Analysis of acoustic and voice quality features for the classification of infant and mother vocalizations

Jialu Li\textsuperscript{a,b,*}, Mark Hasegawa-Johnson\textsuperscript{a,b}, Nancy L. McElwain\textsuperscript{a,c}

\textsuperscript{a}Beckman Institute, University of Illinois, Urbana, IL 61801, USA
\textsuperscript{b}Department of Electrical and Computer Engineering, USA
\textsuperscript{c}Department of Human Development and Family Studies, USA

Abstract

Classification of infant and parent vocalizations, particularly emotional vocalizations, is critical to understanding how infants learn to regulate emotions in social dyadic processes. This work is an experimental study of classifiers, features, and data augmentation strategies applied to the task of classifying infant and parent vocalization types. Our data were recorded both in the home and in the laboratory. Infant vocalizations were manually labeled as CRY (cry), FUS (fuss), LAU (laugh), BAB (babble) or SCR (screech), while parent (mostly mother) vocalizations were labeled as IDS (infant-directed speech), ADS (adult-directed speech), PLA (playful), RHY (rhythmic speech or singing), LAU (laugh) or WHI (whisper). Linear discriminant analysis (LDA) was selected as a baseline classifier, because it gave the highest accuracy in a previously published study covering part of this corpus. LDA was compared to two neural network architectures: a two-layer fully-connected network (FCN), and a convolutional neural network with self-attention (CNSA). Baseline features extracted using the OpenSMILE toolkit were augmented by extra voice quality, phonetic, and prosodic features, each targeting perceptual features of one or more of the vocalization types. Three web data augmentation and transfer learning methods were tested: pre-training of network weights for a related task (adult emotion classification), augmentation of under-represented classes using data uniformly sampled from other corpora, and augmentation of under-represented classes using data selected by a minimum cross-corpus information difference criterion. Feature selection using Fisher scores and experiments of using weighted and unweighted samplers were also tested. Two datasets were evaluated: a benchmark dataset (CRIED) and our own corpus. In terms of unweighted-average recall of CRIED dataset, the CNSA achieved the best UAR compared with previous studies. In terms of classification accuracy, weighted F1, and macro F1 of our own dataset, the neural networks both significantly outperformed LDA; the FCN slightly (but not significantly) outperformed the CNSA. Cross-examining features selected by different feature

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

\*Corresponding author at: Beckman Institute, University of Illinois, Urbana, IL 61801, USA. jialuli3@illinois.edu (J. Li).

CRediT authorship contribution statement

Jialu Li: Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Visualization. Mark Hasegawa-Johnson: Conceptualization, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition. Nancy L. McElwain: Data curation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
selection algorithms permits a type of post-hoc feature analysis, in which the most important acoustic features for each binary type discrimination are listed. Examples of each vocalization type of overlapped features were selected, and their spectrograms are presented, and discussed with respect to the type-discriminative acoustic features selected by various algorithms. MFCC, log Mel Frequency Band Energy, LSP frequency, and F1 are found to be the most important spectral envelope features; F0 is found to be the most important prosodic feature.

**Keywords**

Emotion classifier; Infant vocalizations; Infant-directed speech; Convolutional neural networks; Self-attention; Global feature; Feature selection

1. **Introduction**

Infant vocalizations, especially expressions of emotions, often serve communication purposes in dyadic social processes. Basic emotional infant outbursts, such as cry, fuss, laugh, babble, and screech, can convey meaningful information to parents and other caregivers; for example, a loud and harsh cry is perhaps due to hunger, while a burbling laugh may imply satisfaction. Infants’ emotional expressions and regulation of emotions change dynamically over time as infants grow and acquire the ability to interact with their caregivers in more complex ways. Parental interactions with the child (infant-directed speech, playful speech, shared laughter) help children to develop social and emotional abilities that will support positive mental health outcomes into childhood and beyond.

To permit more intensive, larger-scale studies of these dynamic emotion-relevant interactions between infants and parents, this paper describes a comparative study of different algorithms trained for the automatic classification of the vocalization types of infants and parents. A corpus of infant–parent interactions, including interactions recorded both at home and in a laboratory setting, was manually coded for vocalization type for both the infant and parent. Infant vocalizations were segmented in time, and one of the following six mutually exclusive labels was assigned to each vocalization: **CRY**, **FUS** (fuss), **LAU** (laugh), **BAB** (babble), or **SCR** (screech; labeling conventions provided in Section 3.3). Three classifier architectures were trained and tested using features extracted from the **OPENSMILE toolkit** (Eyben et al., 2010): linear discriminant analysis (LDA), a fully-connected neural network (FCN), and a convolutional neural network with self-attention (CNSA). The available data in our own corpus were augmented by external datasets in two ways. First, we selected additional training examples from datasets including Google AudioSet (Gemmeke et al., 2017) and FreeSound (Font et al., 2013), using both uniform sampling and weighted sampling (Gujral et al., 2019). Second, we experimented with pre-training of the CNSA using an external adult emotion recognition corpus (Livingstone and Russo, 2018). The baseline acoustic feature vector was expanded and contracted, in two sets of experiments. First, we designed complementary features covering aspects of voice quality and speech timbre not well represented in the baseline feature vector, including low-level descriptors of auditory roughness and formant frequencies, and additional measures of pitch and loudness. Second, we used Fisher scores to select the features most helpful for each vocalization type.
detection task; by thus reducing the dimension of the feature vector, we were able to reduce model overfitting.

Apart from improving classification results, we also investigated the most discriminative features among the augmented feature set, with multiple feature selection methods, including Fisher score, Extratree classifiers, minimum redundancy maximum relevancy, and Chi-square scores. For each feature selection algorithm, we investigated the relationship between the minimal number of features selected, from 10–100, vs. accuracy and macro F1 scores for multiclass classifications to test the robustness. Top 100 features were also tested for one-vs-one, and one-vs-all classifications. To understand the features used for distinguishing among different classes, we studied top 30 features selected by each algorithm overlapped between one-vs-one and one-vs-all classifications for each class. Then we further cross-examined those overlapped features among different algorithms. We also used t-SNE (Van der Maaten and Hinton, 2008) to visualize the hidden structure by selecting top five features from all pairs of one-vs-one classifications.

Caregiver vocalizations were also segmented in time, and were separately labeled using a mutually exclusive, collectively exhaustive list of vocalization types including ADS (adult-directed speech), IDS (infant-directed speech), PLA (playful noises), RHY (singing and rhythmic speech), WHI (whisper), and LAU (laugh; labeling conventions provided in Section 3.3). Several studies (Thiessen et al., 2005; Pegg et al., 1992) have shown that infants prefer to listen to IDS rather than ADS. Classifiers trained to distinguish adult speech emotions have been shown to be effective for distinguishing between IDS and ADS (Inoue et al., 2011; Mahdhaoui and Chetouani, 2011). In this study, we further tested the ability of an adult speech emotion classifier to distinguish between manually-coded positive and neutral emotions in IDS.

