Employing self-supervised learning models for child speech maturity classification

Madurya Suresh[1]*, Theo Zhang*[2]*, Kasia Hitczenko*[3]*, Alejandrina Cristia*[3]*, Margaret Cychosz*[1]

[1]Department of Linguistics, University of California, Los Angeles, USA [2]Department of Computer Science, University of California, Los Angeles, USA [3]Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, France

mcychosz4@ucla.edu

Abstract

Speech technology systems continue to struggle with many downstream tasks for child speech due to the small size of training corpora, and the difficulties that the physical properties of child speech pose. We apply two self-supervised learning model architectures—Transformer-based and a convolutional neural network (CNN)—to an outstanding child speech classification task: speech maturity classification between cry, laughter, canonical (consonant+vowel), and non-canonical (just consonant or vowel). Models were trained on a large database of maximally ecologically-valid vocalizations (N=53,359) from daylong recordings of children acquiring more than 25 languages in the U.S., Bolivia, Mexico, France, and Oceania. The Transformer outperformed the CNN (unweighted average recall=74.2% to 57.1%). We then train the models on a dataset that the state-of-the-art employed (N=7,687). Results demonstrate the promise of SSL models for the small, imbalanced datasets common to child speech.

Index Terms: child speech, speech-like vocalization, self-supervised learning, speech development

1. Introduction and related work

In the first months and years of life, children's speech becomes increasingly adult-like. Children start producing sounds that demand multiple, simultaneous vocal tract constrictions, and are eventually able to sequentially combine consonants and vowels. Thus, technology to detect child speech maturity—the ability to automatically detect the stage of a child's speech development—holds great clinical and educational promise. Recently-developed tools are starting to help identify infants and children at-risk of language delay and disorder years before current behavioral techniques permit and allow naturalistic samples of children's speech behavior to be incorporated into speech-language therapy programs [1].

The bottleneck to progress in this domain has been the lack of largescale, carefully-annotated child speech datasets. However, in recent years, weakly- or self-supervised learning (SSL) models have begun to overcome the size limitations of child speech corpora and perform many downstream child speech classification tasks such as adult-child speaker diarization [2], infant cry detection [3], and infant vocal classification [4, 5, 6]. SSL models such as autoregressive predictive coding [7], Wav2vec2.0 [8], and hidden unit BERT (HuBERT) [9] function by first pre-training on large amounts of unannotated audio or images and then fine-tuning on a smaller amount of annotated data, a technique that has allowed them to mitigate concerns about relatively small amounts of training data and improve upon these tasks in child speech technology and classification [4, 10].

In this work, we attempt one of the most important downstream tasks in child speech technology: child speech maturity classification, or the ability to distinguish between linguistic (i.e. speech-like) versus non-linguistic (i.e., cry or laughter) vocalizations, and to further distinguish between canonical vocalizations (containing a consonant-vowel transition) versus non-canonical (containing just a consonant or vowel) vocalizations [6, 11]. Children who have a higher proportion of canonical vocalizations in their speech are at a more mature stage of speech development [12]. So a reliable canonical vs. non-canonical classifier would allow researchers and clinicians to automatically assess the stage of a child's speech development and efficiently diagnose speech delay.

Previous studies have attempted child vocalization classification on infant and child speech datasets using a variety of deep learning approaches. [4] employed an SSL model pre-trained on infant-centered home audio recordings, where vocalizations were labeled as cry, fuss, and babble. In a layer-wise analysis, the authors found best 3-way classification performance for middle layers, which were encoded with higher-level phonetic features. However, this approach did not attempt to distinguish between the types of speech-like vocalizations that a child made (i.e. canonical vs. non-canonical), and therefore did not attempt to classify the *maturity* of children's speech, a key outstanding objective within child speech technology. An alternative, more fundamental step within automated vocal classification is the binary classification task of infant cry detection. [13] trained a support-vector machine from a modified AlexNet with a convolutional neural network architecture and achieved an average F1 score of .59. This line of work also demonstrates how cry detection models trained on in-lab samples rarely extrapolate to more ecologically-valid datasets——[14] find F1s of 0.656 for in-lab samples that drop precipitously to 0.236 when applied to more naturalistic speech datasets.

