The Speaker and Language Recognition Workshop (Odyssey 2024)
18-21 June 2024, Quebec City, Canada

Joint Language and Speaker Classification in Naturalistic Bilingual Adult-Toddler Interactions

Satwik Dutta[1], *Iván López-Espejo*[1,2], *Dwight Irvin*[3], *John H. L. Hansen*[1]

[1]Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, USA
[2]Department of Signal Theory, Telematics and Communications, University of Granada, Spain
[3]Anita Zucker Center for Excellence in Early Childhood Studies, The University of Florida, USA
{satwik.dutta,john.hansen}@utdallas.edu, iloes@ugr.es, dwirvin@coe.ufl.edu

Abstract

Bilingual children at a young age can benefit from exposure to dual language, impacting their language and literacy development. Speech technology can aid in developing tools to accurately quantify children's exposure to multiple languages, thereby helping parents, teachers, and early-childhood practitioners to better support bilingual children. This study lays the foundation towards this goal using the Hoff corpus containing naturalistic adult-child bilingual interactions collected at child ages 2½, 3, and 3½ years. Exploiting self-supervised learning features from XLSR-53 and HuBERT, we jointly predict the language (English/Spanish) and speaker (adult/child) in each utterance using a multi-task learning approach. Our experiments indicate that a trainable linear combination of embeddings across all Transformer layers of the SSL models is a stronger indicator for both tasks with more benefit to speaker classification. However, language classification for children remains challenging.

1. Introduction

By the year 2050, Hispanic/Latino(a) children will constitute about 30% of the total population of children under 8 in the United States[1]. There is an estimated 40 million individuals who speak Spanish in the home [1], although the amount of Spanish and English that children are exposed to at home varies widely [2]. We know that children who are second language learners can benefit from dual language exposure at young age (e.g., increased language and literacy skills in both languages, cognition) [3, 4] and that these benefits can continue on later in life (e.g., job opportunities in the interdependent global economy) [4]. We also know that when young children enter the classroom and have lower English proficiency skills they may miss opportunities to engage in talk with peers and classroom adults [5], and it is not uncommon for them to be misidentified for special education services [6]. Current tools to capture children's exposure to multiple languages appear to be researcher-developed parent-reported measures [7, 8] that are not equipped to fully capture children's exposure to English and Spanish. With the growth in the number of children who are second language learners, there is an immediate need for a tool that accurately quantifies children's exposure to multiple languages in the home and school. Such a tool could: *1)* help teachers and parents better gage Spanish *vs.* English input in the home and classroom, and *2)* cultivate/improve home-school partnerships aimed at promoting children's proficiency in both languages across settings. And, for researchers, this tool could further our understanding of how the contributions of multiple languages in different settings affect children's development. To realize such a tool and by probing into the speech processing pipeline, identifying the speakers and the language spoken in bilingual adult-child interactions are the first important steps.

There has been very recent progress in the field of language identification/classification (LID) [9–13]. While some work keeps focusing on LID from traditional Mel spectra [10] and Mel-frequency cepstral coefficients (MFCCs) [9], some other work explores the use of self-supervised learning (SSL) speech representations/embeddings as speech features [11–13], as we also do in this paper. Specifically, the authors of [11] and [12] exploit popular SSL models like standard wav2vec2 [14], XLS-R [15] or HuBERT [16]. However, [11, 12] carry out identification by *linearly* classifying speech representations out of a *single* Transformer encoder layer of the SSL model, which limits the potentials of the SSL framework. Wang *et al.* [13] deal with this limitation by studying weighted combinations of embeddings from the Transformer encoder layers that constitute the input to an ECAPA-TDNN [17] classifier. However, the latter paper falls into the area of *singing* speech processing, whereas a similar study is missing in the context of regular speech. More importantly, none of the above-referred LID works has also faced, as we do in this paper, bilingual young children's speech, whose specific properties make it worth an analysis (*young children —below 8 years— have very different spoken language skills as compared to older children [18, 19] and adults [20]*). Adult/child speaker classification is another difficult problem, which has been addressed by a recent study [21] by using adapted wav2vec2 and WavLM [22] embeddings for adult/child speaker classification.

Most prior work on language and speaker classification has considered these two tasks separately. In this paper, we use the CHILDES Spanish-English Hoff Corpus [23, 24] to perform joint language (English/Spanish) and speaker (adult/child) classification using a multi-task learning approach (see Fig. 1). Furthermore, exploiting SSL features to jointly predict the language and speaker is essentially uncharted. More importantly, to the best of our knowledge, this is the first work that explores language classification in young children's (2½ – 3½ years) speech on top of adults' speech.

