INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

Online Speech Processing and Analysis Suite

Wikus Pienaar, Daan Wissing

Centre for Text Technology (CTexT)
North-West University
Potchefstroom, South Africa
wikus.pienaar@nwu.ac.za, daan.wissing@nwu.ac.za

Abstract

Proper phonological analyses, descriptions and explanations as well as gaining insight into language variation and change rely heavily upon ample and trustworthy phonetic data. Our **Online Speech Processing and Analysis Suite** is a positive development in just this direction.

Index Terms: acoustic analysis, automatic speech segmentation, Database Searcher, Online application, Phonetic Acoustic Analysis System, normalization

1. Introduction

To be successful, extensive acoustical analyses of speech data necessitates pre- and post-processing of recorded speech, including automatic segmentation and transcription of speech files, retrieval of acoustical data and vowel normalization. These processes can be time-consuming and labor-intensive.

In this paper, we propose an online processing and acoustic analyses suite to assist and speed up these procedures. An indexed speech database is included for the enhancement of the search for specific phonemes and patterns.

2. Background

Initially the pre- and post-processes consisted of four separate offline applications. In this section, a short description of each of these offline applications will be presented, followed by a more complete discussion in Section 3.

Firstly, the **automatic speech segmentation application** (ASS) consists of a graphical user interface (GUI) and a system developed by Van Niekerk [1], specifically intended for under-resourced languages. Wissing *et al.* [2] describes this system as follows:

The Default & Refine algorithm of Davel and Barnard [3] is used to extract grapheme-to-phoneme rules from the existing Afrikaans pronunciation dictionary of Davel and De Wet [4] to convert the orthographic transcriptions into their phonetic forms. Hidden Markov Models is then estimated to find an alignment between the phone sequence and the audio sequence using forced alignment (Van Niekerk and Barnard [5]).

Secondly, **PHONAAS** (Phonetic Acoustic Analysis System) [6], consisting of a simple GUI and a combination of Praat and R scripts that runs in the backend for extracting formants and creating formant-graph data from recorded sound files.

Thirdly, **W-NORM** [7] does vowel normalization and plotting. It is an offline application, based on the online Vowel Normalization and Plotting Suite, developed by Thomas and Kendall [8].

A combination of Perl and Praat scripts are used in the **Database Searcher** to retrieve start and end times of phonemes. The data is then stored in a SQL Lite database for extraction of data with a number of parameters, as selected in the GUI.

3. Design and implementation

In order to create an online web-application that is both user friendly and easily accessible, we developed an **application programming interface** (API). This suite is comprised of each of the four applications mentioned in (2).

3.1. API

The API handles all incoming requests by validating the parameters and headers and then sends the parameters and data to the applicable scripts.

A REST API is used for this application. Masse [9] describes **Representational State Transfer** (REST) as a technical description of how the World Wide Web works as well as a REST API as a type of web server which enables access to resources and models of a system's data and functions.

3.2. Applications

3.2.1. Automatic Speech Segmentation

The workflow for the Automatic Speech Segmentation system is as follows:

- Select the sound files (in .wav format) to process. For long recordings, it works better to split the files into smaller chunks. By doing so, a snowball effect will be prevented in cases of erroneous segmentation.
- Add an exact orthographic transcription in the text box for each of the recordings.
- Choose if silences must only be added on full stops and commas, or automatically.

By clicking the *Process* button, the sound files, orthographic transcription and silence parameter are sent to the API. After successful validation it is sent to ASS (Van Niekerk [1]).

After processing, a zipped file containing the phonetic transcriptions (in TextGrid format) and the sound files (in .wav format) are returned.

The returned TextGrid(s) consists of five tiers:

- The phonetic transcription with phoneme boundaries
- Syllable boundaries
- Word boundaries with orthographic transcription
- Phrase boundaries

- Original orthographic transcription

3.2.2. PHONAAS

In PHONAAS the user can select the following:

- Sound file in .wav format and matching TextGrid with phoneme boundaries
  - Type of phonemes to extract formant data from: monophthongs, diphthongs, long vowels
  - Additional data in the output (optional): Centre of Gravity (of fricatives), duration, F0, HNR, intensity (of vowels)
  - Formant settings (optional): maximum formant value in Hertz (default 5500 Hz) and number of formants (default 5)
  - Create formant graphs of vowels (optional)

The sound file, TextGrid and parameters are sent to the API for validation and then to Praat scripts for processing. If the option for graphical output is selected, the results of the Praat scripts are sent to an R script that creates the formant graphs.