This paper proposes a convolutional neural network with self attention (CNSA) that accepts a fixed-dimensional global feature vector, extracted per vocalization, and reshapes it into an image, in order to take advantage of the parameter-sharing characteristics of the CNSA architecture. To the best of our knowledge, this is the first paper that uses global features as input to a CNSA for the classification of infant vocalizations. We evaluate our classifiers on both benchmark dataset (the Cry Recognition In Early Development (CRIED) (Marschik et al., 2017) database) and our own corpus. For benchmark evaluation, we demonstrate that our CNSA architecture, trained with a weighted sampler and a Fisher-score-selected 1000-dimensional global feature vector, achieves better unweighted average recall (UAR) than previous published studies of the CRIED corpus.

The paper is organized as follows: Section 2 reviews related work. Section 3 describes the CRIED dataset and the acquisition and annotation process of our infant–parent spoken corpus. Section 4 describes our process of preparing acoustic features, transfer learning techniques, architecture of multiple models, and various feature selection algorithms. Section 5 provides classification results. Section 6 analyzes the features that are confirmed by multiple feature selection algorithms for distinguishing among different classes, as a method of better understanding some of the salient acoustic differences among vocalization types. Section 7 concludes.
2. Related work

The categories of infant and adult vocalizations described in this paper are based on categories that have been studied in several different contexts, including both automatic and interactive acoustic analysis. Various artificial neural networks (ANNs) have been shown to be effective in automatic vocalization classification tasks. A typical workflow for these tasks is to first collect infant and caregiver audio data at home using the Language Environment Analysis device (Dongxin et al., 2014, LENA), or videotape their interactions in a lab while they complete interactive tasks, then extract relevant acoustic and prosodic features and train the ANN classifiers. Analysis of both infant and adult vocalizations has also been performed interactively to discover acoustic features that can help to detect particular types of infant vocalization (especially CRY), and to better characterize the difference between infant-directed and adult-directed speech (IDS and ADS).

Section 2.1 reviews a large number of recent studies, including 6 CNN-based models, that have developed analysis or automatic classification of infant vocalizations into vocalization types similar to those used in our experiments. Section 2.2 reviews a small representative sample of studies that have attempted to measure the salient acoustic features of infant-directed speech.

2.1. Analysis and classification of infant vocalizations

Studies of infant vocalizations have included those that report on the range of variation of measured acoustic features, and those that seek to automatically classify vocalization types, with many studies attempting both tasks.

Kent and Murray (1982) studied utterance durations, change of fundamental frequency (F0), change of formant frequencies, and variations of source excitations in the vocal tract for infant vocalizations, and discovered that infant vocalizations often have abrupt F0 shift, vocal fry and noise segments. Papaeliou et al. (2002) studied acoustic patterns of infant vocalizations expressing emotional and communicative functions, and their results show that the features best distinguishing between these two classes are functional descriptors of the time-domain trajectory of F0, including its peak and final values, standard deviation, skewness, kurtosis, duration, and the ratio of standard deviation of F0 to the duration of the vocalization. Reggiannini et al. (2013) studied automatic detection of infant cry, and discovered the necessity of a three-fold categorization: (1) modal cry (voiced, with F0 between 200 Hz and 1000 Hz), (2) voiced fricative cry, and (3) high-voice cry (F0 between 1000 Hz and 5000 Hz). Our screech (SCR) category is inspired by the high-voiced cry category of Reggiannini et al. (2013).

Ji et al. (2021b) presented a comprehensive survey of automatic classification of infant vocalizations. We summarized 12 studies of automatic classification of infant vocalization types, among others, in Table 1. Details of these studies are described in the paragraphs that follow.
2.1.1. Multiple infant vocalizations classifiers—Petroni et al. (1995) recorded 230 crying episodes from 16 infants who were 2–6 months old in the Montreal Children’s Hospital. The target classes, $y_{\text{Petroni}}$, were

$$y_{\text{Petroni}} \in \{\text{ANGER, FEAR, PAIN}\}.$$

which were labeled according to the child’s activity during the recording: pain when children received routine immunization, fear when they saw a jack-in-the-box, and anger when the child’s head was restrained, respectively. Petroni et al. experimented with 5 different ANNs, including feed-forward neural network with full (FF) and tessellated (FT) connections, recurrent neural network (RNN), time-delay neural network and cascade correlation (CC) network. This experiment compared all the ANNs’ performances with input features of mel-cepstrum coefficients and mel filter-bank energy coefficients; FF fed with mel-cepstrum coefficients achieved the best accuracy of 79.4%.

Warlaumont et al. (2010) video and audio-taped two or three 20mins sessions of 6 infants and their caregivers’ joint play when infants were 3, 6, and 9 months old. The target classes are VOCANT (normal-pitch sounds), SQUEAL (high-pitched sounds), and GROWL (low-pitched sounds). Warlaumont et al. fed one second long spectrograms into hybrid self-organizing maps (SOM), that were then fine-tuned as a supervised single-layer perceptron. The model achieved 46.0% accuracy overall.

Yamamoto et al. (2013) collected 57 long recordings of one 1.5-month-old infant, containing 1172 infant vocalizations. These were labeled as CRY or NON-CRY. Examples of CRY were then presented to the infant’s caregiver, who labeled each recorded cry according to its cause: DIS (discomfort), HUN (hunger), or SLE (sleepy). Automatic labeling of CRY vs. NON-CRY was performed by attempting to transcribe the utterance using a standard adult automatic speech recognizer (ASR): segments that could not be orthographically transcribed by the ASR, and segments with an F0 discontinuity of at least 150 Hz in 0.1 s, were labeled as cry. The sub-categories DIS, HUN and SLE were labeled using K-nearest-neighbors (KNN) applied to the principal components analysis (PCA) of a 32-sample magnitude FFT, with majority voting between consecutive frames in the same recording.

2.1.2. SVM-based infant vocalizations classifiers—Schuller et al. (2019) computed baseline accuracies for the 2019 Interspeech Paralinguistic challenges; the same dataset was used in the experiments reported in Goxztolya (2019) and Yeh et al. (2019). The BabySounds Sub-Challenge required proposed models to automatically classify 12445 vocalizations collected from 2–36-month-old infants without any speech development delays using LENA. The target classes, $y_{\text{BabySounds}}$, include

$$y_{\text{BabySounds}} \in \{\text{CAN, NON-CAN, CRY, LAU, OTH}\}.$$

where the symbols CAN and NON-CAN denote canonical babbling (includes consonants) and non-canonical babbling, respectively, and OTH denotes “other.” Schuller et al. (2019) implemented support vector machine (SVM) with linear kernel. The feature set includes
the ComParE acoustic feature set, Bag-of-Audio-Words (BoAW), and audio-encoder based deep representations of the input audio using the AuDEEP toolkit. The baseline accuracy for the BabySounds Sub-Challenge is 58.7%. Goxztolya (2019) augmented the baseline of Schuller et al. (2019) using Fisher vectors computed from the frame-level MFCC, resulting in a small improvement over the baseline classifier, to 59.5%. Yeh et al. (2019) used the same features and classifier as in the baseline of Schuller et al. (2019), but augmented the training dataset using natural baby cry sounds, and using synthetic examples of all classes generated by an adversarial autoencoder network, resulting in a relatively large improvement over the baseline, to 62.39%.