2. Multi-task Classification Approach

Fig. 2 depicts a block diagram of the proposed classification system, which is discussed in the following subsections.

[1]https://www.pewresearch.org/hispanic/

10.21437/odyssey.2024-12

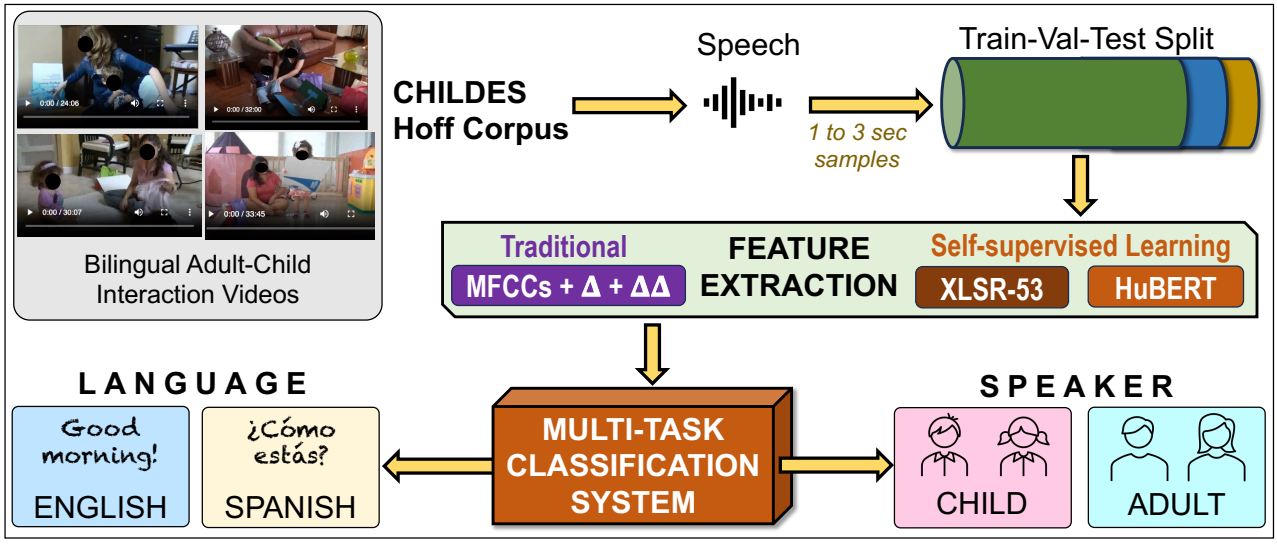

Figure 1: Overall system flow structure for multi-task language and speaker classification in bilingual adult-child interactions.

2.1. Feature Extraction Front-end

We start with 13 MFCC features and their first- and second-order derivatives. Apart from traditional MFCC features, we utilize two state-of-the-art SSL models (left part of Fig. 2): XLSR-53 [25] and HuBERT [16]. Both models consist of *1)* a CNN-based latent feature encoder, and *2)* a Transformer-based context encoder. On the one hand, similarly to wav2vec2, XLSR-53 performs a contrastive task over masked latent speech representations and jointly learns a quantization of the latents shared across languages. XLSR-53 is pre-trained using 56k hours of multilingual data from three datasets: LibriSpeech, Common Voice, and BABEL. On the other hand, HuBERT extracts the hidden units (pseudo-targets) using K-means clustering and predicts targets from the context using cross-entropy loss. This model is pre-trained on 60k hours of unlabeled audio from Libri-Light.

For both SSL models used in this paper, we employ the pre-trained 'LARGE' checkpoints from FAIRSEQ[2]. They comprise $L = 24$ Transformer encoder layers each, which output $D = 1,024$-dimensional embeddings $\mathbf{x}_l \in \mathbb{R}^D$, $l = 1, ..., L$. In this work, we compare two different types of SSL features: *1)* $\mathbf{x}_L$ (denoting the embedding from the last layer), and *2)* the embedding linear combination $\sum_{l=1}^{L} w_l \mathbf{x}_l$, where the weight set $\{w_l > 0;\ l = 1, ..., L\}$ is jointly trained along with the classification back-end.

