



Phonetic/linguistic Web Services at BAS

Thomas Kisler, Florian Schiel, Uwe D. Reichel, Christoph Draxler

Bavarian Archive of Speech Signals,
Ludwig-Maximilians-Universität München, Germany

kisler|schiel|reichelu|draxler@bas.uni-muenchen.de

Abstract

We present recent developments in the collection of phonetic-linguistic web services provided by the Bavarian Archive of Speech Signals (BAS). The BAS back end web services are REST based and can be easily integrated into user applications. Several public web interfaces have been implemented that utilize these back end services to provide easy-to-use access to high-end linguistic and phonetic processing (front end services). In this show&tell we demonstrate the latest front end services of BAS: automatic phonetic segmentation & labelling using the MAUS technique (14 languages), text-to-phoneme conversion (13 languages), automatic phonetic transcription (6 languages), phonetic syllabification (13 languages), and speech synthesis.

Index Terms: automatic segmentation, text-to-phoneme, syllabification, RESTful web service, web interface

1. Introduction

Service-oriented architectures allow for highly decoupled systems, where the back end (web services) can be utilized by an arbitrary front end (web interface, other application, etc.) to carry out specific tasks. The front end provides the necessary data to the back end, which then carries out the actual processing and returns the result to the caller. During the last 4 years a number of phonetic/linguistic tools developed by the Bavarian Archive of Speech Signals (BAS) have been implemented as RESTful web services, following the CLARIN core data model for web services [1]. In our case we utilize REST as an envelope for a process oriented view (in contrast to the the resource oriented view of REST), which results in remote procedure call (RPC) styled REST services. For instance the well-known video annotation tool ELAN [2] uses such a RESTful call to automatically segment a piece of transcribed speech signal. The successful application of web services requires a highly available server infrastructure and a strict adherence to the published interface definition. The BAS web services (amongst others within the CLARIN consortium, see e.g. [3]) define their interfaces both via a Component Metadata file (CMDI, [4]), which in turn follows the core model for CLARIN web services, and the standard description through the Web Application Description Language (WADL). Application users or the application itself can look-up these definitions and form appropriate REST calls to initiate processing on BAS servers.

To illustrate the usefulness of such an approach and allow access to these services to non-technical users, the BAS has implemented several web interfaces as a front end to the RESTful RPC styled back end services. These web interfaces allow users without technical knowledge and without programming skills to apply the BAS services to their locally stored data, simply by using a web browser. It turns out that there exists a huge

and exponentially growing, international demand for these web interfaces. The BAS therefore strives to extend its RESTful services and web interfaces to cover a growing number of service types and languages.

In this show&tell we demonstrate the following services:

- WebMAUS: automatic phonetic segmentation & labelling using pronunciation prediction based on the orthographic transcript ([5]) for 14 languages
- G2P: automatic grapheme-to-phoneme conversion for 13 languages ([6])
- Forced alignment based on SAM-PA transcripts for small and endangered languages (language independent)
- WebMINNI: automatic phone recognition & segmentation for 6 languages
- Pho2Syl: automatic phonetic syllabification for 13 languages
- ChunkPreparation: pre-processing of chunk-segmented and transcribed large video recordings for a more efficient WebMAUS segmentation
- MaryTTS: a free German synthesis with 4 voices based on the MARY TTS system ([7])
- COALA: automatic generation of CMDI-based metadata files to simplify the integration of speech and multimedia corpora into CLARIN repositories.

In the remaining paper we give a very short functional description for the most relevant services that will be demonstrated, followed by relevant URLs to access the web services and web interfaces.

2. Web Interfaces

2.1. WebMAUS – automatic segmentation

Input: speech signal file and orthographic/canonical phonemic transcript

Output: word segmentation or phonetic segmentation

Encodings: SAM-PA, IPA, manner- / place-of-articulation

Batch processing via web-interface: yes

Description: Based on the orthography a statistically weighted lattice for the predicted phonetic pronunciation is calculated, and then combined with language-specific HMMs to estimate the most likely pronunciation and segmentation

Demonstration: Batch processing corpora of signal+text pairs for 14 different languages. Results can be inspected online via the EmuLabeller[8].

2.2. G2P – automatic text-to-phoneme conversion

Input: orthographic transcript or word list

Output: most likely canonical pronunciation, syllable boundaries, lexical stress, for English and German additionally morph segmentation and classification. Phoneme-letter alignment supported.

