Behavior Research Methods (2023) 55:3187–3197
https://doi.org/10.3758/s13428-022-01961-x

Validating a model to detect infant crying from naturalistic audio

Megan Micheletti[1] · Xuewen Yao[2] · Mckensey Johnson[1] · Kaya de Barbaro[1]

Accepted: 18 August 2022 / Published online: 9 September 2022

Abstract

Human infant crying evolved as a signal to elicit parental care and actively influences caregiving behaviors as well as infant–caregiver interactions. Automated cry detection algorithms have become more popular in recent decades, and while some models exist, they have not been evaluated thoroughly on daylong naturalistic audio recordings. Here, we validate a novel deep learning cry detection model by testing it in assessment scenarios important to developmental researchers. We also evaluate the deep learning model's performance relative to LENA's cry classifier, one of the most commonly used commercial software systems for quantifying child crying. Broadly, we found that both deep learning and LENA model outputs showed convergent validity with human annotations of infant crying. However, the deep learning model had substantially higher accuracy metrics (recall, F1, kappa) and stronger correlations with human annotations at all timescales tested (24 h, 1 h, and 5 min) relative to LENA. On average, LENA underestimated infant crying by 50 min every 24 h relative to human annotations and the deep learning model. Additionally, daily infant crying times detected by both automated models were lower than parent-report estimates in the literature. We provide recommendations and solutions for leveraging automated algorithms to detect infant crying in the home and make our training data and model code open source and publicly available.

Keywords Infant crying · Computational model · Deep learning · Naturalistic

With the advent of ambulatory audio recorders in the past two decades, researchers can now collect high-fidelity daylong recordings of a child's everyday life. The Language Environment Analysis system (LENA; Greenwood et al., 2011) is one of the most commonly used "off the shelf" hardware and software systems for quantifying a child's audio environment (Cristia, Lavechin, et al., 2020b). One major benefit of this system is that it combines a daylong audio-recording with automated analyses to detect markers of activity relevant to child development. Although it is used primarily as a "speech pedometer" to detect adult word counts, child vocalizations, and parent–child conversational turns, it also includes algorithms that detect non-speech vocalizations, like a child laughing and crying (Greenwood et al., 2011). Given that individual differences in daily infant crying have been shown to predict infant reactivity and emotion regulation (Stifter & Spinrad, 2002), as well as caregiver mental health (Miller et al., 1993) and parenting behaviors (Barr et al., 2014), an automated algorithm to detect infant crying in the home is a powerful tool for researchers interested in child development and family systems. As a result, examining the performance of different automated cry algorithms is a critical research step. The current study extends the initial validation of a novel deep learning cry detection algorithm introduced by our team in Yao et al. (2022) by testing its performance relative to the LENA algorithm in assessment scenarios relevant to developmental researchers interested in automatically detecting infant cries from naturalistic child-centered audio.

Detecting infant crying episodes in the real-world environments where they typically occur is a challenging engineering problem. Many published cry algorithms are based on "clean" in-lab datasets where extraneous sounds are minimized. Models trained on real-world datasets generally have poor crying classification performance, and models trained on in-lab datasets do not generalize to real-world scenarios (Yao et al., 2022). Additionally, high-quality labeled real-world cry datasets have not been publicly available until recently (Yao et al., 2022). LENA's

✉ Megan Micheletti
m.micheletti@utexas.edu

[1] Department of Psychology, The University of Texas at Austin, 108 E Dean Keeton St, Austin, TX 78712, USA

[2] Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA

automated cry detection algorithm was trained and tested on real-world audio (Gilkerson et al., 2008), but a recent systematic review revealed that no external studies have reported on the validity of LENA's cry classifier (Cristia et al., 2020a). Indeed, we know of no published work to date that has reported specifically on the performance of LENA's cry algorithm, including LENA technical reports (Xu et al., 2009).

Given the limited information on the performance of LENA's cry algorithm, our team developed an open-source deep learning (DL) algorithm for detecting infant cries from daylong naturalistic audio recordings (Yao et al., 2022). Because deep learning architectures introduced in the 2010s have greatly improved classification accuracy in several domains, from natural language processing to speech recognition (Deng & Yu, 2013), it is possible that our deep learning cry detection model will outperform LENA's cry detection model, in particular because deep learning was not yet widely adopted during the development of LENA's cry algorithm in the 2000s. Yao et al. (2022) do not consider or report the accuracy of LENA in their work, nor do they compare their model's performance with LENA. Given the widespread adoption of the LENA system, we consider it valuable to examine and compare the accuracy of DL and LENA cry algorithms in scenarios important to developmental researchers.

One such scenario is the ability to detect temporally precise estimates of individual crying episodes, not just summaries of total crying time per day. For example, developmental researchers might be interested in automatically detecting a dynamic process like caregiver responsiveness to infant distress (Hubbard & van Ijzendoorn, 1991). In order to capture such a dynamic process unfolding over minutes rather than hours or days, a cry detection model would have to provide precise cry onset and offset times. Despite the importance of this possible use case, very little attention has been paid to precise infant cry onsets and offsets in the home, in part because few algorithms exist to automatically and accurately capture the timing of these episodes. A valid cry algorithm should perform well at timescales ranging from minutes to hours.

Another scenario important to researchers is the ability to detect objective estimates of crying across development. Objective reports are valuable because the majority of infant cry estimates in the literature are subjective parent reports (Wolke et al., 2017) or manual annotations from audio recordings (Barr et al., 1988). Manually annotating infant cries is a labor-intensive process not scalable to large volumes of data and subjective reports of crying are often biased. For example, parent reports have been shown to differ significantly from manual annotations, sometimes up to 60 min per day (Cabana et al., 2021; Salisbury et al., 2001). Parent mental health is also associated with reports of infant negative emotionality, such that mothers with higher levels of depression and anxiety report more infant crying (McGrath et al., 2008; Petzoldt et al., 2014), although the directionality and strength of this relationship has long been under investigation (Richters & Pellegrini, 1989). For example, infants of depressed mothers may cry more than infants of non-depressed mothers, exacerbating challenges to maternal mental health (Milgrom et al., 1995). Additionally, the trajectory of infant crying across development is thought to peak at 6 weeks with approximately 120 min of crying per day before falling to less than 60 min per day by 10 weeks (Barr, 1990; Wolke et al., 2017). However, this trajectory is based on parent report, and no study to date has confirmed this developmental trend in infant crying using automated model outputs.