2.1.3. CNN-based infant vocalizations classifications—Anders et al. (2020) collected 833 infant vocalizations, more than 4 s each, from infants less than 9 months old, from multiple online sound libraries. They labeled each utterance with target class labels \( \gamma_{\text{Anders}} \) including:

\[
\gamma_{\text{Anders}} \in \{\text{CRY, FUS, BAB, LAU, OTH}\}
\]

Anders et al. (2020) computed mel-scaled spectrograms given the input audio, then classified the spectrograms using a CNN-based architecture, achieving 72.0% accuracy in this five-class classification task.

Ebrahimpour et al. (2020) used a day-long home LENA audio dataset by Pretzer et al. (2019). Twelve-hour recordings of 19 infants at ages of 3, 6, 9, and 18 months were used for study. Utterances were classified as \texttt{INFANT} vs. \texttt{ADULT}. Infant utterances were classified as \texttt{VOC} (vocal) vs. \texttt{NON-VOC} (non-vocal), and babble was further classified as \texttt{CAN} vs. \texttt{NON-CAN}. Adult utterances were classified as \texttt{IDS} vs. \texttt{ADS}. Ebrahimpour et al. (2020) explored CNN-based networks, some using the inception architecture (Szegedy et al., 2015), observing the raw audio waveform. CNN with the inception module performed better in \texttt{VOC} vs. \texttt{NON-VOC} and achieved accuracy of 90.49%. CNN without the inception module achieved better accuracies in \texttt{INFANT} vs. \texttt{ADULT} (97.01%), \texttt{CAN} vs. \texttt{NON-CAN} (77.03%), and \texttt{IDS} vs. \texttt{ADS} (60.28%).

Turan and Erzin (2018) used the Cry Recognition In Early Development (CRIED) dataset developed by Marschik et al. (2017) in Interspeech 2018 paralinguistic challenge held by Schuller et al. (2018). CRIED dataset consists of 5587 vocalizations collected from 20 healthy infants. The target classes are

\[
\gamma_{\text{Turan}} \in \{\text{NEUTRAL, FUS, CRY}\}
\]

In this study, Turan and Erzin (2018) preprocessed the audio using high-pass FIR filter with a basic voice activity detection (VAD) algorithm followed by extracting spectrograms. Then spectrograms were passed to three layers of CNN with ReLU activations followed by three “PrimaryCaps”. Each “PrimaryCaps” was a convolutional capsule layer with 32 channels of convolutional 8D capsules (Hinton et al., 2011). The last “PrimaryCaps” then passed its output to a parallel bank of three “CryCaps” (one 16D capsule per class), which fed their computations to softmax classification. The best accuracy and unweighted average recall
(UAR) scores achieved after VAD and spectrogram equalization were 86.1% and 71.6% respectively.

Ji et al. (2021a) used two datasets: Baby Chillanto dataset (Reyes-Galaviz et al., 2008) and Baby2020 dataset. Baby Chillanto dataset, collected by the National Institute of Astrophysics and Optical Electronics, CONACYT, Mexico, contains 2267 crying sounds from infants under 9 months old; Baby2020 dataset, collected by the authors themselves, consists of 5540 crying samples from more than 100 healthy infants under 3 months old. The target classes are

\[
\gamma_{\text{BabyChillanto}} = \{\text{ASPHYXIA, DEAF, HUNGER, NORMAL, PAIN}\}
\]

\[
\gamma_{\text{Baby2020}} = \{\text{HUNGER, SLEEPY, WAKEUP}\}
\]

Ji et al. (2021a) proposed a CNN-based graph convolutional network (GCN). ResNet50 (He et al., 2016) was first used to extract features from spectrograms, and the extracted features were used to construct GCN based on similarities among features. GCN then learned the feature representation for each feature of each node over several layers before being used for classification. Ji et al. (2021a) constructed GCN for infant classifications in both semi-supervised and supervised ways. Given 80% training data and 20% testing data, semi-supervised GCN greatly improved CNN by 7.36% and 3.59% for BabyChillanto and Baby2020 dataset respectively.

Maghfira et al. (2020) used the Dunstan Baby Language that was discovered by Priscilla Dunstan and her research team in 2006 Dunstan (2012). Dunstan discovered five words used by all infants, regardless of the language, culture, and race, in their first three months life. These five words were used as target classes,

\[
\gamma_{\text{Dunstan}} = \{\text{NEH, OWH, EH, EAIRH, HEH}\}
\]

NEH means hungry, OWH means sleepy, EH means infant wants to burp, EAIRH means infant experiences stomach cramp, and HEH means infant feels discomfortable. In this experiment, 315 cry sound audios extracted from Dunstan Baby Language video were used. Spectrograms were fed into three layers of CNN followed by max-pooling, then CNN outputs were flattened and passed to two stacked RNNs for classification. The best accuracy was 86.03% achieved by using the categorical cross-entropy loss.

Ji et al. (2020) performed infant vocalizations classifications on both Dunstan Baby Language and Baby Chillanto dataset. In this study, Ji et al. (2020) proposed a multi-stage CNN model using hybrid feature set to reduce the number of classes to classify at each stage based on the classification accuracy of each class. At first stage, the spectrograms CNN is used to classify five categories, and three classes with the lowest classification accuracy will be further examined. At second stage, the waveform CNN is used to classify those three classes, and two classes with the lowest classification accuracy will be further examined.
At final stage, prosodic (F0, formant frequencies F1–F5, and signal energy) CNNs are used to classify those two classes. The results show that hybrid-feature multi-stage CNN on Dunstan Baby Language dataset achieved accuracy of 88.22%, which outperformed 3 CNNs late fusion model by 4.14%. Multi-stage CNN model achieved 95.10% on Baby Chillanto dataset, which outperformed models proposed by previous studies.

2.2. Analysis and classification of adult vocalizations

Adults recognize the emotional content of speech with accuracy of 60%–100% if the speaker and listener are adult members of the same culture (Johnson et al., 1986; Scherer, 1996). If the speaker and listener share no cultural context, the emotional dimension of activity is nevertheless well-recognized, but evaluative meaning may be lost (Van Bezooijen et al., 1983). Measurable acoustic correlates of emotion include patterns of F0, duration and energy, spectral tilt, formant frequencies, jitter and shimmer (Williams and Stevens, 1972; Scherer, 1996). With such a complex range of acoustic correlates, the study of emotion has benefitted greatly from the release of large standard corpora (Dallaert et al., 1996), and by accompanying international algorithm competitions (Schuller et al., 2009). For example, the RAVDESS corpus (Ryerson Audio–Visual Database of Emotional Speech and Song, Livingstone and Russo, 2018) contains 1440 speech audio files, uniformly representing eight emotions: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. Twenty-four actors (12 male and 12 female) vocalized two lexically-matched statements twice for each of these eight emotions, with two levels of intensity for each emotion except neutral.

