



Employing self-supervised learning models for cross-linguistic child speech maturity classification

Theo Zhang¹, Madurya Suresh², Anne S. Warlaumont⁴, Kasia Hitczenko⁵, Alejandrina Cristia⁶, Margaret Cychosz^{2,3}

¹Dept. of Computer Science, UCLA, USA ²Dept. of Linguistics, UCLA, USA ³Dept. of Linguistics, Stanford University, USA ⁴Dept. of Communication, UCLA, USA ⁵Department of Computer Science George Washington University, USA ⁶Département d'études cognitives, ENS, France,

mcychosz@stanford.edu

Abstract

Speech technology systems struggle with many downstream tasks for child speech due to small training corpora and the difficulties that child speech pose. We apply a novel dataset, SpeechMaturity, to state-of-the-art transformer models to address a fundamental classification task: identifying child vocalizations. Unlike previous corpora, our dataset captures maximally ecologically-valid child vocalizations across an unprecedented sample, comprising children acquiring 25+ languages in the U.S., Bolivia, Vanuatu, Papua New Guinea, Solomon Islands, and France. The dataset contains 242,004 labeled vocalizations, magnitudes larger than previous work. Models were trained to distinguish between cry, laughter, mature (consonant+vowel), and immature speech (just consonant or vowel). Models trained on the dataset outperform state-of-the-art models trained on previous datasets, achieved classification accuracy comparable to humans, and were robust across rural and urban settings.

Index Terms: Cross-linguistic automatic speech recognition, speech development, infancy, wav2vec2, spontaneous speech

1. Introduction and related work

In the first years of life, children's speech becomes increasingly adult-like. By about 6-7 months of age, infants start producing sounds that contain both consonant and vowel elements, forming what are known as canonical syllables [1]. Canonical syllables continue to make up an increasing proportion of children's vocal productions over the subsequent years, tracking advances in speech maturity. Technology to detect this child speech maturity holds great clinical and educational promise, including the potential ability to identify children at-risk of delay/disorder years before current behavioral techniques permit [2]. Child speech does not, however, develop the same in all languages. Phonological structures vary dramatically across the world's languages [3], and children's early speech reflects this diversity [4]. Research in this area has traditionally been constrained by somewhat narrow, linguistically unrepresentative datasets that do not capture the full spectrum of child speech across different linguistic and acoustic environments.

The bottleneck to progress has been the lack of largescale, carefully-annotated child speech datasets. However, weakly- or self-supervised learning (SSL) models have begun to overcome these size limitations, performing various child speech classification tasks [5, 6, 7, 8, 9]. SSL models function by first pre-training on large amounts of unannotated audio or images and then fine-tuning on a smaller amount of annotated data. Here we attempt child speech maturity classification at the utterance level, distinguishing between linguistic (speech-like) versus non-linguistic (cry or laughter), and between ma-

ture/canonical (containing a consonant-vowel transition) versus immature/non-canonical (containing just a consonant or vowel) vocalizations [9, 10]. A reliable classifier at the individual vocalization level would allow automatic assessment of a child's speech development stage as well as facilitating other analyses such as determining the rates of adult responding to different child vocalization types.

Previous studies have attempted child vocalization classification, but many focused on basic tasks such as cry detection [11, 12] and struggled with naturalistic datasets with low accuracy due to limitations and uniformity of the training data. Existing research has been fundamentally limited by datasets that are geographically, linguistically, and acoustically homogeneous, providing only a restricted view of early childhood communication. Little work has examined how SSL models classify child speech collected from diverse languages, or in the more realistic contexts in which children actually learn language (c.f. [7]). Prior work testing these models is largely limited due to the lack of diverse age, language, and recording environment data. This is a critical gap because languages differ in the structure of their sound inventories [3], and there is significant cross-cultural variation in the acoustic environments in which children develop language—some children spend the majority of their time outdoors meaning that children are systematically exposed to differing amounts of acoustic properties that compete with the speech signal such as wind interference.

Our approach fundamentally transforms this landscape by introducing a dataset that captures child vocalizations from 25+ languages, spanning radically different acoustic ecologies—from dense, urban centers to remote communities. Not only do we create a more realistic and diverse testing dataset from the novel corpora, we also systematically examine model performance across (e.g. U.S., France) versus rural (e.g. Vanuatu, Bolivia) settings. We build on the Interspeech 2019 Computational Paralinguistics Challenge [10], which achieved moderate success (baseline Unweighted Average Recall = 58.7%).