2.2. Classification Back-end

The classification back-end (right part of Fig. 2) is primarily composed of *1)* a linear bottleneck layer followed by layer normalization, *2)* four Conformer [26] blocks, *3)* a statistics pooling layer, and *4)* two parallel branches with a stack of linear layers (with ReLU and softmax) for language (English/Spanish) and speaker (adult/child) classification. Each Conformer block comprises two feedforward layers with half-step residual connections —sandwiching the multi-headed self-attention and convolution modules— followed by layer normalization. The back-end model uses Adam as an optimizer and trains with the objective of reducing cross-entropy loss with equal weights to both classification tasks.

[2] https://github.com/facebookresearch/fairseq

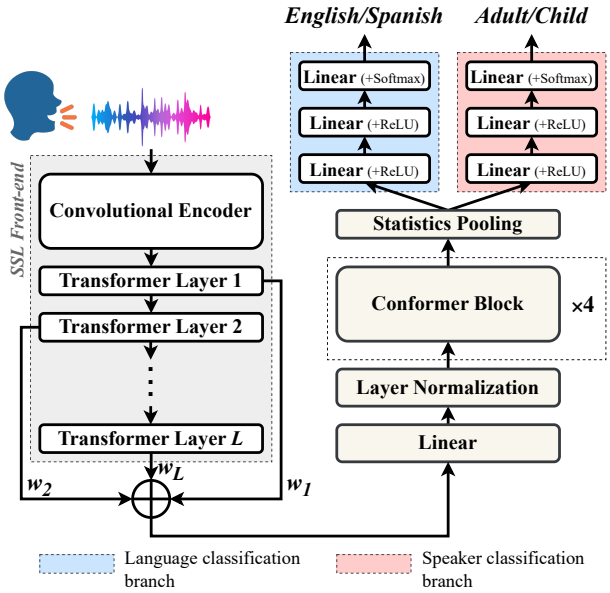

Figure 2: Block diagram of the proposed joint language and speaker classification system.

3. Corpora

For this paper, while the CHILDES Spanish-English Hoff Corpus is considered primary, we also employ a reduced version of the Common Voice (CV) dataset [27] to pre-train English/Spanish language classifiers. Specifically, for model pre-training purposes, we use 100 hours of English speech and 100 hours of Spanish speech. We then use these pre-trained models to initialize classification back-ends (except for the speaker classification branch) for further experimentation using Hoff,

Table 1: Split of the Hoff corpus into Train, Validation, and Test based on language and speaker.

Language	**Speaker**	**Train** (52 families)		**Validation** (6 families)		**Test** (6 families)	
		Utterances	*Time (hrs)*	*Utterances*	*Time (hrs)*	*Utterances*	*Time (hrs)*
Spanish	Adult	20,610	11.71	2,360	1.36	1,754	1.02
English	Adult	22,959	12.76	3,046	1.69	3,564	1.92
Spanish	Child	6,445	3.61	758	0.45	803	0.45
English	Child	10,465	5.90	923	0.49	1,437	0.76
TOTAL		60,479	33.98	7,087	3.99	7,558	4.15

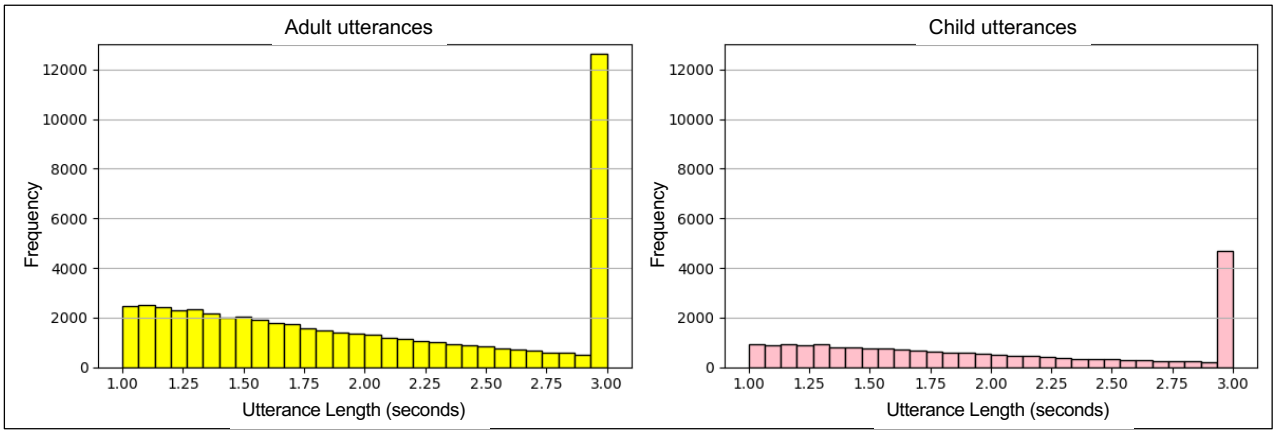

Figure 3: Distribution of the length of adult and child utterances.

which is denoted as "*CV pre-training*".