With few exceptions (e.g. [6]), previous models were trained on datasets consisting of children within narrow age ranges (e.g. under 14 months), exposed to a single language (typically English), recorded in relatively uniform and unrealistic acoustic environments (inside of quiet homes and buildings). Little work has been conducted examining how SSL models classify infant and child speech collected from more diverse languages, or in more realistic and variable contexts such as outdoors. This is a critical gap since languages differ systematically in the structure and properties of their sound inventories [15] and there is wide cross-cultural variation in the acoustic environments in which children develop speech and language. It is thus unclear how previous child speech classification models extrapolate to more diverse and realistic datasets.

[11] introduced the Interspeech 2019 Computational Paralinguistics Challenge for the child speech maturity classi-

fication task that we propose. Baseline models employed Bag of Audio Words features [16] and support vector machines with acoustic features stemming from low-level descriptor contours (Unweighted average recall (UAR) for the baseline model=58.7% with marginal improvement seen within the challenge e.g. UAR=62.4% in [17]). Critically, these models were trained over a large, diverse, and ecologically-valid corpus of child vocalizations from daylong recordings gathered in the U.S., Bolivia, Mexico, and Papua New Guinea (henceforth "BabbleCorpus"), and are thus currently considered the state-of-the-art for this task. Given the promise of SSL models such as wav2vec2.0 for low-resource datasets [8], the approaches outlined in [11] and [17] require updating. SSL models are only starting to be applied to the task of child speech maturity classification, though their architecture has improved upon more basic classifications such as linguistic versus non-linguistic [4], suggesting they could similarly outperform current models for the task at hand.

2. Child speech corpora

2.1. Corpora construction

We employ two different corpora for model training: BabbleCorpus (N=7,687 labeled vocalizations)—the training dataset for the current baseline models—and the significantly-expanded SpeechMaturity dataset (N=53,359 vocalizations). BabbleCorpus contains vocalizations from 46 typically-developing children (aged 2-36 months) exposed to a range of mostly genetically-unrelated languages such as English, Spanish, Tsimane, Tseltal, Yélî Dnye, and Quechua [18, 19, 20, 21, 22, 23]. SpeechMaturity contains vocalizations from 96 children (aged 3-72 months; 90% typically-developing) exposed to the above languages as well as languages such as French, Ninde, and Simbo.

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 speaker diarization upon each recording using either the proprietary Language ENvironment Analysis (LENA) system [24] or Voice Type Classifier, an open-source model trained to identify vocalizations and speakers from long-form, child-centered recordings [25]. N=100 (BabbleCorpus) or N=300 (SpeechMaturity) vocalizations/child were sampled from each recording except for [18] where the vocalizations were hand-segmented. Each vocalization was divided into smaller bits (modal length=500 ms), to remove identifying information, and posted to a public citizen science crowdsourcing website for annotation. After brief training, citizen scientists then listened to clips and classified each as "crying," "laughing,", "canonical," "non-canonical," or "junk" (e.g. no sound, adult speech, etc.). Each clip received at least 3 annotations; only vocalizations where the majority of the annotators agreed on the label were kept. See [26, 27] for further detail.

2.2. Corpora pre-processing for modeling

For precise comparison of our approaches with previous models that were trained using BabbleCorpus, we first replicate the counts and samples of the train/test split from the original 2019 Paralinguistics Challenge for BabbleCorpus (Table 1; train/test=52/48). We apply a more traditional 80/20 train/test for SpeechMaturity. Given the large class imbalance (in particular the under-representation of "laughing"), we up-sampled all of the "laughing" clips from a supplemental child speech corpus, collected and labeled using the same methodology, from many of the same children in SpeechMaturity (N=795 samples added). We then down-sampled all other classes until they approached just three times the size of the smallest class ("laughing") and were approximately equally balanced (train+test=12,000 samples/class; see Table 1).