2. Child speech corpora

2.1. Corpora construction

We employ two different training corpora: BabbleCorpus (N=11,304 labeled vocalizations from 6 languages)—the training dataset for the current baseline models—and the significantly-expanded SpeechMaturity (N=64,636 vocalizations from 25+ languages) which has never been employed for this task before. BabbleCorpus contains vocalizations from 46 typically-developing children (2-36 months) exposed to a range of mostly genetically-unrelated languages (English, Spanish, Tsimane', Tselal, Yéli Dnye, and Quechua [13, 14, 15, 16, 17, 18, 19]). SpeechMaturity contains vocalizations from 222 chil-

dren (aged 3-72 months; 90% typically-developing) exposed to 25+ languages, such as French, Ninde, and Simbo, from 6 different regions. Note that the entire SpeechMaturity corpus contains N=242,004 labeled vocalizations but the “laughing” category is under-represented. To mitigate model performance issues due to class imbalance, we thus down-sampled all classes (explained below) except “laughing,” to approximately 3x the total “laughing” clips (see Table 1 for counts).

Data for both corpora were collected using small audio recording devices which the child wore over the course of an entire day (6-16 continuous hours). Vocalizations were extracted by performing voice type diarization upon each recording using either the Language ENvironment Analysis (LENA) system [20] or Voice Type Classifier [21]. N=100 (BabbleCorpus) or N=300 (SpeechMaturity) vocalizations/child were sampled from each recording except for [13] where all infant vocalizations were hand-segmented. Each vocalization was divided into smaller clips (modal length=500 ms), to remove identifying information, and posted to a public citizen science crowdsourcing website for human annotation. After brief training, citizen scientists listened to the audio clips and classified each as “crying,” “laughing,” “canonical,” “non-canonical,” or “other/junk” (e.g. no sound, animal sounds, etc.). See [22, 23] for further detail. Each vocalization clip was annotated by at least 3 distinct annotators on the platform. For BabbleCorpus, the original corpus creators only kept vocalizations where a *strong* majority of human annotators ($\geq 66\%$) agreed on the label; the rest of the vocalizations were discarded. For SpeechMaturity, however, we generated two sub-datasets: (1) SpeechMaturity-Cleaned (53,089 clips), which included only vocalizations where likewise a strong majority ($\geq 66\%$) of human annotators had agreed on the label and (2) SpeechMaturity-Uncleaned (62,000 clips) which included *all* clips from SpeechMaturity-Cleaned, as well as those that were unreliable and/or too difficult to classify and only resulted in the *highest number* of annotator agreements, not necessarily $>51\%$ of annotator agreement (e.g. for a clip with 6 annotations, 3 of which were ‘canonical,’ 2 were ‘laughing,’ and 1 was ‘crying’ would be classified as ‘canonical’). This is a critical distinction between BabbleCorpus—the current state-of-the-art dataset for this task—and SpeechMaturity: BabbleCorpus represents an idealized set of training data that is not representative of the high variability in children’s vocalizations because a portion of the data were removed in post-processing. SpeechMaturity-Cleaned represents a similarly idealized set of test data but across a more diverse spread of ages, languages, and recording environments, while SpeechMaturity-Uncleaned represents the truest test of how the proposed models would extrapolate to new samples. We compare the effect of this data removal step by training the proposed models upon the idealized SpeechMaturity-Cleaned but evaluating model performance upon both SpeechMaturity-Cleaned and -Uncleaned.

2.2. Corpora pre-processing

For precise comparison of our approaches with previous models that were trained using BabbleCorpus, we replicate the train/dev/test split from the original 2019 Paralinguistics Challenge for BabbleCorpus (Table 1; train/dev/test=35/32/33, with child-disjunct folds). We apply a more traditional 80/10/10 for SpeechMaturity, again with child-disjunct folds, and ensuring that all child ages and languages were approximately evenly represented across folds. Audio clips were converted to mono audio arrays, resampled to 16kHz, and 0-padded around the center such that all arrays contained

9217 elements, which was the maximum length after mono conversion and resampling. This ensured all model inputs were uniform, and no data were lost through truncation. Scripts to replicate our models are available at github.com/spoglab-stanford/w2v2-pro-sm/tree/main/speechbrain/recipes/W2V2-LL4300-Pro-SM.