Acoustic results in Microsoft Excel format as well as plain text format and the phoneme graphs (if applicable) are returned from the system.

3.2.3. W-NORM

The following options are available in W-NORM:

- Input data containing formant values for the applicable vowels
- Selection of type of results: individual vowels, speaker means or group means
- Selection of normalization method (Bark, Lobanov etc.)
- Selection of the standard deviation type and multiplier
- Choosing of labels, colors, size, scaling and output format of the plots

After setting the above mentioned options, the API first performs validation on the input data and parameters and then sends it to an R script (utilizing the R package *Vowels 1.2* developed by Thomas and Kendall in 2018 [10] for the normalization process).

An image of the normalized plot of the vowel space and the plotted data in Excel are returned to the user after successful processing.

3.2.4. Database Searcher

The data for the Database Searcher is pre-processed by using Perl and Praat scripts for the retrieval of start and end times for n-grams (1 gram – 8 gram) from the phonetic transcriptions created by the system described in 3.2.1 above, as well as for every word in the orthographic transcription. These start and end times are then stored in a simple SQL database. The sound files and TextGrids to be used for data retrieval are stored on the server.

For the search and retrieval of data the following options can be selected:

- Dataset(s) to search in.
- Search pattern:
  - Word-level search, including the use of SQL wildcard characters
  - N-gram phoneme search
- Additional options: gender of speaker, age group, ethnic group

The API validates the parameters and sends it to the Praat and Perl scripts. The retrieved data consists of separate sound files for each matched phoneme, syllable or word, and one sound file where the separate sound files are combined into one file.

4. Conclusions and future work

In this paper, we proposed an online processing and acoustic analyses suite to assist and speed up the pre- and post-processing of speech recordings.

ASS is currently only available for Afrikaans. Future work includes the expansion of the speech models for more languages, as well as the continuous expansion of the speech database.

5. Acknowledgements

This paper was made possible with the support from the South African Centre for Digital Language Resources (SADiLaR), a research infrastructure established by the Department of Science and Technology of the South African government.

Ian Bekker as voice artist in the video, Rico Koen for his assistance with the front- and backend of the website and Benito Trollip for his contribution.

6. References

[1] D. R. Van Niekerk, *Automatic speech segmentation with limited data,* Master of Engineering, North-West University, Potchefstroom Campus, 2009.

[2] D. P. Wissing, W. Pienaar and D. R. Van Niekerk, "Palatalisation of /s/ in Afrikaans," *Stellenbosch Papers in Linguistics Plus,* vol. 48, pp. 137-158, 2015.

[3] M. Davel and E. Barnard, "Pronunciation prediction with default and refine," *Computer Speech and Language,* vol. 22, no. 4, pp. 374-393, 2008.

[4] M. Davel and F. De Wet, "Verifying pronunciation dictionaries using conflict analysis," in *Proceedings of Interspeech*, Tokyo, Japan, 2010.

[5] D. R. Van Niekerk and E. Barnard, "Phonetic alignment for speech synthesis in underresourced languages," in *Proceedings of Interspeech*, Brighton, UK, 2009.

[6] W. Pienaar and D. P. Wissing, "PHONAAS: Phonetic Acoustic Analysis System," North-West University; Centre for Text Technology (CTexT), 30 June 2015. [Online]. Available: https://hdl.handle.net/20.500.12185/366. [Accessed 23 March 2019].

[7] W. Pienaar and D. P. Wissing, "W-NORM, A graphical user interface for plotting and normalisation," North-West University; Centre for Text Technology (CTexT), 30 June 2015. [Online]. Available: https://hdl.handle.net/20.500.12185/385. [Accessed 23 March 2019].

[8] E. R. Thomas and K. Tyler, "The Vowel Normalization and Plotting Suite," 11 November 2015. [Online]. Available: http://lingtools.uoregon.edu/norm. [Accessed 27 March 2019].

[9] M. Masse, REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces., Sebastopol: O'Reilly Media Inc, 2011.

[10] E. R. Thomas and K. Tyler, "Vowels: Vowel Manipulation, Normalization, and Plotting in R. R package, v.1.2," 5 March 2018. [Online]. Available: http://lingtools.uoregon.edu/norm/. [Accessed 27 March 2019].