Table 1: *Class distribution for child speech corpora. Counts refer to original train/test splits for previous classifiers trained with BabbleCorpus. SpeechMaturity counts reflect class balances after up- and down-sampling.*

	BabbleCorpus	
	Train	Test
Crying	243	263
Laughing	46	62
Canonical	444	604
Non-canonical	1437	1370
Junk	1826	1392
Total	3996	3691
	SpeechMaturity	
Crying	9830	2420
Laughing	3491	869
Canonical	9762	2488
Non-canonical	9766	2484
Junk	9838	2411
Total	42687	10672

Previous attempts at this classification task for BabbleCorpus varied in whether or not they employed up-sampling and the up-sampling technique employed [17, 28, 29, 11]. The current state-of-the-art, for example, up-sampled in part by supplementing with additional "laughing" files from a different speech corpus [30], collected using different recording methodologies, from different children than BabbleCorpus in way that was not explicitly reported nor replicable. To remain conservative, we thus trained the both the CNN and Transformer model architectures on two BabbleCorpus datasets, one with and another without up-sampling: "canonical" was up-sampled to 100% the size of "non-canonical" and "laughing" was up-sampled using all available samples in the supplementary corpus, to 58.5% of "non-canonical."

Audio clips were converted to mono audio arrays, resampled to 16kHz (as necessary), 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 modeling are available at *blinded for review*.

3. Model architecture

3.1. Convolutional neural network design

The first SSL classification model that we built employed a traditional convolutional neural network (CNN) architecture and design for audio classification. We commenced training with a deep residual network, ResNet34 (a 34-layer model pretrained on the ImageNet dataset [31], accessed using TorchVision [32]). Audio clips were converted to Mel-scaled spectrograms, decomposed using a short-time Fourier transform, and overlaid with a Mel-scale filterbank (N=128 bins). Each re-

sulting spectrogram was subsequently converted to a grayscale image and fed into the transfer-learned CNN for classification. Data were processed in batches of 16 by randomly sampling from all patches during training. Grid searches were conducted for learning rates and batch sizes. Training was conducted synchronously on 4 CPU cores using PyTorch for 10 epochs, employing cross-entropy loss as the loss function and the Adam optimizer algorithm (learning rate=1e-4). A fully connected layer was added to the Resnet34 model architecture in order to output the probabilities for all 5 classes. An additional softmax activation layer then normalized the probabilities followed by a thresholding step (set threshold of 0.5) to convert the probabilities for each class into binary predictions. The final resulting output was a tensor with 5 binary labels: 1 if the sample belonged to a class and 0 if not.

3.2. Transformer-based design

The second SSL classification model that we built employed a Transformer-based design. This transduction model architecture involves minimal or no recurrence or convolutions (unlike the CNN) and instead relies on the attention mechanism ("self-attention") linking the encoder and decoder [33]. In traditional sequential computation architectures, such as CNN, larger distances between inputs and outputs require a larger number of operations, which inherently limits the system's ability to learn dependencies. Transformer architectures instead permit a constant number of operations linking inputs and outputs, regardless of distance. They also require far fewer task-specific training examples for model fine-tuning than traditional encoder-decoder constructions and have thus become the state-of-the-art for several natural language and speech classification tasks [34]. In particular, the ability of these models to generalize from smaller, more limited datasets make them promising for the current classification task.

Model training commenced with wav2vec-base, which passes outputs from a CNN feature extractor through a Transformer architecture to develop contextualized speech representations [8]. Wav2vec-base was pre-trained on 960 hours of unlabeled LibriSpeech [35] with 12 transformer layers, hidden dimension of 768, inner dimension of 3,072, and 8 attention heads [8]). The pre-trained wav2vec-base model can be fine-tuned on labeled data. To fine-tune the model for this supervised classification task, data were inputted into the wav2vec-base feature extractor. Feature outputs were then inputted into the pre-trained Transformer architecture for classification in train/test batches as outlined in Table 1. Data were processed in batches of 32, again via random sampling during training. Training was conducted synchronously on 12 CPU cores for 10 epochs (learning rate=3e-5). The best-performing epoch was then used for testing.