The present study focuses on evaluating the performance of two cry algorithms – LENA and our novel DL model introduced in Yao et al. (2022) – in assessment scenarios relevant to developmental researchers. In particular, given the importance of accurate cry estimates across short and long timescales, we examine individual differences in infant crying at multiple timescales including 24 h, 1 h, and 5 min. We also evaluate the second-by-second accuracy of model-based cries. Lastly, we compare LENA and DL cry outputs to objective human annotations and to previously reported large-scale parent reports of infant crying in the literature to examine the extent to which these various methods may alter what is known about the trajectory of daily infant crying across development.

Methods

Audio data were collected in the context of a broader study leveraging wearable sensors to examine the dynamics of mother-infant interaction in the home (de Barbaro et al., 2022, manuscript submitted for publication). Mother–infant dyads were recruited via convenience sampling from the Austin, TX metropolitan area. We advertised the study using Facebook, online university event calendars, and fliers posted in community health centers. As our broader study was designed to examine the dynamics of typical mother–infant interaction in the home, infants with congenital birth defects or a genetic condition as well as dyads with a primary residence located over 20 miles from the university were excluded (as we needed to travel to homes to deliver sensors). Multiple birth infants were excluded because their audio could capture multiple infants' crying and not be comparable to other infants.

Study protocol

First, a 90-min study introduction was conducted in participants' homes. The study was conducted in English or Spanish depending on the household's primary language. Research assistants showed mothers how to use multiple wearable devices, including the LENA device, a small child-safe recorder worn in a vest over the infant's clothing. The LENA device captures the infant's voice as well as other voices and sounds in the environment up to 10 feet away. Mothers were instructed to remove the vest during infant bathing and sleep, but keep it next to the infant and recording. Research assistants videotaped mothers and infants completing a series of structured and unstructured assessments while wearing their sensors.

At the end of the introductory session, mothers were provided with two LENA 2.0 devices and a charging station. LENA 2.0 has increased storage capacity and can record up to three 24-h recordings on one device. Mothers were asked to record for an additional 72 h over the course of the following week, with at least one continuous 48-h period during which they planned to be with their child (i.e., the child was not at daycare, typically the weekend). This 72 h is the corpus from which we selected recordings for the present study. Parents were instructed to follow their daily routines, including leaving the home if needed. To maximize protocol adherence, mothers selected their recording windows in advance and we sent text message reminders to record. Following the session, mothers were compensated $100 for their participation and received a small gift for their infant.

Participants

Eighty-seven dyads were enrolled in the study and audio data were collected from $N = 77$ dyads. Given that crying rates have been shown to vary over the course of the day (James-Roberts & Halil, 1991), we chose to analyze 24-h recordings in this study to make daylong crying estimates comparable across participants. At least one continuous 24-h recording was collected from 55 participants. We selected a subset of 24-h recordings to include in the present study given the time-intensive nature of annotating infant cries. We stratified participants with 24-h recordings by age and then quasi-randomly sampled participants to ensure a broad range of infant ages. From $N = 55$ recordings, we annotated crying in a final sample of $N = 27$ participants resulting in 27 annotated 24-h recordings (648 h in total). Table 1 depicts sample characteristics.

Table 1 Participant characteristics ($n = 27$)

	n (%)	M (SD), range
Mother age, years		30 (5), 22–42
Infant age, months		3.7 (1.7), 0.9–7.0
Infant gestational age, weeks		38 (2), 31–41
Infant sex, female	15 (55%)	
Race/Ethnicity		
Black	2 (7%)	
Black, Hispanic	4 (15%)	
Hispanic	2 (7%)	
Multiracial	1 (4%)	
White, Hispanic	4 4 (15%)	
White, Non-Hispanic	14 (52%)	
Maternal Education		
High school or less	3 (11%)	
Some college or trade school	6 (22%)	
College	9 (33%)	
Graduate School	9 (33%)	
Family Status		
Married	22 (82%)	
Single Parent	2 (7%)	
Living with a partner without marriage	3 (11%)	
Household Income		
Under $25k	2 (7%)	
$25k–49k	4 (15%)	
$50k–74k	6 (22%)	
$75k–99k	7 (26%)	
Over $100k	8 (30%)	
Primary Household Language		
English	24 (89%)	
Spanish	3 (11%)	
Number of other children in the home		1 (1), 0–5

Cry detection

Human annotations Eight trained research assistants (RAs) used ELAN (https://archive.mpi.nl/tla/elan) to annotate infant crying from the WAV audio output from LENA. For annotation purposes, individual RAs were tasked with coding a non-overlapping continuous 6-h section of each participant's 24-h audio recording (vs. coding a continuous 24 h of the same participant, which could lead to coding fatigue). This meant that up to four RAs could code each 24-h recording. Similar to existing coding schemes (Hubbard & Van Ijzendoorn, 1991), cries had a minimum duration of 3 s and were combined if within 5 s. Fusses did not have a minimum duration. All neighboring crying and fussing sounds occurring within 5 s of one another were combined, and fussing and crying annotations were collapsed into a single category labeled "crying." All other sounds and silence were collapsed into a second category labeled "not crying." To test RA interrater reliability, a subset of 6-h sections (n = 12, 72 h in total) were coded by multiple RAs and their codes were compared. RAs achieved a Cohen's kappa score of 0.78. This value corresponds to 95% observer accuracy for a "highly variable" two-class coding scheme, i.e., a coding scheme with two annotations in which one annotation is much more prevalent than the other (Bakeman & Quera, 2011, p. 165). Given that "crying" (vs. "not crying") was annotated only 4% of the time on average, our coding scheme meets the criteria for a highly variable coding scheme.

LENA Recordings were processed via LENA software to obtain episodes of infant crying, which included cries and fusses (Gilkerson et al., 2017). As detailed above, LENA has not published accuracy data on their cry algorithm. However, "crying" is a sub-category of "child vocalization" in LENA's algorithm, so we report briefly on the accuracy of LENA child vocalization markers below. In a recent systematic review of studies validating LENA, Cristia et al. (2020a) reported an average correlation of r = .77 between LENA and human annotations of child vocalizations from N = 5 studies. In an extensive, independent evaluation of LENA's key markers, Cristia et al. (2020b) found a strong association between child vocalization counts identified by LENA and human annotations (r = .65 in clips with some speech). They found that on average LENA missed four child vocalizations per 1–2-minute clip, resulting in an error rate of – 47%, suggesting LENA had a tendency to underestimate child vocalization counts in each clip relative to human annotations.