Encodings: SAM-PA

Batch processing via web-interface: yes

Description: The transcription and syllabification is carried out by C4.5 decision trees trained on grapheme, resp. phoneme windows of varying length. For German and English POS and morph labels are used as additional features. Word stress is assigned by Bayes classification in a learning-by-analogy framework.

Demonstration: Batch processing of text corpora with several output selections and file formats for 13 different languages.

2.3. WebMINNI – automatic phone recognition

Input: speech signal file

Output: most likely phonetic transcript and segmentation

Encodings: SAM-PA, IPA, manner- / place-of-articulation

Batch processing via web-interface: yes

Description: While the WebMAUS services require textual input, this service performs a phonetic recognition based only on the speech signal applying phonetic HMMs and a phonotactic bi-gram model.

Demonstration: Batch processing of signals files of 6 different languages.

2.4. Pho2Syl – phonetic syllabification

Input: canonical or spontaneous speech transcription (BAS Partiture Format)

Output: syllabified transcription. free choice, whether or not syllable and word boundaries are synchronized.

Encodings: SAM-PA

Batch processing via web-interface: yes

Description: Syllabification is carried out by C4.5 decision trees or rule in terms of sonority hierarchy.

Demonstration: Processing of BAS Partiture Format files of 13 different languages.

3. Useful URLs

- **BAS web interface:** <http://clarin.phonetik.uni-muenchen.de/BASWebServices>
- **BAS web services metadata (CMDI):** https://clarin.phonetik.uni-muenchen.de/BASRepository/WebServices/BAS_WebServices.cmdi.xml
- **BAS web services help page:** <http://clarin.phonetik.uni-muenchen.de/BASWebServices/services/help>
- **CLARIN Component Metadata model (CMD):** <http://www.clarin.eu/content/component-metadata>
- **CLARIN ERIC:** <http://clarin.eu/>
- **CLARIN-D:** <http://de.clarin.eu/>

4. Target Users

The BAS services are most popular among phoneticians and (field) linguists who need to speed up the documentation and annotation process of their speech recordings. Recently, we also received feedback from colleagues of other disciplines, such as:

ethnology, anthropology, sound change, socio-phonetics, historical linguistics, dialectology, speech recognition, speech synthesis, video indexing, and computer assisted learning. As only one of many examples, anthropologists apply the WebMAUS service to the soundtracks of video recordings depicting interaction with speakers of endangered languages. Since for most of these languages no writing system exists, these colleagues use the language independent forced alignment or the WebMINNI service to time-align their video recordings ([9]).

Regarding further processing of results, the WebMAUS service allows the transformation of results into the new emuDB database format, which allows complex database queries in Emu Query Language (EQL), and thus highly sophisticated statistical analysis chains in R language ([10], R packages emuR and wrassp). See as an example the use case 'Field researcher in Australia ...' in [11].

5. Acknowledgements

We thank the European CLARIN ERIC and the German CLARIN-D consortium funded by the German Ministry of Science and Education for establishing the infrastructure, as well as hundreds of users' valuable feedback to improve our services.

6. References

- [1] M. Windhouwer, D. Broeder, and D. Van Uytvanck, "A CMD core model for CLARIN web services," in *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation*, 2012, pp. 41–48.
- [2] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2006.
- [3] Clarin web services. Last accessed: 2015-04-16. [Online]. Available: <http://clarin.eu/content/web-services>
- [4] BAS web service CMDI. Last accessed: 2015-04-16. [Online]. Available: http://clarin.phonetik.uni-muenchen.de/BASRepository/WebServices/BAS_WebServices.cmdi.xml
- [5] F. Schiel, "Automatic Phonetic Transcription of Non-Prompted Speech," in *Proc. of the ICPHS*, San Francisco, August 1999, pp. 607–610.
- [6] U. D. Reichel, "PermA and Balloon: Tools for string alignment and text processing," in *Proc. Interspeech*, Portland, Oregon, 2012, p. paper no. 346.
- [7] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," in *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, August 29 - September 1, Perthshire, Scotland*, 2001, pp. 131–136.
- [8] R. Winkelmann and G. Raess, "Introducing a web application for labeling, visualizing speech and correcting derived speech signals," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014.
- [9] J. Strunk, F. Schiel, and F. Seifart, "Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014.
- [10] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014, last accessed: 2015-04-16. [Online]. Available: <http://www.R-project.org>
- [11] Use case 'Field researcher in Australia ...'. Last accessed: 2015-04-16. [Online]. Available: <http://clarin.phonetik.uni-muenchen.de/BASWebServices/#/help>