Survey papers on speech emotion recognition (SER) (El Ayadi et al., 2011) indicate a consensus among researchers that global features, extracted from the entire duration of an utterance, achieve higher-accuracy emotion recognition than frame-based features. Global features have the disadvantage of failing to represent some fine-grained temporal information, but they have the advantage of more effectively summarizing acoustic and prosodic features of the emotion throughout each utterance, e.g., because not every frame expresses emotion. Global features also typically lead to algorithms that are less computationally complex than local features, and that may have fewer trainable parameters, and hence be less susceptible to overfitting. Tzinis and Potamianos (2017) found a mechanism to integrate the best features of global and local features: global features (functional statistics of the low-level descriptors) were computed over windows of 3 s duration, overlapping by half, then processed by an LSTM in order to compute a single emotion score for the duration of an 8s utterance. Applying this algorithm to the IEMOCAP corpus (Busso et al., 2008) resulted in a UAR below the state of the art, but a weighted accuracy above the state of the art (64.16%). Zhang et al. (2018) tried a similar strategy, but using a CNN instead of an LSTM: a 375-dimensional global feature vector was re-shaped into a 15 × 25 image, concatenated to a spectrogram, and processed by a CNN, because a CNN uses fewer trainable parameters than a fully-connected network. The result achieved 91.42% UAR and 91.78% accuracy on the EmoDB (Burkhardt et al., 2005): results that are considerably better than those achieved using local features, but not equal to the global-feature state of the art for that corpus. Aldeneh and Mower-Provost (2017) proposed a method that reduces the overfitting tendencies of local-feature recognizers by the use of a convolutional neural network with self-attention (CNSA). Their CNSA architecture,
observing only spectrogram inputs, achieved a UAR equal to that of the state-of-the-art global-feature baseline (61.8%) on the IEMOCAP corpus.

Adults talking to very young children tend to speak with higher pitch, and larger pitch range, than adults talking to other adults (Garnica, 1977). The exaggerated prosody of infant-directed speech (IDS) (Fernald, 1985) serves a number of functions beneficial to healthy development, including maintaining the infant’s attention, and communicating affect (Sachs, 1977). Van de Weijer (2001) showed that the vowel triangle spaces, defined by the F1 and F2 values of vowels /a/, /i/ and /u/, are expanded in infant-directed speech (IDS) compared with adult-directed speech (ADS) in function words; the opposite pattern is observed for content words. Kalashnikova and Burnham (2018) studied three hypothesized correlates of IDS (pitch, affect and vowel hyperarticulation), measuring their correlation with the infant’s age (7–9 months old) and vocabulary size, and found that all three correlates are exaggerated in IDS at all ages compared with ADS. Kalashnikova and Burnham (2018) also presented a theoretical argument to the effect that vowel hyperarticulation is the only significant component affecting infant vocabulary development, while pitch and affect have no such benefit. Automatic classification of IDS vs. ADS has been shown, in several studies, to be far more difficult than the automatic classification of infant vocalization types: the results of Ebrahimpour et al. (2020), 60.28% accuracy, shown in Table 1, are typical of the state of the art.

3. Data

3.1. Benchmark data: CRIED database

For benchmark evaluation, we tested all of our models on the CRIED database developed by Marschik et al. (2017) consisting of 5587 vocalizations of 20 infants (10 males and 10 females). All infant vocalizations were recorded and video-taped in a bi-weekly interval from the 4th until the 16th week of life. All vocalizations were extracted from sequences in which the infant was lying in a cot in supine position without any interactions of the environment. Two experts in the field of early speech-language development annotated those vocalizations based on the audio–video clip as one of three classes: “neutral/positive” (NEU), “fussing” (FUS) and “crying” (CRY). We followed the Leave-One-Subject-Out (LOSO) training/testing dataset partitions based on the Crying Sub-Challenge of the Interspeech 2018 paralinguistic challenge (Schuller et al., 2018). The numbers of samples for each class in train/test split are (1) neutral/positive (2292 samples in LOSO-training, 2172 samples in LOSO-testing), (2) fussing (368 samples in LOSO-training, 441 samples in LOSO-testing), and (3) crying (178 samples in LOSO-training, 136 samples in LOSO-testing).

3.2. Experimental data

Data used in this paper were collected as part of two separate studies of typical social and emotional development during the first years of life. Families for both studies were recruited from the community with study flyers distributed at child care centers, pediatric clinics, and local community organizations or online forums visited by families of infants (e.g., public libraries, parenting groups). First, the Toddler Development Project (TDP) was a pilot study of 15 families with children between the ages of 13 and 24 months (mean = 17.67 months,
standard deviation = 3.5 months), which included LENA recordings in the home at a single time point. Second, the Infant Development Project (IDP) was a study of 119 families, which followed infants across the first year of life at 3, 6, 9, and 12 months. Both home recordings and laboratory visits were conducted at each time point, and for the purposes of the current study, we report on subsamples of participants from each study (see Table 2). All study procedures were approved by the Institutional Review Board at the University of Illinois at Urbana-Champaign (TDP protocol #16764; IDP protocol #17748).

Of the 171 h of LENA data included in the current study, 95 h of data (from 6 participating families) were studied by Xu et al. (2018). For the current study, an additional 76 h of audio were acquired and transcribed, from an additional 30 families (4 from TDP; 26 from IDP). The TDP LENA recordings used were based on full-length recordings (about 16 h across multiple days), pre-processed using the LENA segmentation software. Segments labeled by LENA as containing child or adult vocalizations were verified by human coders, and retained for further annotation if the LENA annotation was correct. The IDP LENA recordings were based on ten-minute segments, extracted automatically from day-long recordings by selecting the segments with the highest frequency of LENA-annotated infant or adult vocalizations. After the ten-minute segments were automatically selected, their original LENA annotations were erased, and each segment was re-coded manually with start and end times of all child and adult vocalizations.

IDP LAB audio recordings were also acquired in the laboratory during two semi-structured interactive tasks. At 3 months, mothers and infants were observed in an 8-min “play” task, in which mothers were asked to play with their infants as they might if they had a few moments at home alone with their baby. Age-appropriate toys were provided, although mothers could decide whether or not to use them. At 3, 6, and 9 months, mothers and infants were also observed during a 6-min “still face” procedure (Tronick et al., 1978), in which the infant was seated in a bouncy seat or high chair, and the mother was seated directly across from the infant with no toys present. The “still face” involved three 2-min episodes: (a) natural play, in which the mother and infant engaged in face-to-face interaction, (b) still face, in which mother was signaled by a knock to cease all verbal and physical interaction and look at the infant with a neutral expression, and (c) reunion, in which the mother was signaled by a second knock to resume interaction with the infant. All lab sessions were video taped for later coding, and independent teams of trained coders annotated infant and maternal behaviors, including vocalizations and facial expressions. Table 2 shows the number of LENA and laboratory recordings, grouped according to infants’ ages, used in the analyses.

### 3.3. Annotations

For the TDP data, each LENA audio was automatically segmented by the LENA system, which labeled instances of focal infant vocalizations, and of vocalizations by male and female adults and other children. Manual annotation then affixed a vocalization type label to each verified child or adult vocalization. Annotators were also asked to adjust the length of each segment as appropriate, in order to make sure that the start and end times of each segment are correct to within 100 ms, and to delete segments that were LENA false positives (segments that LENA labeled as containing a vocalization by the specified person, but which
actually contained only background noise). No attempt was made to find and label LENA false-negative segments.

In contrast, 10-min segments from the IDP LENA data were re-segmented manually, without reference to the original LENA automatic segmentations.