3.1. CHILDES Spanish-English Hoff Corpus (Hoff)

The CHILDES Spanish-English Hoff Corpus[3] [23, 24] contains videos of interactions between Spanish- and English-speaking caregivers and children collected at child ages ranging from 2½ to 3½ years as part of a language and literacy development longitudinal study of US-born children raised in Spanish-speaking homes. These interactions were captured during toy playing and book-reading activities, with most recordings in the participants' homes. A total of 64 families with bilingual children were considered for this study. Each family had one or more recordings collected at child ages 2½, 3, and 3½ years. While transcripts are provided, not all utterance-level timestamps are correct/available. Based on the transcripts, each downsampled audio recording (initially converted from video to 16 kHz audio) was then split into smaller utterances: *1)* utterances more than 3 s were split into samples of 3 s or less, and *2)* utterances less than 3 s were kept as they are. Note that all utterances less than 1 s were discarded. Around 80% of utterances are below 3 s. The final distribution of utterance length is displayed in Figure 3 and the experimental split is shown in Table 1. Utterances coded in the transcripts as spoken in Spanish or English were only considered, discarding the rest. The use of this corpus was approved in accordance with the TalkBank Code of Ethics[4].

[3] https://childes.talkbank.org/access/Biling/Hoff.html

[4] https://talkbank.org/share/ethics.html

4. Experimental Results and Discussion

When evaluating model performance, we use system-wide metrics such as balanced accuracy (BAcc) and equal error rate (EER). Given an experiment, the criterion for selecting the best model for testing purposes was maximum summed (across both tasks) validation BAcc when employing early-stopping with a 4-epoch patience. Language and speaker classification results are summarized in Tables 2 and 3.

From Table 2, it can be seen that, generally, the use of SSL-based features clearly outperforms the utilization of traditional speech features like MFCCs. Irrespective of the features, language classification performance is lower for children than for adults. Many factors might contribute to this: young children are still developing their speech/language skills, their speech might not have precise speech representations as expected in adults' speech, and children's spoken responses are usually shorter. That being said, note that this performance gap between children and adults tends to be larger for SSL features than for MFCCs. We reasonably hypothesize that this might be explained by the SSL models being pre-trained by means of just adults' speech.

Moreover, it is shown that learning a linear combination of embeddings out of all the Transformer encoder layers provides better language and speaker classification performance in comparison with using speech representations from the last layer of the SSL models. This outcome is significantly true for XLSR-53, but also for HuBERT. Specifically, and considering "*Hoff training from scratch*", XLSR-53_W (HuBERT$_W$) shows a BAcc relative improvement of around 8.8% and 2.7% (3.1% and 2.2%) with respect to XLSR-53_L (HuBERT$_L$) in terms of overall language and speaker classification, respectively. Sim-

Table 2: Language and speaker classification BAcc and EER results (%) on Hoff. Best results are marked in boldface.

Training	**Front-end**	**Language Classification**						**Speaker Classification**	
		Child		*Adult*		*Overall*		*Overall*	
		BAcc	EER	BAcc	EER	BAcc	EER	BAcc	EER
Hoff training from scratch	MFCC	61.36	36.17	64.81	33.11	63.74	34.08	81.27	14.12
	XLSR-53$_L$	56.09	43.37	72.28	30.34	67.27	34.36	82.46	11.38
	XLSR-53$_W$	61.43	43.78	78.06	26.23	73.17	31.54	84.71	9.35
	HuBERT$_L$	66.25	**28.13**	79.23	19.96	75.10	22.43	91.46	7.26
	HuBERT$_W$	67.55	33.28	81.65	18.61	77.44	23.01	**93.47**	6.11
CV pre-training + Hoff fine-tuning	MFCC	60.23	39.13	72.64	26.12	68.93	30.00	84.62	13.94
	XLSR-53$_L$	55.14	39.43	68.10	28.86	64.08	32.10	87.00	17.06
	XLSR-53$_W$	67.64	29.59	**83.90**	**15.28**	**78.92**	**19.57**	92.86	**5.06**
	HuBERT$_L$	63.26	30.05	79.68	16.76	74.57	20.73	90.93	7.93
	HuBERT$_W$	**67.68**	28.40	81.82	16.61	77.43	20.23	92.07	5.89

$(\cdot)_L \longrightarrow$ last layer, $\mathbf{x}_L$; $\quad (\cdot)_W \longrightarrow$ trainable linear combination of embeddings, $\sum_{l=1}^{L} w_l \mathbf{x}_l$

ilarly, such BAcc relative improvements are, approximately, 23.2% and 6.7% (3.8% and 1.3%) when considering "*CV pre-training + Hoff fine-tuning*". Note that pre-training using CV is helpful for MFCC and XLSR-53$_W$, the latter showing the best overall performance.