Deep learning (DL) model Our team developed a DL cry detection model in order to improve real-world infant cry detection. We trained a support-vector machine (SVM) classifier using a combination of acoustic features and deep spectrum features generated from a modified AlexNet (see Yao et al., 2022 for full model details). AlexNet is a popular convolutional neural network (CNN) architecture with five convolutional layers and three fully connected layers (Krizhevsky et al., 2012). We modified the input and output layer to accommodate the size of mel-scaled spectrograms of the audio (used as input) and our case of binary classification. Batch normalization was also included to facilitate training.

The DL model was trained on a balanced subset of 66 h of data sampled from 24 participants' audio recordings, including 7.9 h of annotated crying (training dataset kappa score: 0.85). In Yao et al. (2022), we tested our DL model on the training dataset using leave-one-participant-out cross validation, as is standard in the machine learning community. Additionally, to test the performance of the model on raw, continuous audio data, which represent its true use-case, we also tested model accuracy on a secondary testing dataset of 17 non-overlapping participants' continuous raw 24-h audio recordings (secondary testing dataset kappa score: 0.80). Results showed that the model performed well on both the sampled dataset and the 24-h continuous test dataset (F1= .61 and .61, respectively).

The present study compares LENA and DL cry detection models across 27 participants, eight of which overlap with the participants included in the training data in Yao et al. (2022). In order to test the DL's predictive performance on unseen data, we used a leave-one-participant-out method to retrain the DL model for each of the eight overlapping participants by excluding that participant from the training dataset. Similar to the human annotations, the DL model produces two labels: "crying" and "not crying". Not crying includes all non-cry sounds and silence. For purposes of comparison, we translated LENA outputs into the same "crying" and "not crying" classifications.

Data analysis

First, we evaluated the 24-h DL and LENA model results for outliers, defined as more than 1.5*IQR above the third quartile or below the first quartile. We decided to employ a conservative strategy and remove outliers given the potential for model-based outputs to suggest unrealistically high amounts of crying (Gilkerson & Richards, 2020). The DL model outputs for two participants (4.4-month-old female, 2.3-month-old female) were determined to be outliers relative to other DL outputs with 653 and 164 min of crying in 24 h, respectively. Relative to the other LENA outputs, the same 2.3-month-old female had LENA model output that met outlier criteria with 41 min of crying in 24 h. The n = 2 DL and n = 1 LENA model outputs at 24 h, 1 h, and 5 min were excluded from subsequent analyses. Although outlier exclusion is not common in machine-learning papers, we excluded outliers here to show researchers how to make practical use of models on a corpus of child-centered audio and illustrate the effect of outliers on different performance metrics.

Next, we evaluated the accuracy of cry onsets and offsets detected by LENA and DL models. We selected accuracy metrics common in both psychology and machine learning communities, namely, Cohen's kappa, F1, precision, and recall. We used Cohen's kappa to consider the moment-by-moment accuracy of each model's outputs relative to human annotations using a statistic familiar to the psychology community. We also calculated classification accuracy (F1), positive predictive value (precision), and sensitivity (recall) at each timescale of interest. At the 1-h and 5-min timescales, where we had an F1, precision, and recall value at each time interval, we calculated average values per participant before averaging across all participants to minimize possible between-person effects. We also tested infant age as a predictor of accuracy, specifically Cohen's kappa. All accuracy metrics were calculated at the second level and provide a fine-grained analysis into model performance.

Next, we calculated Pearson correlation coefficients for crying (in minutes) between human annotations and our LENA and DL models. Although our data violated the assumptions for Pearson's correlation (specifically normality and homogeneity of variance at the 1-h and 5-min timescales), this is likely because of the high frequency of zero values at these timescales. Visual inspection of our raw data revealed that the relationship between human-annotated estimates and cry algorithm estimates was linear and consistent across the span of data (i.e., we did not observe a monotonic relationship), so we report Pearson's R rather than a non-parametric correlation. Pearson correlation coefficients were calculated for human vs. LENA crying and human vs. DL crying at 24 h, 1 h, and 5 min. Similar to above, at the 1-h and 5-min timescales, we calculated correlation coefficients per participant before averaging across participants. By calculating correlation coefficients in addition to standard accuracy metrics, we were able to evaluate the extent to which models agreed with human annotations for each infant, rather than average model performance across the entire sample.

Lastly, we compared the amounts of crying detected by our human, LENA, and DL models to existing parent reports in the literature. Parent-reported crying was obtained from an international meta-analysis conducted by Wolke et al. (2017), which showed that infant crying peaks at 6 weeks and falls by 10 weeks. To examine this trend in our data, we grouped our participants into two age bins (younger than 10 weeks and 10 weeks or older) and compared daily crying time (in minutes) across groups.

Results

Moment-by-moment accuracy

We report the second-by-second accuracy metrics of the DL and LENA model in Table 2. The DL model achieved mean F1 = 0.59, with precision = 0.62 and recall = 0.62. The mean kappa score between human and DL-detected cry episodes was 0.53 (SD = 0.23, range 0.06–0.79), corresponding to 90–95% agreement between the ground truth and our DL model using the Bakeman and Quera (2011) criteria for highly variable two-class coding schemes, as described above. LENA achieved a mean F1 = 0.17, with precision = 0.81 and recall = 0.10. LENA achieved a mean kappa score of 0.19 (SD = 0.12, range 0.02–0.54), corresponding to 80% agreement using the same criteria. Figure 1 depicts a representative time series of infant cry episodes across models, selected because it had kappa values consistent with the average value observed across participants. Of note, including outliers did not meaningfully change DL or LENA average F1, precision, recall, or kappa values (see Table 2). However, outliers did affect correlation results (see below).

Correlations

Figure 2 depicts plots comparing human-annotated, DL, and LENA cry amounts at 24 h, 1 h, and 5 min. Correlations at 24 h revealed the DL model outputs to be strongly correlated (r = .70) with human annotations, accurately estimating the duration of infant crying to within 1 min per day (overestimating by only 35 s per day, on average).