For the IDP LAB sessions, annotators inserted code at the start time of each vocalization type. The labeled data file results in intervals of 33 ms. Besides labeling vocalizations as described below, annotators of the IDP lab sessions also labeled mother’s facial expressions in each video frame. The mother’s facial expressions $y_{\text{adultface}}$ were labeled, if visible, as one of the seven categories:

$$y_{\text{adultface}} \in \{\text{FLAT, INTERESTED, SIMPLESMILE, BROADSMILE, ANGRY, DISTRESSED, OTHER}\}$$

Descriptions of mother facial expression codes used in this study are shown in Table 3.

Annotators labeled infant intentional vocalizations in all three sub-corpora (TDP LENA, IDP LENA, and IDP LAB), into one of five distinct vocalization types, $y_{\text{infantvoc}}$:

$$y_{\text{infantvoc}} \in \{\text{CRY, FUS, LAU, BAB, SCR}\}$$

Unintentional vocalizations (vocalizations judged to not be part of an intentional communicative act, e.g., hiccup, burp, cough, sneeze) were removed from the infant vocalization tier, and hence treated identically to any other type of silence. The number of vocalizations in the train and test sets for each class, in each fold of cross validation, are shown in Table 4. Descriptions of the infant vocalization codes are shown in Table 5.

Annotators were asked to label adult vocalizations in the IDP LENA and IDP LAB (but not TDP LENA) datasets, if not silent, into one of the six categories $y_{\text{adultvoc}}$:

$$y_{\text{adultvoc}} \in \{\text{IDS, ADS, PLA, RHY, LAU, WHI}\}$$

The number of adult vocalizations in the train and test sets for each class, in each fold of cross validation, are shown in Table 6.

Descriptions of the adult vocalization codes are shown in Table 7. Twenty-two percent of the IDP lab session recordings were double-coded to access intercoder reliability; coders were blind to which protocols were double-coded. Reliability was assessed via Cohen’s kappa and showed good reliability for all codes: 0.82 (infant vocalizations), 0.83 (mother vocalizations), and 0.74 (mother facial expressions).

The IDP lab data contains only recordings of infants and their adult female caregivers, whereas the LENA datasets (both IDP and TDP) also include recordings of adult male caregivers and older siblings. Adult male vocalizations and older sibling vocalizations were labeled using the labels $y_{\text{adultvoc}}$, but are not used to train or test classifiers, because of their scarcity in the dataset.
4. **Methods**

4.1. **Acoustic features**

For LENA vocalizations, we define each labeled audio segment as one data sample. For IDP lab vocalizations, because of its high-resolution labeling scheme, we define any sequence of continuous intervals labeled as the same class to be one data sample.

We extracted 1582 spectral and prosodic features for each vocalization using the OpenSMILE toolkit (Eyben et al., 2010), based on the configuration file named emobase2010.conf. This particular set of acoustic features was published as the winning strategy for the Interspeech 2010 Paralinguistic Challenge (Schuller et al., 2010). The acoustic feature set is composed of 21 functionals (segment-level summaries) created from each of 76 low-level descriptors (LLDs, the acoustic features and its first derivatives computed in each frame). The LLDs include loudness, 15 Mel-frequency cepstral coefficients (mfcc[0] through mfcc[14]), 8 log power measurements acquired in Mel-frequency bands (logMelFreqBand[0] through logMelFreqBand[7]), 8 line spectral pair (LSP) frequencies computed from 8 linear prediction filter coefficients (lspFreq[0] through lspFreq[7]), the smoothed F0 contour (F0final) and its envelope (F0finEnv), the probability of voicing of the most likely F0 candidate (voicingFinal), the frame-to-frame jitter and shimmer, and the differenced frame-to-frame jitter (jitterDDP). LLDs include both static and delta ((d)) measurements, for example, mfcc(d)[0] means the first derivative of mfcc[0]. From each of these 76 LLDs, 21 different functionals were created, as specified in Table 8, for a total of 1582 features. We use this acoustic feature set as our default feature set.

In order to further reduce classification error rates, we designed a set of complementary LLDs, each representing perceptually salient voice quality, prosodic, or phonetic attributes of one or more of the vocalization types. Zero-crossing rate (ZCR), root-mean-squared energy (RMS), unsmoothed fundamental frequency (F0 voiceProb) and voicing probability (voiceProb) and their deltas were added using the OpenSMILE configuration file IS09_emotion.conf, because those features have been proven to be helpful for laughter detection (Knox and Mirghafori, 2007). Approximate auditory roughness (approx roughness) as described in He et al. (2017) was added to the LLD set based on the hypothesis that auditory roughness, defined as the human perception of harsh sounds (Vassilakis, 2005), would distinguish between CRY and FUS. Furthermore, we used Parselmouth (Boersma and Weenink, 2018), a Python version of Praat (Boersma and Weenink, 2020), to extract more readily interpretable phonetic, voice quality, and prosodic features including the first three formant frequencies (F1 through F3), the level difference (in decibels) between the first harmonic and the first formant (H1-A1; Hanson, 1997), the level difference between the first harmonic and the second harmonic (H1—H2), the harmonic-to-noise ratio (HNR; Yumoto and Gould, 1982; Boersma, 1993), and additional measures of signal energy, pitch, and intensity. The six IS09_emotion.conf features were each summarized by 11 functionals, and the ten complementary features each 20 functionals as shown in Table 8, for a total complementary feature set of 266 features.

In addition to the default and default + complementary feature sets, the CNSA model (training algorithm described in Section 4.4.3) was also tested using a raw (no functional

*Speech Commun. Author manuscript; available in PMC 2022 September 01.*
features) Mel spectrogram input (40 coefficients, 25 ms Hamming window, 10 ms shift). Other classifiers were not tested with raw Mel spectrogram inputs.

All features were Z-normalized over the complete training dataset prior to training any classifiers.

### 4.2. Transfer learning and data augmentation

Like most naturalistic recordings, the dataset used in this study contains a highly imbalanced distribution of vocalization types. The number of samples for each class are as follows, BAB: 25447, FUS: 6946, CRY: 1766, LAU: 1316, SCR: 203; IDS: 6093, PLA: 3401, LAU: 1628, RHY: 795, WHI: 762, and ADS: 308. In order to reduce the deleterious effects of class imbalance, three types of transfer learning/data augmentation were attempted: (1) pre-training of classifier weights on a related task (adult emotion recognition), (2) augmentation of under-represented vocalization types using data uniformly sampled from public corpora, (3) augmentation of all vocalization types using data sampled according to a cross-corpus information similarity measure.

#### 4.2.1. Adaptation from pretrained model—
We pretrained and adapted the proposed CNSA model as described in Section 4.4.3 using RAVDESS (Livingstone and Russo, 2018). We randomly choose one expression of each emotion produced by each actor as a validation test set, and use the rest as training data. The CNSA model was pre-trained using both the default feature set and the raw Mel spectrogram input. Validation accuracies on the RAVDESS validation set were 71.88% for the default feature set, and 77.04% for the raw Mel spectrogram, therefore only the CNSA model with raw Mel spectrogram features was adapted to the LENA and IDP datasets.

#### 4.2.2. Data augmentation: Uniform—
The IDP lab corpus is entirely composed of one-on-one mother-infant interactions recorded in a laboratory setting, therefore, by design, the plurality of adult utterances (46.9%) are of type IDS. In an attempt to improve the F1 score of classifiers of adult vocalization type, therefore, examples of other vocalization types were sampled from other corpora. Adult-directed speech (ADS) was sampled (300 vocalizations, selected uniformly at random) from the Augmented Multi-party Interaction (AMI) corpus (Mccowan et al., 2005, meeting ID: IS1009b). Whispered speech vocalizations (WHI) (500 vocalizations, selected uniformly at random) were sampled from the American-accented portion of the wTIMIT corpus (Lim, 2011).