Table 3: Language and speaker classification BAcc and EER results (%) from comparing the use of trainable, $(\cdot)_W$, and fixed uniform, $(\cdot)_A$, embedding combination weights. Best results are marked in boldface.

Training	**Front-end**	**Language C.**		**Speaker C.**	
		BAcc	EER	BAcc	EER
Hoff training from scratch	XLSR-53$_W$	73.17	31.54	84.71	9.35
	XLSR-53$_A$	75.09	29.76	93.06	7.90
	HuBERT$_W$	77.44	23.01	**93.47**	6.11
	HuBERT$_A$	77.68	20.91	92.42	6.02
CV pre-training + Hoff fine-tuning	XLSR-53$_W$	**78.92**	**19.57**	92.86	**5.06**
	XLSR-53$_A$	77.35	20.24	92.66	5.53
	HuBERT$_W$	77.43	20.23	92.07	5.89
	HuBERT$_A$	78.67	22.08	92.67	6.17

4.1. Analysis of Embedding Combination Weights

We observed that the learnt embedding combination weights did not significantly deviate from their random initialization $w_l^{(0)} \sim \mathcal{U}(0,1)$, $l = 1, ..., L$, where $\mathcal{U}$ stands for uniform distribution. To shed some light on this behavior, we decided to test the case of simply averaging embeddings out of all the Transformer encoder layers of the SSL models (i.e., using fixed uniform weights, $w_l = 1/L \; \forall l$). From Table 3, we can see that, particularly for "*CV pre-training + Hoff fine-tuning*", learning or fixing $\{w_l;\ l = 1, ..., L\}$ does not make a significant difference. Given this outcome, we hypothesize that, as long as a powerful classification back-end is used (as in our case), w_l values are not so critical provided that multiple Transformer encoder layers contribute to shape the final speech representations.

5. Conclusion

With a slowly-increasing focus on analyzing children's speech, it is important to consider a broader cohort of subjects with respect to language diversity, as well as disability and socioeconomic status. Such steps would enable innovative solutions for parents, teachers, as well as practitioners in the US to monitor and better understand the language environments of a large young Hispanic children population in the home and classroom. We hope that this work will be the first step in this direction, and make speech systems more inclusive. Overall, our language and speaker classification solution using naturalistic bilingual adult-child interactions shows the best results using a trainable linear combination of embeddings extracted from XLSR-53. Future work will explore analyzing other corpora as well as fine-tuning SSL models using adult-child interactions.

6. Acknowledgement

This work is sponsored by the U.S. National Science Foundation under Grants #1918032 and #2234916 (Hansen), #1918012 and #2235041 (Irvin), and the Quad Fellowship (Dutta). Dr. Iván López-Espejo was supported by the Spanish Ministry of Science and Innovation under the "Ramón y Cajal" programme (RYC2022-036755-I) as part of this collaboration with CRSS-UTDallas. This study is approved by The University of Texas at Dallas Institutional Review Board under IRB-24-6. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. The authors would like to acknowledge the discussion with Dr. Erika Hoff at the Florida Atlantic University for responding to all queries related to the Hoff corpus.

7. References

[1] Mark Hugo Lopez, Ana Gonzalez-Barrera, and Gustavo López, *Hispanic Identity Fades Across Generations as Immigrant Connections Fall Away*, Pew Research Center, 2017.

[2] Lisa M. López and Matthew E. Foster, "Examining heterogeneity among Latino dual language learners' school readiness profiles of English and Spanish at the end of Head Start," *JADP*, vol. 73, 2021.

[3] Lillian Durán, Cary Roseth, Patricia Hoffman, and M Brooke Robertshaw, "Spanish-Speaking Preschoolers' Early Literacy Development: A Longitudinal Experimen-

tal Comparison of Predominantly English and Transitional Bilingual Education," *Bilingual Research Journal*, vol. 36, pp. 6–34, 2013.