Table 1: *Class distribution for child speech corpora employed in model training*

Class	BC ^a			SM-C ^b			SM-U ^c
	Train	Dev	Test	Train	Dev	Test	Test
Crying	243	163	263	9830	1177	1098	1116
Laughing	46	41	62	3491	388	356	358
Canonical	444	378	604	9762	1262	1226	1674
Non-Canonical	1437	1678	1370	9766	1232	1252	1867
Junk	1826	1357	1392	9838	1226	1185	1185
Total	3996	3617	3691	42687	5285	5117	6200

^aBC: BabbleCorpus - Reflects original train/dev/test splits from previous attempts.

^bSM-C: SpeechMaturity-Cleaned - Reflects data after down-sampling to mitigate class imbalance.

^cSM-U: SpeechMaturity-Uncleaned - Only the test set from SpeechMaturity-Uncleaned was used in experiments. Also reflects data after down-sampling to mitigate class imbalance.

3. Model architectures

We employed three different Wav2Vec2 models of varying size and complexity for the task of child speech maturity classification: *W2V2-base*, *W2V2-LL4300h*, and *W2V2-LL4300-Pro*. The three models were pre-trained in different ways and fine-tuned on either the BabbleCorpus or SpeechMaturity-Cleaned datasets (explained below). *W2V2-base* was pre-trained on thousands of hours of unlabeled LibriSpeech (English) [24] with 12 transformer layers, hidden dimension of 768, inner dimension of 3,072, and 8 attention heads [25]). *W2V2-base* passes outputs from a CNN feature extractor through a transformer architecture to develop contextualized speech representations (see [25] for further description). The second model was *W2V2-LL4300h*, which pre-trained *W2V2-base* on 4300h of daylong home-based audio recordings of children under 5 years old acquiring English [12]. Finally, *W2V2-LL4300h* incorporates the auxiliary task of child speech phoneme recognition within *W2V2-LL4300h* [7]. To achieve this, the hidden features from a model that generates phonetic pseudo-reference transcripts are fused with *W2V2-LL4300h* by adding a linear layer to a middle transformer layer of *W2V2-LL4300h*. It is then trained using CTC to generate hypothesis transcripts that match the pseudo-reference transcripts with minimum cross-entropy. Essentially, the auxiliary task requires *W2V2-LL4300h* to simultaneously learn and incorporate children’s phonetic representations along with predicting the target 5 labels, a change that results in the *W2V2-LL4300-Pro* model. [7]. *W2V2-LL4300-Pro* is more complex compared to *W2V2-LL4300h* due to these additions, and thus is the most complex model upon which we conduct experiments.

All three models were then subsequently fine-tuned on either the BabbleCorpus dataset or the SpeechMaturity-Cleaned dataset (noted in Table 2 with either -BC or -SM). For *W2V2-base*, audio data were inputted into the model’s feature extractor; those features were then inputted into the pre-trained model’s transformer architecture for classification in train/dev/test batches as outlined in Table 1. Data were processed in batches of 32, via random sampling during training. For *W2V2-LL4300h* and *W2V2-LL4300-Pro*, we replicated the

Table 2: State-of-the-art model performance as reported in all prior publications attempting this classification task, all tested on the BabbleCorpus dataset (the only extant at the time). The best Test performance corresponds to Li et al.

Model	UAR (%)		
	Train	Dev	Test
Dataset: BabbleCorpus			
Challenge Baseline ^a [10]	*	54.0	58.7
Yeh et al. ^a [26]	*	61.3	62.4
Gosztolya ^a [28]	*	58.7	59.5
Kaya et al. ^a [29]	*	60.1	61.4
Li et al. [7]	*	70.4	64.6

^aPerformance on Train set not reported in original paper.

training and testing environments with the provided pre-trained checkpoints from [7]. Training for all three models was conducted synchronously on 10 CPU cores for 10 epochs (learning rate=3e-5 for *W2V2-base* and 1e-5 for *W2V2-LL4300h* and *W2V2-LL4300-Pro*). The best-performing epoch for each model was then used for testing.

