# Introducing the Speech Maturity Dataset: Research opportunities for speech scientists and linguistic fieldworkers

Margaret Cychosz[1], Kasia Hitczenko[2], William Havard[3], Loann Peurey[4], Madurya Suresh[1], Theo Zhang[1], and Alex Cristia[4]

*[1]University of California, Los Angeles, [2]George Washington University, [3]Université Grenoble Alpes, [4]École normale supérieure*

Over the first years of life, children's spontaneous vocal productions become increasingly adult-like in their shape and phonetic properties, laying the foundation for later phonological development (Oller, 2000). Yet, as in language development at large, research in this area has been limited to a narrow set of languages and communities, mainly Indo-European from Western(ized) speaker communities, limiting our understanding of cross-linguistic and cross-cultural variation in speech development (Kidd & Garcia, 2022).

To address this issue, we introduce a new publicly-available corpus, the Speech Maturity Dataset (SMD), consisting of 258,914 labeled audio clips extracted from child-centered, longform audio recordings (~8 continuous hours/child). Recordings came from 398 children (209 male, 186 female), aged 2 months to 6 years, from 14 communities (ranging from rich industrialized societies to farmer-forager speaker communities) learning 25+ languages. All clips were manually labeled for speaker and vocalization type by at least 3 citizen scientists (i.e., non-scientific volunteers who devote time to annotate and label scientific data) on Zooniverse, the world's largest citizen science platform. Citizen scientists labeled each clip by vocalization type: laughing, crying, canonical (speech-like vocalization containing an adjacent consonant and vowel), non-canonical (speech-like vocalization without an adjacent consonant and vowel), or junk (silence or non-human sounds). For a subset of the clips (N=110,577), citizen scientists also labeled the speaker type: baby (younger than 3 years), child (3-12 years), female/male adolescent (12-18 years), or female/male adult.

This demonstration and walk-about has two goals. First, albeit already massive, SMD represents the first version of an ongoing collaborative effort between field linguists, phoneticians, and developmental scientists. SMD continues to grow: the citizen science project is still live (LINK REMOVED FOR REVIEW) and we continue to accept new data for annotation into the dataset. So the first objective of our demonstration is to illustrate several case studies of how we helped traditional documentary field linguists, with no background in child language research or large-scale speech corpora, to collect and contribute data to SMD, resulting in several large-scale research collaborations.

The second objective of our demonstration is to illustrate how SMD, which includes a wealth of metadata (child's age, gender, linguistic environment, etc.), lends itself to the development of new tools to automate the processing of largescale, spontaneous speech recordings. We will illustrate how SMD is already used to study child speech development at an unprecedented scale in a wide variety of communities, by computing indices of children's vocal development such as canonical proportion (i.e. the proportion of speech-like vocalizations that contain an adjacent consonant and vowel) or linguistic proportion (i.e. the proportion of vocalizations that are speech-like) (Hitczenko et al., 2023). We will end by showcasing how we used SMD to train supervised vocalization-type classifiers in an effort to make software dedicated to largescale speech corpus processing free, open-source, and reproducible.

References

Hitczenko, K., Bergelson, E., Casillas, M., Colleran, H., Cychosz, M., & Cristia, A. (2023). The development of canonical proportion continues past toddlerhood. *Proceedings of the International Congress of the Phonetic Sciences*. International Congress of the Phonetic Sciences, Prague, CZ.

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, *42*(6), 703–735. https://doi.org/10.1177/01427237211066405

Oller, D. K. (2000). *The emergence of the speech capacity*. Lawrence Erlbaum Associates.