#### 4.2.3. Data augmentation: Non-uniform—
Infant vocalization types in the LENA + IDP lab dataset are even more imbalanced than adult vocalizations: 71.3% of samples are BAB, 19.5% are FUS, and the remaining three classes (CRY, LAU, SCR) are collectively 9.2%. In an attempt to augment infant vocalizations, therefore, we combined audio clips from categories of baby cry, baby laughter and baby babbling in the Google AudioSet (Gemmeke et al., 2017) as well as baby cry, baby laughter, baby fussing, and baby babbling found in the Freesound database (Font et al., 2013). Uniform sampling of data from AudioSet and Freesound caused reduced precision and recall for every vocalization type: apparently, the acoustic characteristics of CRY, FUS, LAU, and BAB available in public
datasets are quite different from the acoustic characteristics of the same vocalization types in our dataset. In order to make more effective use of cross-corpus data augmentation, therefore, we selected samples from external corpora using the in-domain/out-domain pointwise divergence measure proposed by Moore and Lewis (2010). Given a speech sample \( s \), let \( P_I(c|s) \) be the probability that utterance \( s \) is classified as vocalization type \( c \), as computed by an in-domain classifier, \( I \), trained on our spoken corpus. Similarly, \( P_O(c|s) \) is the probability computed by an out-of-domain classifier, \( O \), trained using labeled examples from the external corpus (Google AudioSet or Freesound). The Kullback–Leibler divergence between the measures \( P_I \) and \( P_O \) is

\[
D_{KL}(P_I \parallel P_O) = \sum_{s, c} P_I(s, c) \log \frac{P_I(c \mid s)}{P_O(c \mid s)}
\]  

(1)

We are interested in samples that minimally change \( D_{KL}(P_I \parallel P_O) \), i.e., samples for which \( \log \frac{P_I(c \mid s)}{P_O(c \mid s)} \approx 0 \). Indeed, we are only really interested in the degree to which the two classifiers agree about the correct classification. If \( \hat{c} = \text{argmax } P_I(c|s) \) is the classifiers output of the in-domain classifier, then we are interested the augmentation dataset \( \mathcal{D}_A \) defined by:

\[
\mathcal{D}_A = \left\{ s \in O : \log \frac{P_I(\hat{c} \mid s)}{P_O(\hat{c} \mid s)} < \text{threshold} \right\}
\]  

(2)

A small number of experiments using validation test data suggest that adding dataset \( \mathcal{D}_A \) to the training data provides only a small benefit to the F1 score of the resulting classifier, and that the benefit is maximized with a threshold of approximately 0.1. Experiments were conducted using augmentation of all four classes available from external sources (CRY, FUS, LAU and BAB), and using augmentation of the three non-majority classes (CRY, FUS and LAU); augmentation using all four classes resulted in higher F1 score, therefore only results from that experiment are presented here. Table 9 shows the exact number of external infant vocalizations selected for each class using three-fold cross-validation tests of different models. As shown, although BAB is the majority class in the original training set, the goal of maximum F1 score was achieved by selecting a larger number of tokens from this class than from any other class, apparently because other classes suffered from larger differences between the audio feature distributions of in-domain vs. out-domain examples.

#### 4.3. Auxiliary task: Infant-directed speech emotion classification

In order to perform downstream research on the long-term development of emotional and social interaction between an infant and an adult caregiver, it would be useful to automatically label the emotional valence of infant-directed utterances. In our training and test datasets, there were no examples of IDS that coders perceived to have any emotional tone other than “happy” or “neutral”, therefore, in order to explore the possibility of automatic emotion classification, a binary (neutral vs. happy) emotion classifier was fine-tuned over vocalizations labeled as ids. Manual annotations did not specify this emotion label in the audio, but emotional annotations were provided in the accompanying video: we defined vocalizations as neutral when the simultaneous mother’s facial expression was...
labeled as $y_{\text{adultface}} = \text{FLAT or INTERESTED}$, and we defined IDS vocalizations as happy when the simultaneous facial expression was labeled as $y_{\text{adultface}} = \text{SIMPLE SMILE or BROAD SMILE}$. Descriptions of facial expression codes are shown in Table 3. Cross-validation experiments determined that higher accuracy was obtained by augmenting the training dataset, for this task, using examples from vocalization types other than IDS, as follows: vocalizations of type ADS were labeled as NEUTRAL, while those of types RHY, PLA, or LAU were labeled as HAPPY. The resulting number of vocalizations in training and test data, for each fold of cross-validation, are shown in Table 10.

4.4. Models

Three different classifier architectures were tested: linear discriminant analysis (LDA), a fully-connected neural network (FCN), and a CNSA architecture applied to a global feature vector. Classifiers were tested using the default features, the default + complementary features, and the raw Mel spectrogram; not all classifiers were tested for all feature sets, as described below. Classifiers were tested for the tasks of classifying child vocalization type and adult vocalization type, using both in-domain and augmented training datasets; not all classifiers were tested for all datasets, as described below.

4.4.1. Linear discriminative analysis (LDA)—Linear discriminant analysis (LDA) was trained and tested with default and default + complementary feature sets. Procedures for training LDA were similar to those reported in Xu et al. (2018), using both the complete feature vector, and a feature vector compacted using the feature selection step reported in that paper. LDA using the full feature vector gave higher cross-validation accuracy in all experiments, and is therefore reported here.

4.4.2. Feed-forward fully-connected neural network (FCN)—The FCN is a two-layer feed-forward neural network model, with its parameter count limited in order to avoid over-training. After comparing several configurations in cross-validation, we set the number of hidden nodes to 128, for both the default and default + complementary feature sets. We choose leaky rectified linear unit (LReLU) as the activation function for both layers. In the output layer, LReLU outputs were fed as input to a softmax nonlinearity, and weights were trained to minimize cross-entropy.

4.4.3. Convolutional neural network with self-attention (CNSA)—The third architecture tested in this paper is a CNSA model, based on the models built by Aldeneh and Mower-Provost (2017) and An et al. (2019), but accepting global feature inputs in a manner similar to that proposed by Zhang et al. (2018). Fig. 1 shows the overall architecture of our proposed model.

The 2d CNSA was trained and tested using default, default + complementary, and raw Mel spectrogram features. In order to apply a 2d CNSA to the default and default + complementary feature vectors, the vector is first zero-padded so that its dimension is an integer multiple of 40, then resized into a matrix with 40 rows (40 × 25 for selected top 1000 features, 40 × 40 for the default features, 40 × 47 for default + complementary, 40 × 160 for COMPARE). We further zero-padded those vectors to match the size of the largest kernel (40
Although it is possible to reduce the size of the kernels for smaller feature vectors, we keep the sizes of kernels consistent across different feature sets. The resulting matrix is not an image, but studies such as Zhang et al. (2018) have demonstrated that the patterns in a very long feature vector may be better learned by resizing it to a matrix, in this way, and then applying a 2d CNN. The CNN uses fewer trainable parameters per hidden node than the FCN, permitting us to train a deeper network with the same amount of training data. Many of the input features are partially redundant. This redundancy is distributed unevenly through the 40 × 40 grid, in patterns that are difficult to predict in advance, but which may be learnable using convolution kernels followed by self-attention.