4. Results

Here we present results from our proposed model architectures for BabbleCorpus and SpeechMaturity, as well as previous attempts at this classification task for BabbleCorpus. Following previous classification models for this task (e.g. [28]), model performance was evaluated using the UAR, a metric that is often preferred when the sample class ratio is imbalanced as it is here [36]. UARs for all models are presented in Table 2. In all cases, models evaluated over the test sets performed significantly better than chance (e.g. 20% UAR), indicating success at the task. Overall classification performance is best for the Transformer-based model trained on the larger SpeechMaturity dataset, achieving a UAR of 74.2%.

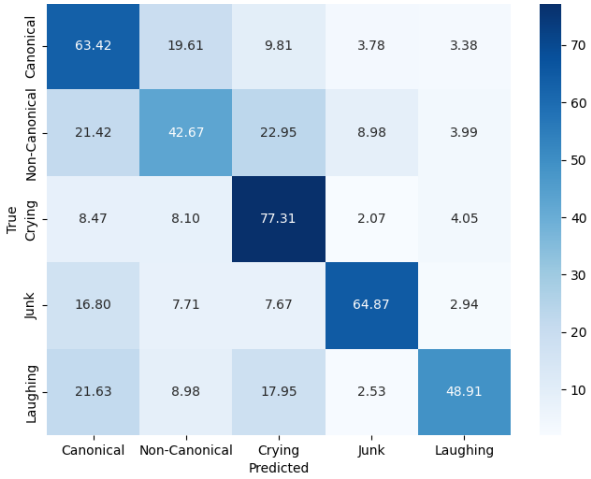

Figure 1: *Confusion matrix of the test set predictions for CNN model trained on SpeechMaturity.*

Table 2: *Unweighted average recall (%) for the baseline, state-of-the-art, and proposed model architectures.*

Model	Train	Test
Dataset: BabbleCorpus		
Challenge Baseline [11]	54.0	58.7
Yeh et al. [17]	61.3	62.4
Gosztolya [29]	58.7	59.5
Kaya et al. [28]	60.1	61.4
Proposed Models		
Dataset: BabbleCorpus		
CNN:BabbleCorpus-noUpSample	92.7	41.9
CNN:BabbleCorpus-UpSample	46.7	39.26
Transformer:BabbleCorpus-noUpSample	35.8	34.7
Transformer:BabbleCorpus-UpSample	77.3	47.5
Dataset: SpeechMaturity		
CNN:SpeechMaturity	67.7	57.1
Transformer:SpeechMaturity	77.7	74.2

Both the CNN and Transformer-based model architectures had stronger classification performance when trained over the larger SpeechMaturity dataset than BabbleCorpus-UpSample or -noUpSample, but the difference was most notable for the Transformer-based model (41.9% to 57.1% improvement for the CNN versus 47.5% to 74.2% improvement for Transformer-based structure). Figures 1 and 2 show the confusion matrices for the CNN and Transformer-based models trained on SpeechMaturity, respectively.

The Transformer-based model outperformed the CNN on all categories except "Crying," where it performed the same (UAR=approximately 77% for both). The largest differences in performance between models were for the speech categories of "Canonical" and "Non-canonical" achieving 85.49% and 64.73% UAR respectively, and thus far surpassing the state-of-the-art as trained on the much smaller and less representative BabbleCorpus (UAR=67.7% for "canonical" and 42.5% for "non-canonical" in [17]). Although SpeechMaturity was

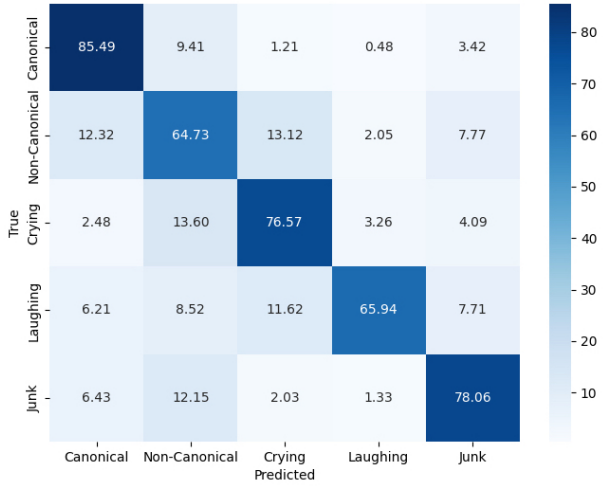

Figure 2: *Confusion matrix of the test set predictions for the Transformer-based model trained on SpeechMaturity.*

magnitudes larger than BabbleCorpus, and thus stronger performance was anticipated, SpeechMaturity also contained more variable data (46 different children in BabbleCorpus versus 96 in SpeechMaturity and 6 languages in BabbleCorpus versus more than 25 in SpeechMaturity). Thus, the stronger overall performance in SpeechMaturity suggests that rather than harming the model, the potentially increased variability in the dataset was offset by its larger size. Overall these results exemplify how Transformers' architectures allow these models to efficiently extrapolate patterns over limited labeled datasets, such as the ones employed here.

Table 3: *By-class comparison (unweighted average recall %) for the previous state-of-the-art and our Transformer-based model trained on up-sampled BabbleCorpus. Statistics refer to test sets.*

	Yeh et al. [17]	**Proposed Model**
Crying	66.8	21.7
Laughing	65.8	4.8
Canonical	67.7	47.4
Non-canonical	42.5	78.6
Junk	64.4	85.1

5. Discussion

In this study, we applied SSL techniques to the task of child speech maturity classification from two ecologically-valid child speech corpora. These corpora represent children acquiring a wide array of languages (e.g. French, Tseltal Mayan) in a variety of acoustic environments (e.g. outdoors with wind interference, indoors with electronic noise). Thus, success at this classification task would indicate an ecologically-valid robustness currently lacking in much child speech technology work trained over more limited datasets. To this task, we applied two distinct model architectures: a more traditional CNN as well as a Transformer-based system; previous work has demonstrated that the latter can be especially robust to smaller datasets [33]. We applied both model architectures to the large, comprehensive SpeechMaturity dataset (96 children and N=53,359 audio clips). When trained over this larger dataset, the Transformer-based model outperformed both the CNN trained on the same dataset, and a Transformer trained on the smaller BabbleCorpus. The combination of this algorithm and the SpeechMaturity training dataset (a dataset that did not exist when previous models for this classification task were trained) is thus the state-of-the-art for this speech classification task.

As is apparent from the by-class results (Table 3), "laughing" negatively impacted the Transformer-based model's overall performance on BabbleCorpus-UpSample, while the model's UARs for "Non-canonical" and "Junk" were actually higher than the previous state-of-the-art in [17]. Previous authors attempted to class-balance, in particular augmenting the "laughing" class. But previous models took a different approaches [17, 28, 29, 11]. For example, [17] in part augmented "laughing" by adding samples from a corpus collected using a different methodology, consisting of completely different children [30]. Because each of the previous attempts at this task took a different augmentation approach, we remained conservative and trained our BabbleCorpus models without augmentation (though we evaluated the impact of up-sampling).

Going forward, we intend to apply different data augmentation strategies to the SpeechMaturity dataset to evaluate the effects upon model performance. Deeper analysis into different augmentation strategies for the "laughing" class, which remains under-represented in our current training dataset even after the up-sampling that we employed, will be especially important since our by-class analyses suggests that the "laughing" class under-sampling may have reduced overall model performance. We additionally intend to apply a layer-wise analysis in subsequent modeling, as recent work applying models pre-trained on adult speech to child speech datasets has demonstrated that phonetic representations of child speech may be more robustly represented in middle layers than top layers [4].