Table 2 Deep learning (DL) and LENA model performance relative to human annotations

	Precision	Recall	F1	Cohen's kappa
24-hour model performance with outliers				

DL (n = 27)	M = 0.60, SD = 0.23
	Range = 0.07–0.90
		M = 0.64, SD = 0.20
		Range = 0.08–1
			M = 0.58, SD = 0.21
			Range = 0.12–0.83
				M = 0.52, SD = 0.24
				Range = 0.07–0.79

LENA (n = 27)	M = 0.82, SD= 0.28
	Range = 0–1
		M = 0.11, SD = 0.11
		Range = 0–0.45
			M = 0.19, SD = 0.15
			Range = 0–0.61
				M = 0.20, SD = 0.13
				Range = 0.02–0.54

24-hour model performance excluding outliers				

DL (n = 25)	M = 0.62, SD = 0.21
	Range = 0.10–0.90
		M = 0.62, SD = 0.19
		Range = 0.08–0.85
			M = 0.59, SD= 0.2
			Range = 0.14–0.83
				M = 0.53, SD = 0.23
				Range = 0.06–0.79

LENA (n = 26)	M = 0.81, SD = 0.28
	Range = 0–1
		M = 0.10, SD = 0.09
		Range = 0–0.45
			M = 0.17, SD= 0.13
			Range = 0–0.61
				M = 0.19, SD = 0.12
				Range = 0.02–0.54

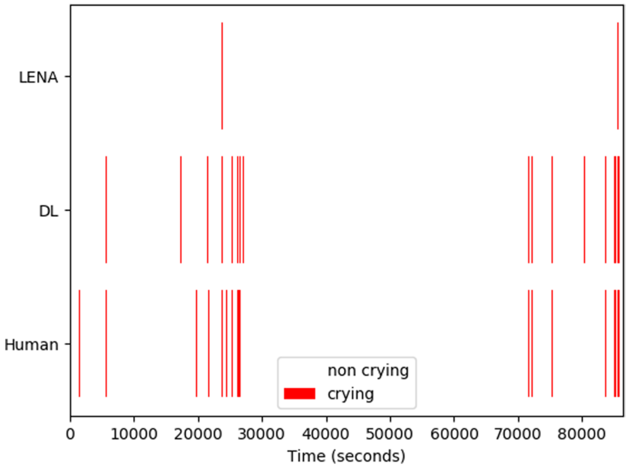

Fig. 1 Representative time series of infant crying detected by LENA, deep learning (DL), and human annotations. This 7-month-old female was selected because individual kappa scores (DL = 0.65, LENA = 0.19) were representative of the broader sample's average kappa scores (DL = 0.53, LENA = 0.19). Model accuracy metrics calculated for this 24-hour recording: DL F1 = 0.66, precision = 0.67, recall = 0.64; LENA F1 = 0.16, precision = 0.96, recall = 0.09

Correlations at 1 h (r = .86) and 5 min (r = .79) revealed very strong correlations with human annotations. Model predictions were accurate to within seconds of ground truth data (underestimating infant crying by only 5 s per 1 h and 0.6 s per 5 min, on average). Evaluating correlations between human annotations and LENA, we found strong correlations at 24 h (r = .62) and 1 h (r = .75). On average, LENA underestimated infant crying by 51 min in 24 h and 2 min in 1 h. LENA was moderately correlated with human annotations at 5 min (r = .58), underestimating infant crying by 10 s on average. All reported correlations were positive and significant at the < .001 level. Mean, standard deviation, and range of crying outputs in minutes at each timescale are presented in Table 3.

Although outliers did not change DL accuracy metrics, they did impact DL correlations. We observed weaker correlations between DL and human annotations at the 24-h timescale when including the n = 2 outliers detected in the DL output (r = .13 vs .70 at 24 h, r = .85 vs. .86 at 1 h, and r = .78 vs. .79 at 5 min), suggesting the importance of testing for and removing outliers when using this model. In contrast, we did not observe meaningful changes in the correlations between LENA and human annotations across timescales when including the n = 1 outlier detected from LENA output (r = .66 vs. .62 at 24 h, r = .76 vs. .75 at 1 h, and r = .59 vs. .58 at 5 min).

Developmental trends

Figure 3 shows developmental trends in amount of daily crying for human, DL, LENA, and parent-report (estimates drawn from Wolke et al., 2017). Consistent with parent-report estimates, we found that infants cried significantly more when less than 10 weeks compared to 10 weeks or more in human (B = – 23.45, p < .001), LENA (B = – 2.65, p = .047), and DL (B = –15.09, p = .031) outputs. However,

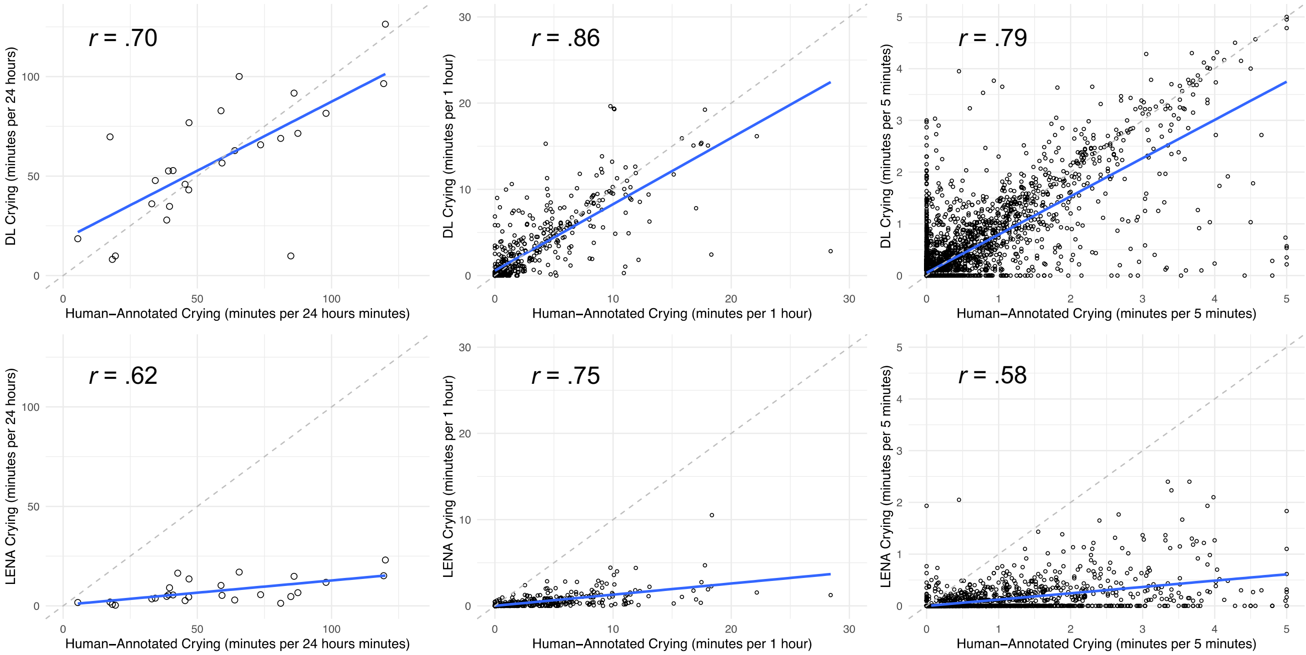

Fig. 2 Correlations of LENA and deep learning (DL) cry durations with human-annotated cry durations at varying timescales

Table 3 Mean (SD) and range of cry amounts (in minutes) across timescales

	24 h	1 h	5 min
Human	58.6 (31.9), 5.5–120.0	2.4 (3.9), 0–28.4	0.2 (0.6), 0–5
DL	57.5 (30.4), 8.2–126.3	2.4 (3.7), 0–19.7	0.2 (0.6), 0–5
LENA	7.2 (5.8), 0.3–23.0	0.3 (0.8), 0–10.5	0.02 (0.1), 0–2.2

unlike Wolke, our participants did not show a peak in crying at 6 weeks in human, DL, or LENA outputs.