4. Results

Following previous classification models for this task (e.g. [26]), model performance was evaluated using the unweighted average recall (UAR), a metric that takes the mean of recall values for each class, giving equal importance to each class regardless of its size. It is thus well-suited for multi-class classification tasks, especially when classes are imbalanced as they are here [27] (Table 3). Results highlight two key contributions of this work, one regarding data and the other model architecture. Across models and test sets, performance increases when fine-tuning on the SpeechMaturity dataset compared to BabbleCorpus, with increases that go well beyond simply matching training and test sets. For example, each model type gained between 7 and 30% when the fine-tuning set was SpeechMaturity-Cleaned and the test sets were either of the other two. As for architecture, we replicate the observations that pretraining on relevant data and incorporating a phonetic task improves performance when compared with *W2V2-base* [7], but further show that the relative gain depends crucially on the fine-tuning dataset (e.g., max. 26.4% for BabbleCorpus versus max. 4.6% for SpeechMaturity, both when testing on BabbleCorpus). Finally, our conditions significantly surpassed previous state-of-the-art performance (best UAR for *W2V2-LL4300-Pro-SM* at 74.2%), even when holding the test set constant (same model, 68.6% versus 54-64.3%; see Table 2).

For by-category results, Figure 1 shows the final confusion matrix over the SpeechMaturity-Cleaned test set for *W2V2-LL4300-Pro-SM* which was fine-tuned on SpeechMaturity-Cleaned. The results of *W2V2-LL4300-Pro-SM* far surpass the state-of-the-art (trained on the smaller, less representative BabbleCorpus) which achieved a UAR=67.7% for the speech category “canonical” and UAR=42.5% for “non-canonical” in [26], while *W2V2-LL4300-Pro-SM* achieved a UAR=84.3% on “canonical” and UAR=62.9 on “non-canonical.”

4.1. Classification accuracy of models versus human annotators

To set a benchmark for model performance and rigorously assess the performance of our model in a way that pre-

Table 3: UAR (%) across fine-tuning (rows) and testing sets (columns) for our best-performing models. Performance metrics listed in the BC (BabbleCorpus) column correspond to the same Test data as in the Test column in Table 2.

Model	UAR (%)		
	BC ^a	SM-C ^b	SM-UC ^c
Fine-tuning Dataset: BabbleCorpus			
W2V2-base-BC	34.4	29.9	29.4
W2V2-LL4300h-BC	60.2	50.0	48.1
W2V2-LL4300-Pro-BC	60.8	51.7	49.8
Fine-tuning Dataset: SpeechMaturity-Cleaned			
W2V2-base-SM	64.2	71.0	68.4
W2V2-LL4300h-SM	66.6	73.8	71.5
W2V2-LL4300-Pro-SM	68.6	74.2	71.9

^a BC: BabbleCorpus testing set.

^b SM-C: SpeechMaturity-Cleaned testing set.

^c SM-UC: SpeechMaturity-Uncleaned testing set.

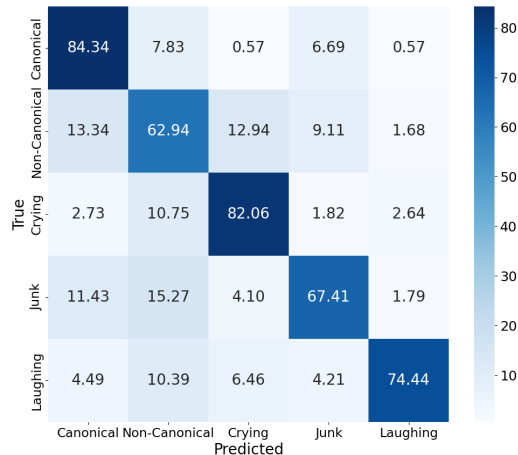


Figure 1: Confusion matrix of the SpeechMaturity-Cleaned test set predictions for *W2V2-LL4300-Pro* finetuned on SpeechMaturity-Cleaned.