Finally, we intend to evaluate model performance for specific speech corpora represented in SpeechMaturity. Two elements could potentially result in outsize- or under-performance for some models. First, approximately half of the children in SpeechMaturity were from socio-cultural backgrounds where they spent the majority of their time outside, with multiple different adult and, especially, child speakers. Thus, these children's recordings would contain larger amounts of interference from environmental noises such as wind and animals, as well as greater levels of speech overlap between multiple speakers that includes the target child. These facts could affect model performance in a corpus-specific manner (something that we could additionally examine via layer-wise analysis of acoustic features). The second reason why we should look for corpus-specific differences is that the children were exposed to a range of languages differing in syllable complexity and phonological inventory content and size (e.g. at least 42 distinct phonemes in Yélî Dnye versus just over 20 in Spanish). Given that some, but not all, of children's early speech patterning reflects ambient language characteristics in ways that are still being examined [27], it will be important to evaluate how the classification accuracies of canonical and non-canonical vary by corpus to ensure that these metrics are equally robust across languages.

6. References

[1] 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.

[2] 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.

[3] 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.

[4] J. Li, M. Hasegawa-Johnson, and N. L. McElwain, "Analysis of Self-Supervised Speech Models on Children's Speech and Infant Vocalizations," 2024.

[5] 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.

[6] 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.

[7] Y.-A. Chung and J. Glass, "Generative Pre-Training for Speech with Autoregressive Predictive Coding," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, 2020, pp. 3497–3501.

[8] 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.

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[10] R. Jain, A. Barcovschi, M. Yiwere, P. Corcoran, and H. Cucu, "Adaptation of Whisper models to child speech recognition," in *Proceedings of Interspeech 2023*, Dublin, Ireland, 2023.

[11] 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.

[12] D. Oller, *The Emergence of the Speech Capacity*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[13] 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.

[14] X. Yao, M. Micheletti, M. Johnson, E. Thomaz, and K. De Barbaro, "Infant Crying Detection In Real-World Environments," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Singapore, Singapore: IEEE, May 2022, pp. 131–135.

[15] I. Maddieson, *Patterns of Sounds*, ser. Cambridge Studies in Speech Science and Communication. Cambridge [Cambridgeshire] ; New York: Cambridge University Press, 1984.

[16] M. Schmitt and B. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[17] 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.

[18] M. Casillas, P. Brown, and S. Levinson, *Casillas HomeBank Corpus*, 2017.

[19] M. Cychosz, *Cychosz HomeBank Corpus*, 2018.

[20] E. Bergelson, *Bergelson Seedlings HomeBank Corpus*, 2017.

[21] A. Cristia and H. Colleran, *Long-Form, Child-Centered Recordings Collected in Malekula in 2016-2018*, 2018.

[22] A. Warlaumont, G. Pretzer, S. Mendoza, and E. Walle, *Warlaumont HomeBank Corpus*, 2016.

[23] C. Scaff, J. Stieglitz, and A. Cristià, *Daylong Recordings from Young Children Learning Tsimane in Bolivia*, 2018.

[24] 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 lTR-05-2, 2009.

[25] M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, and A. Cristia, "An open-source voice type classifier for child-centered daylong recordings," 2021.

[26] 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.

[27] K. Hitczenko, 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.

[28] 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.

[29] 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.

[30] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, Mar. 2017, pp. 776–780.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[32] P. Marcel, T. Pfister, S. Kornblith, T. Nando, M. Auli, A. Smola, J. Uszkoreit, and B. Steiner, "TorchVision: Datasets, Transforms, and Models for PyTorch," 2022.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Advances in neural information processing systems*, vol. 30, pp. 5998–6008, 2017.

[34] Y. Zhang, B. Li, H. Fang, and Q. Meng, "Spectrogram Transformers for Audio Classification," in *2022 IEEE International Conference on Imaging Systems and Techniques (IST)*. Kaohsiung, Taiwan: IEEE, 2022, pp. 1–6.

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books."

[36] A. Keesing, Y. Koh, and M. Witbrock, *Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech*, Aug. 2021.