LENA consistently underestimated daily infant crying relative to human, DL, and Wolke outputs from 3 to 30 weeks. To examine if age systematically predicted model performance, we tested age as a predictor of Cohen's kappa score. For the LENA model, age did not significantly predict kappa scores ($B = -0.003$, $p = .852$). For the DL model, younger infants had higher kappa scores than older infants ($B = -0.061$, $p = .019$). Figure 4 depicts kappa scores by age, showing that the DL model performs better for younger infants relative to older infants and, on average, the DL model performs better than the LENA model for all ages.

Discussion

The present study demonstrates the validity of a novel deep learning model for detecting infant crying from child-centered audio recordings. Broadly, we found that both LENA and DL models showed at least "acceptable" convergent validity ($r > .50$) with human annotations of infant crying at each timescale (Abma et al., 2016). However, the DL model published by Yao et al. (2022) dramatically outperformed LENA in moment-by-moment accuracy, with recall values nearly six times higher than LENA. LENA's low recall values led to a dramatic underestimation of crying relative to human coders, missing almost one hour of crying per day. Below, we review these results in detail and provide practical recommendations for researchers interested in automated cry algorithms.

Although both DL and LENA model outputs captured individual differences in infant crying, we found stronger correlations for DL (vs. LENA) at each timescale. LENA showed acceptable convergent validity (as defined by $r > .50$) at all timescales and the DL model displayed strong convergent validity (as defined by $r > .70$) at all timescales (Abma et al., 2016). Correlations were highest at 1 h, then 5 min, then 24 h for the

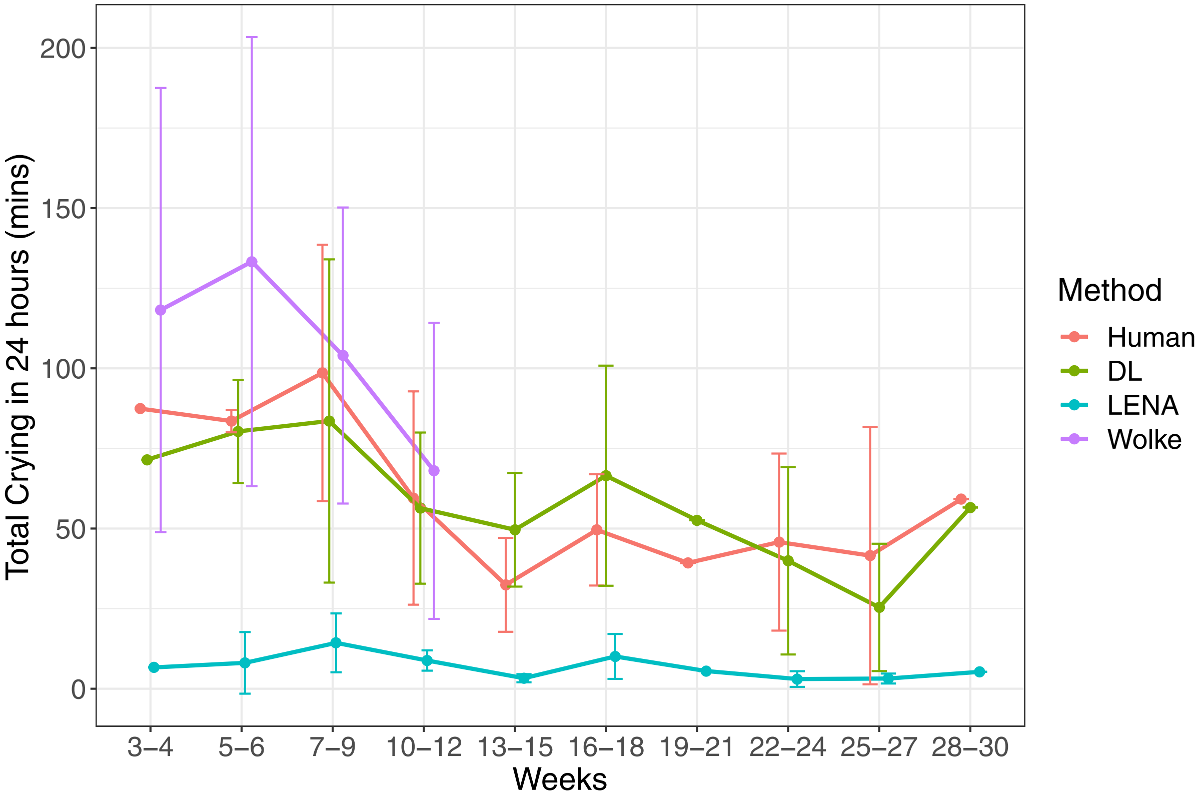

Fig. 3 Developmental trend in crying across cry detection methods. Mean and standard deviation of minutes crying per day by cry method. Sample size by weeks for human, deep learning (DL) and LENA cry detection methods: 3 to 4 weeks ($n = 1$), 5 to 6 ($n = 2$), 7 to 9 ($n = 3$), 10 to 12 ($n = 3$), 13 to 15 ($n = 3$), 16 to 18 ($n = 5$), 19 to 21 ($n = 1$), 22 to 24 ($n = 3$), 25 to 27 ($n = 3$), 28 to 30 ($n = 1$). Wolke et al. did not report cry values past 12 weeks