[4] Ellen Bialystok, "Bilingual education for young children: review of the effects and consequences," *International Journal of Bilingual Education and Bilingualism*, vol. 21, pp. 666–679, 2018.

[5] Cecilia Jarquin Tapia, Sarah Surrain, and Stephanie M Curenton, "The Importance of Dyadic Classroom Conversations for Dual Language Learners," *The Reading Teacher*, vol. 75, pp. 777–781, 2022.

[6] Else V Hamayan, Barbara Marler, Cristina Sanchez-Lopez, and J Damico, "Reasons for the Misidentification of Special Needs among ELLs," *Special education considerations for English language learners: Delivering a continuum of services*, pp. 2–7, 2007.

[7] Kelly Bridges and Erika Hoff, "Older Sibling Influences on the Language Environment and Language Development of Toddlers in Bilingual Homes," *Applied Psycholinguistics*, vol. 35, pp. 225–241, 2014.

[8] J Marc Goodrich, Christopher J Lonigan, Beth M Phillips, JoAnn M Farver, and Kimberly D Wilson, "Influences of the home language and literacy environment on Spanish and English vocabulary growth among dual language learners," *ECRQ*, vol. 57, pp. 27–39, 2021.

[9] Victoria YH Chua, Hexin Liu, Leibny Paola Garcia Perera, Fei Ting Woon, Jinyi Wong, Xiangyu Zhang, Sanjeev Khudanpur, Andy WH Khong, Justin Dauwels, and Suzy J Styles, "MERLIon CCS Challenge: A English-Mandarin code-switching child-directed speech corpus for language identification and diarization," in *Interspeech*, 2023, pp. 4109–4113.

[10] Shashi Kant Gupta, Sushant Hiray, and Prashant Kukde, "Spoken Language Identification System for English-Mandarin Code-Switching Child-Directed Speech," in *Interspeech*, 2023, pp. 4114–4118.

[11] Shangeth Rajaa, Kriti Anandan, Swaraj Dalmia, Tarun Gupta, and Eng Siong Chng, "Improving spoken language identification with map-mix," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[12] Weiwei Lin, Chenhang He, Man-Wai Mak, and Youzhi Tu, "Self-supervised Neural Factor Analysis for Disentangling Utterance-level Speech Representations," in *ICML*, 2023.

[13] Xingming Wang, Hao Wu, Chen Ding, Chuanzeng Huang, and Ming Li, "Exploring universal singing speech language identification using self-supervised learning based front-end features," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," 2020, vol. 33, pp. 12449–12460.

[15] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al., "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Interspeech*, 2022, pp. 2278–2282.

[16] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman 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.

[17] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Interspeech*, 2020, pp. 3830–3834.

[18] Satwik Dutta, Sarah Anne Tao, Jacob C Reyna, Rebecca Elizabeth Hacker, Dwight W Irvin, Jay F Buzhardt, and John HL Hansen, "Challenges remain in Building ASR for Spontaneous Preschool Children Speech in Naturalistic Educational Environments," in *Interspeech*, 2022, pp. 4322–4326.

[19] Satwik Dutta, Dwight Irvin, Jay Buzhardt, and John HL Hansen, "Activity focused speech recognition of preschool children in early childhood classrooms," in *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 2022, pp. 92–100.

[20] Matteo Gerosa, Diego Giuliani, and Fabio Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, pp. 847–860, 2007.

[21] Rimita Lahiri, Tiantian Feng, Rajat Hebbar, Catherine Lord, So Hyun Kim, and Shrikanth Narayanan, "Robust Self Supervised Speech Embeddings for Child-Adult Classification in Interactions involving Children with Autism," in *Interspeech*, 2023, pp. 3557–3561.

[22] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE J-STSP*, vol. 16, pp. 1505–1518, 2022.

[23] Michelle K. Tulloch and Erika Hoff, "Filling lexical gaps and more: code-switching for the power of expression by young bilinguals," *J. Child Lang.*, vol. 50, pp. 981–1004, 2023.

[24] Brian MacWhinney, *The CHILDES Project: The database*, Psychology Press, 2000.

[25] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised Cross-lingual Representation Learning for Speech Recognition," in *Interspeech*, 2021, pp. 2426–2430.

[26] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Interspeech*, 2020, pp. 5036–5040.

[27] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber, "Common Voice: A Massively-Multilingual Speech Corpus," in *LREC*, 2020, pp. 4218–4222.