Exploring the Onset of Phonetic Drift in Voice Onset Time Perception

Metadata in Open Science Framework (OSF) [`url-redacted-for-review`]

This document provides a description of the dataset used in the project *Exploring the Onset of Phonetic Drift in Voice Onset Time Perception.* The columns in the dataset are defined below:

- **PARTICIPANT –** A categorical variable indicating the participant identifier, a three-digit numerical code ranging from 101 to 940 and assigned semi-randomly (in blocks) before participants were placed into an exposure task condition.
- **CONDITION** – A categorical variable indicating the exposure task condition to which a participant was assigned. There were four possible values for this variable, corresponding to the four exposure task conditions implemented in the study. Conditions 1–3 involved exposure to Tagalog language audio, while Condition 4 was an active control that involved exposure to audio of ocean waves instead, taken from YouTube. The values for this variable are defined as follows:
    - **1** – Condition 1 (crosslinguistic mapping) – These participants heard plosive-initial Tagalog stimuli and judged via keyboard press whether the initial plosive was voiced or voiceless in a two-alternative forced-choice (2AFC) identification paradigm.
    - **2** – Condition 2 (emotion identification) – These participants heard plosive-initial Tagalog stimuli and classified the emotion of the speaker as "POSITIVE", "NEGATIVE", or "NEUTRAL" via a three-way keyboard response option.
    - **3** – Condition 3 (unrelated task) – These participants performed a distractor math task from Gordon et al. (1993) in which they were shown three numbers in a vertical stack and judged whether the numbers were numerically equidistant ("SAME") or not ("DIFF"). During this task they heard plosive-initial Tagalog stimuli, either interleaved with the math trials (condition 3a) or played ambiently across the trials uninterrupted (condition 3b).
    - **4** – Condition 4 (active control) – These participants performed the same math task as in Condition 3, with ocean wave audio played across the trials as in Condition 3b and no exposure to Tagalog.
- **GENDER** – A categorical variable indicating the self-identified gender of participants, either "M" (male), "F" (female), or "Other".
- **AGE** – A variable indicating the self-reported age of participants, ranging from 18-73.
- **TRUEVOICING** – A binary variable indicating whether participants reported experience in an L2 with a "true voicing" VOT contrast (such as Spanish, French, or Portuguese). A

value of "1" indicates experience with such a language, and a value of "0" indicates no such reported experience.

- **TRUEVOICERECENCY** – A categorical variable indicating whether participants with prior experience in a "true voicing" L2 (value "1" for TRUEVOICING) had been exposed to said language only in the past ("past") or were currently being exposed to it ("current"), for instance in an L2 classroom setting. This variable was coded as NA for participants with no such prior experience (value "0" for TRUEVOICING).

- **EXPOSURE** – A variable indicating the number of exposures elapsed at the L1 test of the trial in the current row. There were 10 exposures, so values are from 0–10 (foreign language/FL exposures in Conditions 1-3, ocean ambience in Condition 4).

- **L1TASK** – A variable indicating the L1 perception test of the trial in the current row. There were 20 L1 perception tasks, so values are from 1–20.

- **RECENCY –** A categorical variable with two values, indicating whether the trial in the current row is from an L1 post-test ("post-test") immediately following exposure, or an L1 pre-test ("pre-test") after a delay of several hours from the last exposure.

- **PLACE** – A categorical variable with two values, indicating whether the trial in the current row is from a bilabial ("bilabial") or velar ("velar") plosive continuum.

- **TRIALNUMBER** – A variable indicating progression through an L1 test, or the sequential position of the trial in the current row in relation to the other trials. There were 72 trials in each L1 test, so values are from 1-72.

- **CONTINUUM** – A categorical variable with 12 values, indicating the specific plosive continuum of the trial in the current row. There were six bilabial plosive continua ("PB1", "PB2", "PB3", "PB4", "PB5", "PB6") and six velar plosive continua ("KG1", "KG2", "KG3", "KG4", "KG5", "KG6").

- **STEP** – A variable indicating the position within the continuum of the VOT step for the trial in the current row. There were steps in each continuum, so values are from 1-12.

- **SOUNDFILE** – A categorical variable indicating the unique name of the sound file for the plosive step of the trial in the current row. These file names contain both the continuum identity and the number of the step within the continuum (e.g., "KG3_12.wav" is the 12th plosive step in the continuum "KG3"). Each unique sound file was heard by participants 20 times (once in each of the 20 L1 tests).

- **HEARDVOICELESS –** The (binary) dependent variable, indicating whether the plosive step for the trial in the current row was perceived as voiceless (value "1") or as voiced (value "0"). The statistical analyses in the study observe whether the independent variables

significantly predict the likelihood of HEARDVOICELESS being "1" (in other words, the likelihood of plosive tokens being heard as voiceless).

- **RESPONSETIME** – A variable indicating the response time for the trial in the current row. Values that were shorter than 150 ms or 2.5 standard deviations longer than the mean for a given participant (34% of responses) were excluded from the analysis.

Reference

Gordon, Peter C., Jennifer L. Eberhardt, and Jay G. Rueckl. 1993. Attentional modulation of the phonetic significance of acoustic cues. *Cognitive Psychology* 25 (1): 1–42. doi:10.1006/cogp.1993.1001.