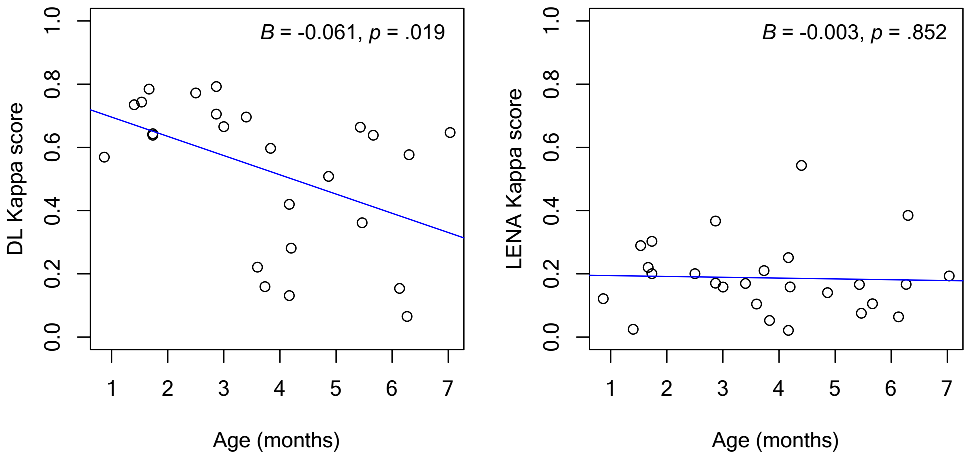

Fig. 4 Kappa scores by age for deep learning (DL) and LENA cry models

DL model. For LENA, correlations were highest at 1 h, then 24 h, then 5 min. Correlations may have been highest at 1-h timescales because misclassifications were less apparent than in 5-min timescales and variability in correlation coefficients was averaged out compared to the 24-h timescales. Although LENA had acceptable correlations, it underestimated infant crying at each timescale, missing 51 min of crying every 24 h. This shows that despite being able to detect relative differences in infant crying across participants, LENA performs poorly when tasked with detecting total amounts of crying.

Although not specific to crying, Cristia et al. (2020b) found a similar correlation coefficient between human and LENA child vocalization counts ($r = .65$), with LENA underestimating child vocalizations relative to humans. This is consistent with our pattern of results and the correlation we observed ($r = .62$) between human and LENA cries across 24-h recordings. Given the relatively high correlations between summaries of LENA outputs and human annotations, researchers using LENA outputs could improve their crying estimates at all timescales (and in particular, at 24 h) by applying a linear transformation to their data (see for example, de Barbaro et al., 2022, manuscript submitted for publication). In particular, the regression coefficients we derived when predicting human annotated crying from LENA cry outputs can be used to “transform” LENA outputs into more accurate estimates of true crying time. For example, if LENA reports 10 min of crying in a 24-h recording, using the regression equation derived from our results would lead to an updated estimate of 64 min of crying per day. Our data suggest that this would be much closer to the true crying times annotated by trained researchers. Table 4 presents regression equations to transform LENA cry outputs summarized from 24-h, 1-h, and 5-min recordings.

The DL model had much stronger moment-by-moment accuracy compared to LENA (kappa = .53 vs. .19, observer accuracy = 90–95% vs. 80%, F1 = .59 vs .17, recall = .62 vs .10, respectively). Note that the Cohen’s kappa values for both models are low because this accuracy statistic penalizes agreement that is “expected”, or could occur by chance (i.e., both models detecting “not crying” because it occurred 96% of the recording time, on average). Simulations have established that kappa values for highly variable, two-class annotated data like our own are substantially diminished relative to less variable higher-class data, even when coding accuracy is held constant (as detailed in Bakeman & Quera, 2011). Across all accuracy metrics, however, the DL model performed better than LENA, indicating it is particularly suited to developmental researchers interested in dyadic or time-locked analyses, for example, where accurate cry detection at short timescales is critical. Confusion matrices revealed that LENA did not consistently mislabel true cries with any particular label, but rather classified them variously (Figure S1), with overlapping sounds (OLN) and other child (CXN) labels occurring most frequently. Importantly, LENA’s low recall suggests that LENA did not capture approximately 90% of true cries, which is likely leading to the underestimates of overall cry detection. Despite its low recall, LENA had higher precision than our DL model (precision = .81 vs. .62, respectively), meaning that when crying was predicted by LENA it was correct a majority of the time, which is consistent with the LENA developers’ focus on precision over sensitivity (Gilkerson & Richards, 2020). However, the balance between LENA’s precision and recall was poor, meaning that the model was overly cautious in its prediction

Table 4 Regression equations to transform LENA data

Timescale	Regression equation [a]
24 h	y = 3.0675x + 34.412
1 h	y = 3.277x + 1.3154
5 min	y = 2.7567x + 0.1376

[a] x = LENA minutes of crying, y = human-annotated minutes of crying

strategy. For these reasons, kappa also showed lower agreement between LENA and human codes relative to DL.

Of note, we did observe inconsistencies in model performance across individuals in both models (see Table 2 for ranges). It is possible that some participants have higher levels of noise in their audio, which could contribute to weaker model performance. For example, the $n = 2$ outliers we identified from DL outputs had high levels of ambient noise in their recordings, which likely interfered with the model's performance and led to an overestimation of infant crying in both cases. Of note, these two outliers did not influence DL model accuracy metrics (precision, recall, F1, kappa), but did weaken convergent validity to human annotations. Although outliers may not be detected using traditional machine-learning performance metrics, model predictions are susceptible to noise in the data. As such, we recommend that researchers at least investigate outliers in their model results from external child-centered audio corpora.

We also found that infant age contributed to inconsistencies in model performance in the DL model, where kappa scores were relatively higher for younger vs. older infants (Fig. 4). In particular, for researchers working with infants under 3 months of age, LENA may not capture a large proportion of infant cries. Additionally, the DL model appears particularly well suited for infants less than 3 months of age. These data suggest that younger infants may have different acoustic properties of crying relative to older infants. However, although age predicted DL kappa scores, we found that kappa scores were higher on average for the DL model vs. LENA across our 1- to 7-month-old sample, suggesting that the DL model still outperforms LENA across development.

Results across both models suggested that crying time decreased after the first 10 weeks of life, consistent with previous reports (Wolke et al., 2017). Notably, the crying times recorded by both automated models were substantially lower than those from parent report published in recent studies. For example, Wolke et al. (2017) reported mean daily crying times up to 130 min/day (M = 133.3, SD = 70.1) in the first 10 weeks of life. The cry algorithms we tested, as well as our human annotators, detected cry durations closer to 60 min/day (M = 61.8, SD = 42.2). This finding expands upon previous work that parents overestimate crying relative to trained human annotators by a factor of two or three (Cabana et al., 2021) by replicating this pattern of results with an automated cry algorithm. Practically speaking, this could mean that the thresholds for colic (typically greater than 3 h per day) should be adjusted when automated tools are used to determine daily cry duration (Wessel et al., 1954).

