web-based perception experiment software Percy.

Florian Schiel

Florian Schiel received his Dipl.-Ing. and Dr.-Ing. degrees from the Technical University in Munich in 1990 and 1993 respectively, both in electrical engineering. His doctoral thesis deals with automatic speaker adaptation in ASR. Since 1993 he was mainly affiliated to the Institute of Phonetics, Ludwig-Maximilians-Universität Munich (LMU), leading the German VERBMOBIL, SmartKom, BITS and SmartWeb project groups. In 1994 and 1997 he spent 6 months each as a research fellow at the International Computer Science Institut (ICSI), Berkeley, California.

Currently he is CEO for BAS Services and is tenured as a senior researcher at the new Institute of Phonetics and Speech Processing at LMU.

Audience Information

The tutorial addresses mainly but not exclusively researchers in the field of phonetics, speech technology research and development, general linguistics, dialectology, sociology and ethnology. University curricula in general do not feature such hands-on tutorials, and we therefore believe that this tutorial thus fills a gap.

References

[DJ04] Christoph Draxler and Klaus Jänsch. SpeechRecorder – a universal platform independent multi-channel audio recording software. In *Proc. LREC*, pages 559–562, Lisbon, Portugal, 2004.

[KSS12] Thomas Kisler, Florian Schiel, and Han Sloetjes. Signal processing via web services: the use case WebMAUS. In *Proceedings Digital Humanities*, pages 30–34, Hamburg, 2012.

[Sch99] F. Schiel. Automatic phonetic transcription of non-prompted speech. In *Proc. ICPhS*, pages 607–610, San Francisco, 1999.

[Sch15] F. Schiel. A statistical model for predicting pronunciation. In *Proceedings ICPhS*, Glasgow, United Kingdom, August 2015.

[WR14] Raphael Winkelmann and Georg Raess. Introducing a web application for labeling, visualizing speech and correcting derived speech signals. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA).