Guide to the Supplementary Information for the paper

"Weak biases emerging from vocal tract anatomy shape the repeated transmission of vowels"

The Supplementary Information for this paper is structured in two main categories described below, and a table summarizing the files can be found on the next page.

1. Summary of the supplementary files

File name	Type	Size	Description
SIGuide.pdf	**PDF**	318 KB	This guide to the Supplementary Information.
SupplementaryInformationFile.pdf	**PDF**	5.9 MB	The Supplementary Information document containing Supplementary Figures, the Supplementary Tables and the Supplementary Methods.
SupplementaryData1.txt	TAB-separated **CSV**	178 KB	Cross-linguistic corpus of vowel realizations.
SupplementaryData2.csv	TAB-separated **CSV**	3 KB	Participant info.
SupplementaryData3.tsv	TAB-separated **CSV**	273 KB	Midsagittal hard palate shape (MSHPS) tracings.
SupplementaryData4.tsv	TAB-separated **CSV**	77 KB	"Classical" anthropological measures for the ArtiVarK participants.
SupplementaryData5.csv	TAB-separated **CSV**	21.8 MB	Simulation results for the five MSHPSs with multiple replications.
SupplementaryData6.csv	TAB-separated **CSV**	7.7 MB	Simulation results for all MSHPSs (one replication) + Bézier parameters describing the MSHPS.
SupplementaryResults1.html	**HTML**	30 MB	All the statistical analyses and plots, except for those using the 3D intra-oral scans.
SupplementaryResults2.html	**HTML**	11.8 MB	The statistical analyses and plots using the 3D intra-oral scans.
SupplementarySoftware1.zip	**ZIP** archive (multiple)	6.6 MB	ZIP archive containing all the data and Rmarkdown script needed to reproduce SupplementaryResults1.html.
SupplementarySoftware2. zip	**ZIP** archive (multiple)	11 KB	ZIP archive containing the Rmarkdown script and some of the data (but not all, due to privacy concerns) needed to reproduce SupplementaryResults2.html.
SupplementarySoftware3. zip	**ZIP** archive (multiple)	3.1 MB	ZIP archive containing the scripts, programs and configuration files needed to run the simulations.

2. The supplementary information

This information is needed for a full understanding of the paper and is composed of the following:

2.1. The Supplementary Information File

This is a **PDF document** containing the *Supplementary Figures 1 to 11* (pages 1-11), the *Supplementary Tables 1 to 4* (pages 12-17), and the *Supplementary Methods* (pages 18-20), which are directly referenced from the main paper using "Supplementary Figure X", "Supplementary Table Y" and "Supplementary Methods".

2.2. The Supplementary Results

These are two self-contained **HTML documents** (i.e., all their images, styling and interactive content is included in them, with no need for accessory files) that can be shown by any modern web browser. These documents are directly referenced from the main paper using "*Supplementary Results 1*" and "*Supplementary Results 2*" and contain all the plots and statistical analyses (some not referenced directly from the paper due to space constraints). The first is by far the larger containing most of the analysis, while the second contains only a subset of the analysis based on data that cannot be made publicly available at this time.

Therefore, while "Supplementary Results 1" can be reproduced using "Supplementary Software 1" (see below), "Supplementary Results 2" cannot be reproduced using "Supplementary Software 2" because a key input dataset is not available. This missing dataset is the intra-oral 3D scanning data which, due to its very high resolution, may be potentially used to faithfully reconstruct the inside of our participants' mouth, which could lead to their identification or to revelation of private data.

2.3. The Supplementary Data

These six files contain the primary data used in the analyses reported in the paper (and needed to run the *Supplementary Software*); **Supplementary Data ID** is of the form "Supplementary Data X" and is used to refer to these data from the paper; **File name** is the name of the actual file as it is in the `data` directory (see *Supplementary Software* for details); **Description** is a brief summary of the format and content.

Supplementary Data ID	File name	Description
Supplementary Data 1	`BeckerVowelCorpus.txt.xz`	Contains the cross-linguistic vowel corpus from ref[1], as freely provided by Márton Soskuthy in his GitHub repository (https://github.com/soskuthy/u-fronting). Despite its extension (`.txt`), it really a TAB-separated file with column headers on 1st row. Pease note that in the `data` directory this file is XZ-compressed.
Supplementary Data 2	`participant_info.csv`	Contains various participant info such as the sample they come from (implicitly coded by the first letter of the ID: "A" for ArtiVarK and "T" for ref[2]), the original ID in the original sample, as well as sex, age and broad ethno-linguistic group. It is a TAB-separated file.
Supplementary Data 3	`hard-palate-tracings.tsv.xz`	Contains the tracings of all hard palate shapes, as reported in ref[3], namely for each participant (ID) and tracing we give the (x,y) coordinates of a tracing point, in sequence. It is a TAB-separated file. Pease note that in the `data` directory this file is XZ-compressed.
Supplementary Data 4	`trad_measures.tsv`	Contains the "classic" anthropological measures for the ArtiVarK participants, one measure per column. It is a TAB-separated file.
Supplementary Data 5	`HPshape-replicated-chains.csv.xz`	Contains the data for all the five MSHPSs with replications, giving the participant ID ("condition"), the replication, the chain generation, the vowel, the "seed" acoustics (F1s-F5s), the current target (F1t-F5t), the realized sounds (F1-F5), the articulatory parameter values used to produce this sound, and the acoustic "error" between realization and target (F1e-F5e). It is a TAB-separated file. Pease note that in the `data` directory this file is XZ-compressed.
Supplementary Data 6	`HPshape-nonreplicated-chains.csv.xz`	Contains the data for all MSHPSs with only one replication, giving the participant ID ("condition"), the replication (here, just one), the chain generation, the vowel, the "seed" acoustics (F1s-F5s), the current target (F1t-F5t), the realized sounds (F1-F5), the articulatory parameter values used to produce this sound, the acoustic "error" between realization and target (F1e-F5e) and the four Bézier parameters describing the participant's MSHPS. It is a TAB-separated file. Pease note that in the `data` directory this file is XZ-compressed.

3. The supplementary software

These contain the software, scripts and data needed to reproduce the simulations, analyses and plots on which this paper is based.

3.1. Supplementary Software 1

This is a `ZIP` archive containing the data, `Rmarkdown` script and other ancillary files needed to reproduce the *Supplementary Results 1* document (please note that due to various random components or linked to software version, results might differ slightly):

- The `data` directory contains the input files (`BeckerVowelCorpus.txt.xz`, `hard-palate-tracings.tsv.xz`, `HPshape-nonreplicated-chains.csv.xz`, `HPshape-replicated-chains.csv.xz`, `participant_info.csv` and `trad_measures.tsv`) described above and also available as Supplementary Data; please note that they are TAB-separated CSV files (despite varying extensions) and that some of these files are XZ-compressed to save space (those have an extra `.xz` extension).
- `apa.csl` and `bibliography.bibtex` are needed for the bibliography cited in the *Supplementary Results 1* document, being the APA citation style and the BibTeX bibliography, respectively.
- `landmarks_definition.jpg` and `landmarks_definition.tif` are the low- and high-quality versions (respectively) of illustrative anatomical vocal tract landmarks.
- `SupplementaryResults1.Rmd` is the actual `Rmarkdown` script.

To compile the *Supplementary Results 1*, we suggest a possible workflow as follows:

1. Make sure `R` (https://www.r-project.org/) and `RStudio` (https://www.rstudio.com/) are installed and running. We used `R 3.4.4` (with `OpenBLAS 0.3.6`) and `RStudio Server 1.2.1335` on an `Ubuntu 18.04` machine and, due to possible changes in future version of `R`, `Rstudio` and `R packages`, we strongly recommend the use of the same or of a very similar setup (please see *Appendix I: R session info* in *Supplementary Results 1* for full info).
2. Load `SupplementaryResults1.Rmd` in `RStudio` and simply "Knit" it (https://rmarkdown.rstudio.com/authoring_quick_tour.html).
3. An output document very similar (barring random factors and round errors) to *Supplementary Results 1* should result, together with a directory `cache` (where computationally expensive intermediate results are computed during the 1st run and cached for the following ones), a directory `figures` (where the plots used in the paper are exported both in low- and high-resolution formats), as well as various temporary files

and directories created by the "knitting" process itself. Please note that the 1st knitting will take very long (may take up to 1-2 hours on a high-end multi-core desktop), but even the subsequent "knits" may require 20 minutes or so, depending on the changes made to the `Rmarkdown` script and the caching strategies.