Practical considerations

Ultimately, researchers should consider their research question when deciding how important it is to use the DL model vs. LENA on child-centered audio. Although we recommend using the DL model as it outperformed LENA in accuracy metrics and correlations at all timescales across development, we found that both DL and LENA models can detect crying to an "acceptable" degree of convergent validity to human annotations. This means that model outputs from either the DL or LENA model (only after LENA outputs have been transformed to increase crying estimates) could be used in a number of research situations. For example, both models could estimate amount of infant crying, providing objective reports for infant temperament or colic (Barr et al., 1992) or comparing groups of babies in crying volume, like infants of mothers with depression or anxiety relative to community controls. Both models could also be used to detect relative differences in crying between or within participants, for example, testing the hypothesis that relatively higher values of exposure to infant crying are related to higher levels of parent anxiety (Brooker et al., 2015). In contrast, only the DL model could be used to accurately detect the majority of individual cry episodes in a 24-h, 1-h, or 5-min recording. This is particularly important in situations where onset and offset times of crying are critical to the research question, like maternal sensitivity to infant distress, or if researchers are interested in characterizing "typical" parental responses to randomly selected crying episodes. In addition, researchers working with infants under 10 weeks of age should use the DL model as it outperforms LENA at all ages, but especially in younger infants.

Although we validated the DL model across a number of scenarios in the present study, we recommend additional testing to examine its generalizability to other child-centered audio corpora. A model is only as good as its training data, and performance in one sample may not generalize to other potentially distinct samples. Systematic differences in the acoustics of crying or the structure of "non-cry" background noise – related to children's age, household language, or social-economic circumstances – could all affect model performance. For example, neonates show native language-specific differences in the structure of their cry melodies (Prochnow et al., 2019). As such, we note that our DL model (Yao et al., 2022) was tested on audio collected by 1- to 7-month-old infants from mostly English-speaking homes, who were 52% non-Hispanic White (15% Hispanic White, 7% Hispanic, 7% Black, 15% Hispanic Black, 4% multiracial) and whose mothers had a relatively high educational attainment (33% graduate school, 33% college diploma, 22% some college, 11% high school diploma or less). Our testing data was 89% English-speaking and 11% Spanish-speaking. Although our model did not perform systematically worse on our Spanish-language infants, the extent to which melodic differences (present in other languages or in a larger Spanish-speaking sample) may impact the DL model's performance is unknown. Further evaluating the generalizability of the DL model to other samples is a worthwhile next research step.

Our sample characteristics could also affect differences in the reported accuracy between the LENA and DL model. LENA was trained and tested on a broader age range of children (trained on 1- to 42-month-olds and tested on 2- to 36-month-olds;

Gilkerson et al., 2008) and the DL model was trained and tested in a sample with a higher proportion of highly educated mothers. Mismatches between the LENA training data and the testing dataset used in the present study could contribute to LENA's relatively worse performance on our testing dataset. However, given that our age ranges overlapped with LENA's training data, it is not unreasonable to assess LENA's performance in this younger sample. Moreover, given the importance of crying during infancy in particular (Wolke et al., 2017), we believe the training and testing datasets presented here represent a sample that will be of relevance and interest to the developmental community.

More broadly, we recommend that models trained on one sample be tested and validated on distinct samples, different in infant age or family language for example, as model performance may differ across these samples owing to potential differences in the structure of cries to be learned. As with all models, validating the DL model across different child-centered audio corpora with families diverse in language, sociodemographic factors, family structures, and home environments is necessary to verify its generalizability.

Conclusions

Automated measures of infant crying afford researchers the opportunity to systematically access an everyday, ecologically important signal in its naturalistic setting. In the current study, we validated a deep learning model for detecting infant crying from child-centered audio and showed how it dramatically outperforms LENA in real-world assessment scenarios important to developmental researchers. We also present our training data and open source code to support future model improvements. By leveraging wearable devices and automated algorithms to detect infant crying in its natural setting, we can better access the dynamics of early development to support infants and families.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-022-01961-x.

Code availability Deep learning model code is available at https://github.com/AgnesMayYao/Infant-Crying-Detection.

Author Note This work was supported by the National Institute of Mental Health K01 Award (1K01MH111957- 01A1) to Kaya de Barbaro and a National Science Foundation Graduate Research Fellowship (DGE-1610403) to Megan Micheletti. Additional thanks to Nina Nariman, Brooke Benson, and the families that participated in our research.

Author's contributions K. D. designed the study. M.M., M.J., and K.D. collected and processed data. X.Y. designed the computational model. M.M., X.Y., and K.D. analyzed and interpreted data. M.M. wrote the manuscript with input and approval from all authors.

Funding This work was supported by the National Institute of Mental Health K01 Award (1K01MH111957-01A1) to Kaya de Barbaro and a National Science Foundation Graduate Research Fellowship (DGE-1610403) to Megan Micheletti.

Data availability Training data is available in the de Barbaro Cry Corpus at https://homebank.talkbank.org/access/Password/deBarbaroCry.html.

Declarations

Conflict of interest/Competing interests The authors report no conflicts of interest to disclose.

Ethics approval All procedures were in accordance with the 1964 Helsinki Declaration and its later amendments. The study was approved by the Institutional Review Board of the University of Texas at Austin (ethics approval number: 2017-06-0026). No experiments were preregistered.

Consent to participate Informed consent was obtained from legal guardians.

Consent for publication Legal guardians signed informed consent regarding publishing of their data.

References

Abma, I. L., Rovers, M., & van der Wees, P. J. (2016). Appraising convergent validity of patient-reported outcome measures in systematic reviews: Constructing hypotheses and interpreting outcomes. *BMC Research Notes, 9*(1), 226. https://doi.org/10.1186/s13104-016-2034-2

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. Cambridge University Press.