We used 4 kernels with different widths as listed in Table 12 to extract different information from the same region in the given input feature vectors. We applied max pooling over the outputs of each convolution layer, then we concatenated the outputs of the max pooling layer as a vector. We fed the concatenated hidden vector into a self-attention module followed by a dense layer. In the self-attention module as shown in Fig. 1b, $W_1$ and $W_2$ are two trainable matrices; $H$ is the output from the max pooling layer. $E$ is the output of the entire module, which is fed into the dense layer. Specifically, let

$$Z = \text{softmax}(\tanh(HW_1)W_2),$$

where $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ and $\text{softmax}(Z)$ is normalized along the first dimension of the input matrix, i.e., for $z_{ij}$ equal to the $i$th element the $j$th column of $Z$, $z_{ij} = \exp(z_{ij})/\sum_t \exp(z_{jt})$. Then we compute $E$ as

$$E = Z^TH.$$

Finally, the softmax probabilities of each vocalization type are computed as

$$y = \text{softmax(Dense(E[:]))},$$

where $E[:]$ denotes reshaping to a vector, and Dense() is an FCN with one hidden layer.

To test the effectiveness of introducing a self-attention module, we performed infant and mother vocalizations classifications over the proposed CNN network with and without self-attention. The results shown in Table 11 illustrate that the self-attention module greatly improved macro F1 score by 6.31% and 4.76% for infant and mother vocalizations classifications respectively, while the standard deviations across three-fold validation tests dropped under 1% for both classification tasks. The accuracy and weighted F1 scores are also slightly improved. In the rest of paper, we use “CNSA” to denote our proposed “CNN + self attention” network for brevity.

### 4.5. Feature selection analysis

To better understand how acoustic features are used to classify infant vocalizations, we performed feature selection over the default + complementary feature set for the IDP + TDP
dataset. We choose four feature selection methods from the review given in Li et al. (2017). A brief descriptions of each feature selection method is included below.

4.5.1. **Fisher score**—Fisher score is one of the most-widely used supervised feature selection methods. The Fisher scoring algorithm tends to select features whose values are similar within each class but dissimilar among different classes. The Fisher score $s(f_i)$ of feature $f_i$ is computed as:

$$s(f_i) = \frac{\sum_j n_j (\mu_{ij} - \mu_i)}{\sum_j n_j \sigma_{ij}^2}$$

where $n_j$ is the number of samples for class $j$, $N$ is the number of classes, $\mu_i$ is the mean value of feature $f_i$, and $\mu_{ij}$ and $\sigma_{ij}^2$ are the mean and variance of feature $f_i$ in class $j$ respectively. Features with the largest Fisher scores are selected.

4.5.2. **Extratree classifier**—Extratree classifier proposed by Geurts et al. (2006) is an ensemble method based on decision trees. For building each decision tree, this algorithm first subsamples the training data without replacement. Then, at each level, the best splitting node is selected from a random subset of features based on certain criteria, such as Gini index or entropy. The algorithm repeats previous steps to aggregate multiple decision trees, and the majority vote is used as prediction. We implemented this method using the sklearn package Pedregosa et al. (2011). We used Gini index as the criterion for both decision tree splitting and feature importance ranking.

4.5.3. **Minimum redundancy and maximum relevancy**—Minimum Redundancy and Maximum Relevancy (MRMR), proposed by Peng et al. (2005), is one representative feature selection criteria in the family of information-theory-based methods. Information-theory-based algorithms tend to maximize relevancy by selecting features that have high mutual information with the class labels. A typical workflow is to (1) initialize an empty set, (2) perform forward feature selection given all available features according to certain criteria, (3) greedily choose the feature with the highest score and add it to the selected feature set, and (4) repeat (2)–(3) until the desirable number of features is selected. For feature $F_k$, the MRMR criteria can be written as,

$$J_{MRMR}(F_k) = I(F_k, Y) - \frac{1}{|S|} \sum_{F_j \in S} I(F_k, F_j)$$

where $I(F, Y)$ is the mutual information between feature $F$ and label $Y$, and $S$ is the currently selected feature set. We implemented MRMR using the package developed by Homola (2016).

4.5.4. **Chi-squared score**—Chi-squared (Liu and Setiono, 1995) is a statistical measurement of the independence between a feature and class labels. Given a feature $f_i$ with $M$ different values, its Chi-squared score $\chi^2(f_i)$ can be computed as follows,
\[ \chi^2(f_i) = \sum_{j=1}^{M} \sum_{y=1}^{N} \frac{(n_{jy} - \mu_{jy})^2}{\mu_{jy}} \]

where \(N\) is the number classes, \(n_{jy}\) is the number of samples in the \(y\)th class for which \(f_i\) has its \(j\)th value, and \(\mu_{jy}\) is the expected value of \(n_{jy} \cdot \mu_{jy} = \left(\sum_{j=1}^{M} n_{jy}\right) \left(\sum_{y=1}^{N} n_{jy}\right) / \left(\sum_{j=1}^{M} \sum_{y=1}^{N} n_{jy}\right)\). A higher \(\chi^2(f_i)\) indicates greater importance of the feature \(f_i\) for distinguishing among classes. We implemented Chi-squared score using the sklearn package.

To test robustness of these feature selection methods, we evaluated all four algorithms by varying the number of selected top weighted features from 10 to 100. We also performed one-vs-one and one-vs-all classifications for the top 100 features selected by each algorithm. We used t-SNE (Van der Maaten and Hinton, 2008) to visualize the underlying space of infant vocalizations by selecting the top 5 features from all pairwise one-vs-one classifications for each feature selection algorithm. Further analysis of the top 30 features overlapped among four different feature selection methods are also discussed. Detailed results are included in Section 5.3.

### 4.6. Experimental setup

We implemented LDA using sklearn (Pedregosa et al., 2011) and both of our NN based models using PyTorch (Paszke et al., 2019). We used three-fold cross validation to evaluate all of our experiments. Other experimental configuration details are shown in Table 13. Apart from training all proposed models using default OresSMILE features, we performed various experiments for further improvements on mostly NN based models. Detailed results and discussion are present in Section 5.

### 5. Results

LDA, FCN, and CNSA classifiers were used to classify infant vocalizations. Results for benchmark evaluation on the CRIED dataset are described in Section 5.1. Results for the LDA, FCN and CNSA are presented in Section 5.2. We evaluated our classifiers using multiple metrics, including accuracy (percentage of correct classifications), weighted F1, macro F1, and unweighted-average recall (UAR) scores. Let \(R[k]\) and \(F1[k]\) be defined as the recall and the harmonic mean of precision and recall for the \(k\)th vocalization type respectively. Weighted F1 (WF1) and macro F1 (MF1) are two different methods of averaging \(F1[k]\) across classes, and UAR measures unweighted recall scores across classes:

\[
WF1 = \frac{N_k}{N} \sum_{k=1}^{K} F1[k], \quad MF1 = \frac{1}{K} \sum_{k=1}^{K} F1[k], \quad UAR = \frac{1}{K} \sum_{k=1}^{K} R[k]
\]

where \(N_k\) is the number of test tokens in class \(k\), and \(N\) is the total number of test tokens.
Feature selection analyses are presented in Section 5.3. Classification results for adult vocalization type are discussed in Section 5.4, and emotional valence classification results for infant-directed speech are discussed in Section 5.5.

