Classifying Infant Vocalizations in Audio Recordings through Transfer Learning and Image Processing Techniques

Arun Prakash Singh & Natalia Kartushina

Department of Linguistics and Scandinavian Studies, University of Oslo

The identification and classification of infants' vocalizations (into canonical vs non-canonical babbling vs adult/other) in audio recordings can play crucial role in understanding infants' language and speech development in various learning environments. Recent research indicates that deep learning models such as convolution neural network (CNN) trained from 'scratch' can be utilized to accomplish various audio classification tasks including classifying infants' vocalizations. However, to achieve high classification accuracy, models trained from scratch require large number of vocalization instances. This study addresses this limitation and examines whether pre-trained vision deep learning models, using transfer learning with spectrogram images of vocalization audio clips as input, can achieve better accuracy in classifying infant vocalizations. An open-source labelled database of infant and adult vocalization recorded in a home environment for English-speaking families (infants wore Lena vests for 12-hour periods at 3, 6, 9, and 18 months of age) was used for this task. Despite few audio clips of infant and adult vocalizations being available, the classification accuracies of various vision deep learning models, used in the current study, ranged between 64.50% to 82.77%, which is higher, as compared to previously used models when trained with same number of audio clips, i.e., between 55% and 71.42% of accuracy.