vious work has not, we conducted an agreement analysis on *W2V2-LL4300-Pro-SM*. We assessed inter-human annotator agreement using a weighted Fleiss’ kappa metric for the SpeechMaturity-Cleaned and -Uncleaned datasets. We employed a custom weighting scheme to reflect the relative importance of different category distinctions in the classification task [30], assigning higher weights to the canonical versus non-canonical distinction (the most important linguistic classification distinction in this task), as well as between speech-like categories (canonical, non-canonical) versus non speech-like vocalizations (cry, laugh, and junk). Using this weighted scheme, we found a weighted Fleiss’ kappa of $K = 0.375$ (95% CI: 0.361-0.388) for SpeechMaturity-Cleaned and $K = 0.271$ (0.259-0.284) for SpeechMaturity-Uncleaned for agreement between 5 or fewer annotators. Unsurprisingly, agreement improved when comparing between vocalizations annotated by 3 or fewer annotators for SpeechMaturity-Uncleaned ($K = 0.282$, CI: 0.106-0.458) and -Cleaned ($K = 0.457$, CI: 0.402, 0.512). These results indicate fair to moderate levels of agreement among human annotators—which is unsurprising for a task of this complexity—and suggest a benchmark by which to compare model performance.

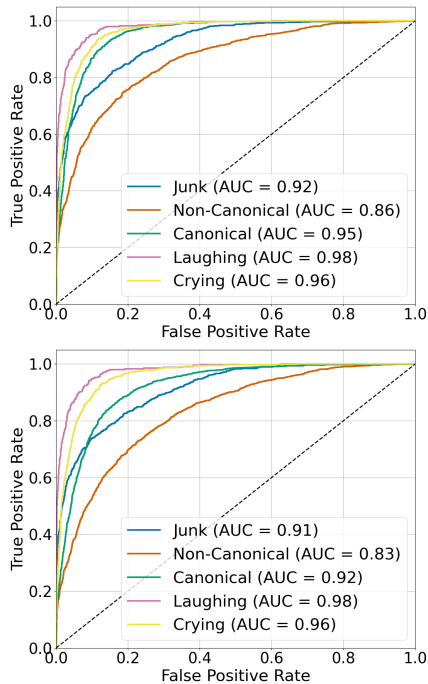


Figure 2: ROC curves achieved by **W2V2-LL4300-Pro-SM** on *SpeechMaturity-Cleaned* (top) and *SpeechMaturity-Uncleaned* (bottom) for classification of canonical (green), non-canonical (red), crying (yellow), laughing (pink), and junk (blue) in comparison to human annotators.

Next, we compared human annotator performance to **W2V2-LL4300-Pro-SM** by calculating weighted Cohen’s kappa between **W2V2-LL4300-Pro-SM** and each annotator of each vocalization over *SpeechMaturity-Cleaned* and *-Uncleaned*. The average weighted Cohen’s kappa was $K = 0.478$ ($SD = 0.012$) for *SpeechMaturity-Cleaned* and $K = 0.406$ ($SD = 0.015$) for *SpeechMaturity-Uncleaned*, indicating moderate levels of agreement between **W2V2-LL4300-Pro-SM** and human annotators. The standard deviation reflects the variability in agreement between the model and different annotators. Most critically, these levels of agreement approach and/or surpass the level of multiple human inter-rater agreement computed above. Figure 2 displays the ROC curves and respective by-category AUC scores. For the model tested on *SpeechMaturity-Cleaned*, the AUC scores indicate excellent to outstanding discrimination across all categories with particularly strong classification performance for canonical vocalizations. For *SpeechMaturity-Uncleaned* we likewise found excellent to outstanding AUC scores.

4.2. Classification accuracy of the SSL models by language learning environment

W2V2-LL4300-Pro-SM tested over *SpeechMaturity-Cleaned* achieved a UAR of 70.7% in urban environments and 67.8% in rural environments, showing a robustness to data from the vastly different language and acoustic learning environments of children from around the world (Table 4). The discrepancy in UARs by environment could be due to a number of reasons including wind interference (due to increased time spent out of doors) or more multi-party and/or overlapping speech, both of which have been documented during psycholinguistic

and ethnographic fieldwork conducted in the language learning environments of children in the rural areas of this corpus (e.g. [31]). Further exploration of the causes of the discrepancies by rural versus urban settings is an avenue for future research.

Table 4: *Distribution and Performance by Language Environment for SpeechMaturity-Cleaned*. Upper= clip counts by split in urban v. rural. Lower= UARs.

