Abstract: Isolating Child Speech using Convolutional Neural Networks

1 Background

Gathering targeted child speech data in both research and classroom settings consequently involves the capture of non-targeted adult speech in the recordings, often as prompts for the child speaker or in conversational turn-taking tasks [1]. In order for data to be utilized in future work, a researcher needs to go through a labor-intensive identification process to isolate the child speech, constraining resources and limiting the growth of extensive child speaker datasets for speech pathology research.

Automating the process of blind source extraction of child speakers would be beneficial to researchers, but any blind source extraction technique would need to be privy to the noise of the environment, such as a classroom, that it is collected in.

The emergence of accessible deep learning toolkits over the last decade has led to rapid advancements in audio and speech application domains. Open source speaker diarization pipelines, such as OpenAI's Whisper [2], have been developed to tackle the "who said what and when" problem of audio feature extraction, but these pipelines struggle to delineate between child and adult speakers in conversational turn-taking tasks due to intrinsic and extrinsic factors that impact post-analysis of child speech.

2 Methods

We propose a CNN architecture designed for audio classification, targeted at the classification of child and adult speakers. The core of our model consists of six convolutional blocks, each of which contains a 2D convolutional layer.

Alongside the standard direct classifications of adult and child speakers of a given audio segment, we examine the impact of including other common speech occurrences in conversational turn-taking task recordings.

In this study, we train the model on the Speech Exemplars and Evaluation Database (SEED), which contains 17,000 single word and sentence utterances gathered from 69 child and 33 adult speakers, gathered in both clinical and classroom settings [5].

3 Results

A series of models were trained to best find the version most applicable to a test environment for researchers. All four model versions have achieved classification accuracy of over 80% on validation datasets through 45 epochs of training.

4 Future Work

Future work will entail benchmarking our classification architecture against a simple linear classifier on Wave2Vec extracted audio features [7], and exploring the combination of these features with those derived from the CNN architecture.

References

[1] M. Speights, K. D. Gilbert, J. MacAuslan, and R. Goldhor, "Automated episode selection of child continuous speech via blind source extraction," in J. Acoust. Soc. Am., vol. 144, no. 3_Supplement, Sep. 2018.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision."

[3] L. Nanni, G. Maguolo, S. Brahnam, and M. Paci, "An Ensemble of Convolutional Neural Networks for Audio Classification," Applied Sciences, vol. 11, no. 13, p. 5796, Jun. 2021.

[4] K. J. Piczak, "Environmental sound classification with convolutional neural networks," 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA, 2015.

[5] Speights Atkins, M., Bailey, D. J., & Boyce, S. E. (2020). Speech exemplar and evaluation database (SEED) for clinical training in articulatory phonetics and speech science.

[6] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in Interspeech 2019, Sep. 2019.

[7] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-training for Speech Recognition," 2019.