Barr, R. G. (1990). The normal crying curve: What do we really know? *Developmental Medicine & Child Neurology, 32*(4), 356–362. https://doi.org/10.1111/j.1469-8749.1990.tb16949.x

Barr, R. G., Kramer, M. S., Boisjoly, C., McVey-White, L., & Pless, I. B. (1988). Parental diary of infant cry and fuss behaviour. *Archives of Disease in Childhood, 63*(4), 380–387. https://doi.org/10.1136/adc.63.4.380

Barr, R. G., Rotman, A., Yaremko, J., Leduc, D., & Francoeur, T. E. (1992). The crying of infants with colic: A controlled empirical description. *Pediatrics, 90*(1), 14–21. https://doi.org/10.1542/peds.90.1.14

Barr, R. G., Fairbrother, N., Pauwels, J., Green, J., Chen, M., & Brant, R. (2014). Maternal frustration, emotional and behavioural responses to prolonged infant crying. *Infant Behavior and Development, 37*(4), 652–664. https://doi.org/10.1016/j.infbeh.2014.08.012

Brooker, R. J., Neiderhiser, J. M., Leve, L. D., Shaw, D. S., Scaramella, L. V., & Reiss, D. (2015). Associations between infant negative affect and parent anxiety symptoms are bidirectional: Evidence from mothers and fathers. *Frontiers in Psychology, 6*, 1875. https://doi.org/10.3389/fpsyg.2015.01875

Cabana, M. D., Wright, P., Scozzafava, I., Olarte, A., DiSabella, M., Liu, X., & Gelfand, A. A. (2021). Newborn daily crying time duration. *Journal of Pediatric Nursing, 56*, 35–37. https://doi.org/10.1016/j.pedn.2020.10.003

Cristia, A., Bulgarelli, F., & Bergelson, E. (2020a). Accuracy of the Language Environment Analysis system segmentation and metrics: A systematic review. *Journal of Speech, Language, and Hearing Research, 63*(4), 1093–1105. https://doi.org/10.1044/2020_JSLHR-19-00017

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., Bunce, J., & Bergelson, E. (2020b). A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior Research Methods*, *53*(2), 467–486. https://doi.org/10.3758/s13428-020-01393-5

de Barbaro, K., Micheletti, M., Yao, X., Khante, P., Johnson, M., & Goodman, S. H. (2022). *Infant crying predicts real-time fluctuations in maternal mental health in ecologically valid home settings* [Manuscript submitted for publication].

Deng, L., & Yu, D. (2013). Deep learning: Methods and applications. *Foundations and Trends® in Signal Processing, 7*(3–4), 197–387. https://doi.org/10.1561/2000000039

Gilkerson, J., Coulter, K. K., & Richards, J. A. (2008). *Transcriptional analyses of the LENA natural language corpus*. https://www.lena.org/wp-content/uploads/2016/07/LTR-06-2_Transcription.pdf

Gilkerson, J., & Richards, J. A. (2020). *A guide to understanding the design and purpose of the LENA® system*. https://www.lena.org/wp-content/uploads/2020/07/LTR-12_How_LENA_Works.pdf

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H. L., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, *26*(2), 248–265. https://doi.org/10.1044/2016_AJSLP-15-0169

Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., & Gilkerson, J. (2011). Assessing children's home language environments using automatic speech recognition technology. *Communication Disorders Quarterly, 32*(2), 83–92. https://doi.org/10.1177/1525740110367826

Hubbard, F. O., & van Ijzendoorn, M. H. (1991). Maternal unresponsiveness and infant crying across the first 9 months: A naturalistic longitudinal study. *Infant Behavior and Development, 14*(3), 299–312. https://doi.org/10.1016/0163-6383(91)90024-M

James-Roberts, I. S., & Halil, T. (1991). Infant crying patterns in the first year: Normal community and clinical findings. *Journal of Child Psychology and Psychiatry, 32*(6), 951–968. https://doi.org/10.1111/j.1469-7610.1991.tb01922.x

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*. https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

McGrath, J. M., Records, K., & Rice, M. (2008). Maternal depression and infant temperament characteristics. *Infant Behavior and Development, 31*(1), 71–80. https://doi.org/10.1016/j.infbeh.2007.07.001

Milgrom, J., Westley, D. T., & McCloud, P. I. (1995). Do infants of depressed mothers cry more than other infants? *Journal of Paediatrics and Child Health, 31*(3), 218–221. https://doi.org/10.1111/j.1440-1754.1995.tb00789.x

Miller, A. R., Barr, R. G., & Eaton, W. O. (1993). Crying and motor behavior of six-week-old infants and postpartum maternal mood. *Pediatrics, 92*(4), 551–558. https://doi.org/10.1542/peds.92.4.551

Petzoldt, J., Wittchen, H.-U., Wittich, J., Einsle, F., Hofler, M., & Martini, J. (2014). Maternal anxiety disorders predict excessive infant crying: A prospective longitudinal study. *Archives of Disease in Childhood, 99*(9), 800–806. https://doi.org/10.1136/archdischild-2013-305562

Prochnow, A., Erlandsson, S., Hesse, V., & Wermke, K. (2019). Does a 'musical' mother tongue influence cry melodies? A comparative study of Swedish and German newborns. *Musicae Scientiae, 23*(2), 143–156. https://doi.org/10.1177/1029864917733035

Richters, J., & Pellegrini, D. (1989). Depressed mothers' judgments about their children: An examination of the depression-distortion hypothesis. *Child Development, 60*(5), 1068–1075. https://doi.org/10.2307/1130780

Salisbury, A., Minard, K., Hunsley, M., & Thoman, E. B. (2001). Audio recording of infant crying: Comparison with maternal cry logs. *International Journal of Behavioral Development, 25*(5), 458–465. https://doi.org/10.1080/016502501316934897

Stifter, C. A., & Spinrad, T. L. (2002). The effect of excessive crying on the development of emotion regulation. *Infancy, 3*(2), 133–152. https://doi.org/10.1207/S15327078IN0302_2

Wessel, M. A., Cobb, J. C., Jackson, E. B., Harris, G. S., & Detwiler, A. C. (1954). Paroxysmal fussing in infancy, sometimes called "colic". *Pediatrics, 14*(5), 421–435. https://doi.org/10.1542/peds.14.5.421

Wolke, D., Bilgin, A., & Samara, M. (2017). Systematic review and meta-analysis: Fussing and crying durations and prevalence of colic in infants. *The Journal of Pediatrics, 185*, 55–61. https://doi.org/10.1016/j.jpeds.2017.02.020

Xu, D., Yapanel, U., & Gray, S. (2009). *Reliability of the LENA. TM Language Environment Analysis system in young children's natural home environment*. https://www.lena.org/wp-content/uploads/2016/07/LTR-05-2_Reliability.pdf

Yao, X., Micheletti, M., Johnson, M., Thomaz, E., & de Barbaro, K. (2022). Infant crying detection in real-world environments. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 131–135). https://ieeexplore.ieee.org/document/9746096

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.