Split/ Metric	Language Environment		
	Urban	Rural	Total
Train Clips	8508	32723	41231
Dev Clips	759	4283	5042
Test Clips	681	4436	5117
Total Clips	9948	41442	51390
UAR (SD) ^a	70.7 (<0.01)	67.8 (<0.01)	-

^aUAR: Unweighted Average Recall (%); SD: Standard Deviation

5. Discussion

SpeechMaturity represents a significant shift in child speech research, challenging existing methodological constraints in computational studies of child speech development. By capturing child vocalizations across 25+ languages and dramatically diverse acoustic environments—from urban centers in industrialized communities to remote communities in Vanuatu and Papua New Guinea—this corpus provides an unprecedented window into the global landscape of early vocal development. The dataset is openly available for other researchers to use [32] and build tools with, allowing new questions about early communication to be answered in a cross-linguistically diverse sample.

To demonstrate the potential of this dataset to resolve longstanding challenges in child speech technology, we applied three transformer models to child speech maturity classification. When trained on the comprehensive *SpeechMaturity* dataset ($N=64,636$ audio clips from 222 children acquiring one or more of 25+ languages), the models outperformed those trained on smaller, less representative datasets (*BabbleCorpus*). This performance gain reflects *SpeechMaturity*’s ecological richness: by including children from vastly different linguistic and acoustic settings, *SpeechMaturity* enables more robust generalizable insights into early speech development. Notably, model performance consistently improved after fine-tuning on *SpeechMaturity*, regardless of model complexity. This underscores the dataset’s fundamental value: it captures variability in child speech that previous, smaller and more limited corpora overlooked. Even the lowest-complexity model, **W2V2-base-SM** which required no additional pre-training or auxiliary tasks and took the least amount of memory to fine-tune and test, performed comparably to either of the more computationally-intensive models, suggesting that the dataset’s diversity may be more impactful for success at this task than overall model architecture or sophistication. The best performing model, **W2V2-LL4300-Pro-SM** fine-tuned on the novel *SpeechMaturity* dataset exceeds previously published state-of-the-art solutions and (1) was robust even on a dataset that consisted of “noisier” clips (*SpeechMaturity-Uncleaned*), (2) had acceptable levels of agreement with human annotators and strong AUC values, and (3) had consistently high levels of performance between rural and urban child rearing environments.