3.2. Supplementary Software 2

This is a `ZIP` archive containing some of the data and the `Rmarkdown` script needed to reproduce the *Supplementary Results 2* document. Please see above the info about *Supplementary Results 1*, but please note that due to privacy concerns, some crucial input data are ***not*** provided, making the actual compilation ("knitting") of this `Rmarkdown` script impossible. Nevertheless, the script itself is provided for inspection and reuse.

3.3. Supplementary Software 3

This `ZIP` archive contains the software needed to run the simulations and is based on ref[4] – please see also its associated GitHub repository https://github.com/ddediu/let-the-agents-do-the-talking for more explanations and details:

- the `Kits` directory contains a text document `needed-kits-Windows-64.txt` describing what software is needed for running the simulations (it was currently tested on `Windows 7 64 bits` with `Java7` and `Python2.7` only) – due to potential licensing issues, we preferred not to include the actual installation kits for 3rd party software here.
- the `agent` directory contains the modified VTL2.1 binary library (`NativeInterface.dll`), the agent Java code (`Agent.jar`), the Bézier hard palate Cython code (`cyBezier.pyd`), the Python script controlling the simulation (`chain.py`), together with configuration files:
    - the `config.csv` is a comma-separated CSV file describing a simulation; while most parameters are self-explanatory, some of the important ones here are:
        - `wav`: should the actual acoustic output be saved as well (may require a lot of disk space)
        - `maxProcesses`: how many parallel simulations to run?
        - `nReplications`: the number of replications for each condition
        - `nChainGen`: how many generations does a chain have?
        - `targets`: the vowels to simulate (as described in `config/targets.csv` below)
        - `nFormants`: the number of formants to consider
        - `iAnatomies`: the anatomies (i.e., MSHPSs) to use (as 0-based indices in the `config/anatomy.csv` file described below)

- the (confusingly named) `config` directory contains two more configuration files referenced from the `config.csv` file, namely:
  - `config/targets.csv`: a comma-separated CSV file containing the description of the target sounds that could be used in a simulation (there may be more than actually used), giving the fixed VTL2.1 parameters as first two rows (i.e., parameters that do not change between vowels and MSHPSs), followed (from row 4th onwards) on separate rows by each sound's identity (i.e., "i" or "schwa"), and the values of the modifiable VTL2.1 parameters (named on the 3rd row) that result in that target sound.
  - `config/anatomy.csv`: a comma-separated CSV file containing the description of the anatomies (i.e., MSHPSs) to be used in the simulation, including those parameters that do not change between MSHPSs; here, only the 4 Bézier parameters (*Palate fronting*, *Palatal concavity*, *Palate angle* and *Alveo-palatal weight*) vary.

Running a (set of) simulation(s) may involve:

a. making sure all the needed software is installed (as described in `Kits/ needed-kits-Windows-64.txt`) and running
b. adding any new target sounds to `agent/config/targets.csv` file (if needed)
c. adding any new anatomies to `agent/config/anatomy.csv` file (if needed)
d. changing the `agent/config/csv` file so that the targets and anatomies we want to simulate are referenced, and to set the number of replications, generations per chain and parallel threads (and/or other parameters) as desired
e. launching the script /agent/chain.py (recommended from the agent directory)
f. after the simulations are run (which might take a very long time, depending on the hardware used and on the number of targets, anatomies, generations, replications and other parameters), the results should be processed using the `Python` script `results/summarize.py`.

For more information and details, please see refs[4,5] and the GitHub repositories https://github.com/ddediu/let-the-agents-do-the-talking and https://github.com/ddediu/larynx-height.

References

1. Becker-Kristal, R. Acoustic Typology of Vowel Inventories and Dispersion Theory: Insights from a Large Cross-linguistic Corpus. (University of California, 2010).
2. Tiede, M. K., Boyce, S. E., Holland, C. K. & Choe, K. A. A new taxonomy of American English /r/ using MRI and ultrasound. *J. Acoust. Soc. Am.* **115**, 2633–2634 (2004).
3. Janssen, R., Moisik, S. R. & Dediu, D. Modelling human hard palate shape with Bézier curves. *PLOS ONE* **13**, e0191557 (2018).
4. Janssen, R. Let the agents do the talking: On the influence of vocal tract anatomy on speech during ontogeny and glossogeny. (Radboud University/Max Planck Institute for Psycholinguistics, 2018).
5. Janssen, R., Moisik, S. R. & Dediu, D. The effects of larynx height on vowel production are mitigated by the active control of articulators. *J. Phon.* **74**, 1–17 (2019).