6. References

- [1] D. K. Oller, *The Emergence of the Speech Capacity*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.
- [2] S.-I. Ng, C. W.-Y. Ng, J. Wang, and T. Lee, "Automatic Detection of Speech Sound Disorder in Child Speech Using Posterior-based Speaker Representations," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 2853–2857.
- [3] I. Maddieson, *Patterns of Sounds*, ser. Cambridge Studies in Speech Science and Communication. Cambridge [Cambridgeshire]; New York: Cambridge University Press, 1984.
- [4] B. de Boysson-Bardies, M. M. Vihman, and B. de Boysson-Bardies, "Adaptation to Language: Evidence from Babbling and First Words in Four Languages," *Language*, vol. 67, no. 2, pp. 297–319, Jun. 1991.
- [5] R. Fan, Y. Zhu, J. Wang, and A. Alwan, "Towards Better Domain Adaptation for Self-Supervised Models: A Case Study of Child ASR," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1242–1252, Oct. 2022.
- [6] A. Gorin, C. Subakan, S. Abdoli, J. Wang, S. Latremouille, and C. Onu, "Self-supervised learning for infant cry analysis," in *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2023, pp. 1–5.
- [7] J. Li, M. Hasegawa-Johnson, and K. Karahalios, "Enhancing Child Vocalization Classification with Phonetically-Tuned Embeddings for Assisting Autism Diagnosis," Jun. 2024.
- [8] N. Al Futaisi, Z. Zhang, A. Cristia, A. Warlaumont, and B. Schuller, "VCMNet: Weakly Supervised Learning for Automatic Infant Vocalisation Maturity Analysis," in *2019 International Conference on Multimodal Interaction*. Suzhou China: ACM, Oct. 2019, pp. 205–209.
- [9] Z. Zhang, A. Cristia, A. S. Warlaumont, and B. Schuller, "Automated Classification of Children's Linguistic versus Non-Linguistic Vocalisations," in *Proceedings of Interspeech 2018*, Hyderabad, India, 2018.
- [10] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl, A. S. Warlaumont, L. Yankowitz, E. Nöth, S. Amiriparian, S. Hantke, and M. Schmitt, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 2378–2382.
- [11] M. Micheletti, X. Yao, M. Johnson, and K. De Barbaro, "Validating a model to detect infant crying from naturalistic audio," *Behavior Research Methods*, vol. 55, no. 6, pp. 3187–97, 2022.
- [12] J. Li, M. Hasegawa-Johnson, and N. L. McElwain, "Towards Robust Family-Infant Audio Analysis Based on Unsupervised Pre-training of Wav2vec 2.0 on Large-Scale Unlabeled Family Audio," in *INTERSPEECH 2023*, Aug. 2023, pp. 1035–1039.
- [13] M. Casillas, P. Brown, and S. Levinson, *Casillas HomeBank Corpus*, 2017.
- [14] M. Cychosz, *Cychosz HomeBank Corpus*, 2018.
- [15] E. Bergelson, *Bergelson Seedlings HomeBank Corpus*, 2017.
- [16] A. Cristia and H. Colleran, *Long-Form, Child-Centered Recordings Collected in Malekula in 2016-2018*, 2018.
- [17] A. Warlaumont, G. Pretzer, S. Mendoza, and E. Walle, *Warlaumont HomeBank Corpus*, 2016.
- [18] C. Scaff, J. Stieglitz, and A. Cristia, *Daylong Recordings from Young Children Learning Tsimane in Bolivia*, 2018.
- [19] M. VanDam, A. S. Warlaumont, E. Bergelson, A. Cristia, M. Soderstrom, P. De Palma, and B. MacWhinney, "HomeBank, an online repository of daylong child-centered audio recordings," *Seminars in Speech and Language*, vol. 37, pp. 128–142, 2016.
- [20] D. Xu, U. Yapanel, and S. Gray, "Reliability of the LENA Language Environment Analysis System in young children's natural home environment," LENA Research Foundation, Boulder, CO, Technical Report ITR-05-2, 2009.
- [21] M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, and A. Cristia, "An open-source voice type classifier for child-centered daylong recordings," *arXiv:2005.12656 [eess]*, 2021.
- [22] M. Cychosz, A. Cristia, E. Bergelson, M. Casillas, G. Baudet, A. S. Warlaumont, C. Scaff, L. Yankowitz, and A. Seidl, "Vocal development in a large-scale crosslinguistic corpus," *Developmental Science*, vol. 24, no. 5, p. e13090, 2021.
- [23] K. Hitzenko, E. Bergelson, M. Casillas, H. Colleran, M. Cychosz, and A. Cristia, "The development of canonical proportion continues past toddlerhood," in *Proceedings of the International Congress of the Phonetic Sciences*, Prague, CZ, 2023.
- [24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books."
- [25] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Proceedings of the 34th International Conference on NeurIPS Systems*, 2020, pp. 12 449–12 460.
- [26] S. L. Yeh, G.-Y. Chao, B. Su, Y.-L. Huang, M.-H. Lin, Y.-C. Tsai, Y.-W. Tai, Z.-C. Lu, C.-Y. Chen, T.-M. Tai, C.-W. Tseng, C.-K. Lee, and C.-C. Lee, "Using Attention Networks and Adversarial Augmentation for Styrian Dialect Continuous Sleepiness and Baby Sound Recognition," in *Proceedings of Interspeech 2019*, Graz, Austria, 2019, pp. 2398–2402.
- [27] A. Keesing, Y. Koh, and M. Witbrock, *Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech*, Aug. 2021.
- [28] G. Gosztolya, "Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 2413–2417.
- [29] H. Kaya, O. Verkholyak, M. Markitantov, and A. Karpov, "Combining Clustering and Functionals based Acoustic Feature Representations for Classification of Baby Sounds," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*. Virtual Event Netherlands: ACM, Oct. 2020, pp. 509–513.
- [30] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.
- [31] C. Scaff, M. Casillas, J. Stieglitz, and A. Cristia, "Characterization of children's verbal input in a forager-farmer population using long-form audio recordings and diverse input definitions," *In-fancy*, vol. n/a, no. n/a.
- [32] K. Hitzenko, L. Peurey, W. Havar, Tey, A. Seidl, C. Semenzin, M. Lavechin, B. Kelleher, L. Hamrick, L. Gautheron, M. Cychosz, M. Casillas, and A. Cristia, "Speech Maturity Dataset: A cross-cultural corpus of naturalistic child and adult vocalizations," under review.