5.1. Infant vocalizations: benchmark evaluation

Table 14 shows accuracy, macro F1 score, and UAR for LOSO-training (Leave One Subject Out) and LOSO-testing in the CRIED dataset for our proposed classifiers, and for models published in previous studies. Because UAR is used as evaluation metric in the Interspeech 2018 paralinguistic challenge (Schuller et al., 2018) for this dataset, we explored additional training strategies for optimizing UAR. We used a weighted sampler by assigning sampling weight for each sample belonging to class \( k \) as \( N/N_k \), which is inversely proportional to the number of samples for class \( k \). In this way, samples of underrepresented classes are given more weight than those of the majority class. We also reduced the dimension of the feature vector to 1000 using Fisher-score feature selection. Both of these training strategies are the subject of ablation studies in Section 5.2, using our default feature set and dataset.

We compare our results to four Interspeech 2018 baseline models, and two competition models. The baseline models are: (1) END2YOU: a CNN followed by subsequent GRUs given raw waveforms of the audio, (2) OpenSMILE acoustic feature set + SVM: SVM is used to classify the Interspeech CompParE feature set (6373 features) extracted by OpenSMILE, (3) OpenXBOX + SVM: SVM is used to classify the Bag-of-Audio-Word (BoAW) learned from CompParE acoustic feature set using the OpenXBOX toolkit, and (4) AUDEEP: SVM is used to classify feature set obtained by unsupervised autoencoder representation learning using the AUDEEP toolkit. Turan and Erzin (2018) proposed Capsule Network (CapsNet) to learn local activation and pose components with additional voice activation detection module given extracted spectrograms for baby cry sound classifications. Huckvale (2018) investigated jointly optimizing temporal (126 parameters for every 10 ms) and summative CompParE features using a combined NN architecture based on LSTM and a dense layer. The best UAR trained under each optimized setting for these studies is shown in Table 14.

Table 14 shows that using an unweighted sampler maximizes accuracy but outputs much lower UAR. Our best accuracy (86.72%) is achieved by training CNSA on the default + comp feature set using the unweighted sampler, which outperformed accuracy of all other previous studies. For both FCN and CNSA, training with either weighted sampler or Fisher scores selection will improve UAR at the expense of reduced accuracy. By using only our default + comp feature set (1848 parameters), we achieved the best UAR (76.15%) with both weighted sampling and Fisher scores selection. For further improving UAR, we explored augmenting our feature set by using the CompParE feature set, and we achieved the best UAR for 77.96% with weighted sampler and Fisher scores selection. This result beats the best UAR proposed in previous studies by about 1%.

5.2. Infant vocalizations: LDA, FCN and CNSA classifiers

Infant vocalization type classifiers are compared, for our TDP and IDP datasets, using three metrics: accuracy (percentage of correct classifications), weighted F1, and macro F1. Table
15 shows mean and standard deviations for three-fold cross validation tests of accuracy, weighted F1 score, and macro F1 score for infant vocalizations. All three measures, for all experiments, have relatively small standard deviation ($\sigma < 3\%$ for accuracy, $\sigma < 0.5\%$ for weighted F1 score and $\sigma < 1.5\%$ for macro F1 score), which indicates that different folds of cross-validation give similar results.

FCN and CNSA significantly outperform LDA (Table 16: matched-pair t-tests, Gillick and Cox, 1989). This finding is contrary to the result reported in Xu et al. (2018), in which it was reported that LDA outperformed a smaller FCN for the classification of child vocalizations using a subset of the same dataset. It is likely that the improved accuracy of FCN in the current study is simply a consequence of the increased dataset size. Both FCN and CNSA models achieved comparable accuracy and weighted F1 score for various experiments, but the FCN tends to outperform the CNSA in terms of macro F1. The FCN trained on the default OpenSMILE feature set already achieves relatively good results; with augmented data, and using the default + complementary features, we were able to improve marginally both its accuracy and F1 scores. Specifically, by augmenting the training datasets with data drawn from the Google AudioSet and Freesound databases, we achieve the better macro F1 score, 56.71% at the cost of lower raw accuracy and lower weighted F1 score. Feature augmentation, using the default and complementary feature sets together, provides marginal benefits to all three metrics by 0.31%, 0.22%, and 0.85% respectively. With both data and feature augmentation, we also see marginal improvements of all three metrics compared with the FCN trained only on default features. We also explored the Fisher-score selection of the top 1000 features, and training with a weighted sampler, as in the benchmark evaluation task. We achieved the best macro F1 score, 58.04%, using Fisher-score feature selection, but with an unweighted sampler. Training with a weighted sampler significantly degraded both accuracy and macro-F1, relative to the unweighted sampler.

Unlike FCN, under the same condition, CNSA achieved comparable accuracy and macro F1 score when trained with weighted sampler + Fisher scores selection. The CNSA was also tested using raw Mel spectrogram features, with weights pre-trained using the RAVDESS corpus; this setting did not outperform the default + complementary feature set trained only on in-domain data. With both data and feature augmentation, CNSA achieved the best accuracy 81.54% out of all classifiers.

5.3. Infant vocalizations: feature selection analysis

Four feature selection algorithms (Fisher score, Extratree, MRMR and Chi-squared) were used to select small feature vectors (10-, 20-, 50-, and 100-dimensional vectors) for each classification problem. Fig. 2 shows accuracy and macro F1 scores as a function of feature selection algorithm and feature dimension. The dashed lines indicate the accuracy and macro F1 of a 46-feature overlap set, whose composition is described in the next paragraph. As shown, this 46-feature overlap set is usually a little better than the 50-dimensional feature vectors selected by any one algorithm, but not as good as the 100-dimensional vectors. In general, as the number of selected features increases, accuracy and F1 score increase. MRMR has the best performance for up to 50 selected features; Extratree is best if there are 100 features. Compared with LDA, FCN generally has higher accuracy but lower F1
score, for these small feature vectors; with 1000 or more features, FCN outperformed LDA as shown in Section 5.2.

It is possible that some acoustic features may contain useful information about just one vocalization type, and might therefore be missed in a 5-class analysis like that shown in Fig. 2. In order to select the most important features for each vocalization type, while also limiting the number of such selected features, we performed a multi-step voting procedure in order to construct an overlap feature vector containing the 46 features shown in Table 17. First, we compared one-vs-all and one-vs-one binary classification problems. The top 30 features were selected for every one-vs-one binary classification problem, e.g., for the LAU-vs-SCR problem. Features that did not also show up in one of the two one-vs-all problems (LAU-vs-all or SCR-vs-all) were discarded. This process was repeated for all four feature selection algorithms: the overlap vector contains the features that were selected, in this way, by at least three of the four feature selection algorithms. To ensure robustness of this overlap feature set, we tested it in LDA and FCN 5-way classifiers. The mean and standard deviation for accuracy and macro f1 score are 74.49% ± 0.24% and 43.07% ± 0.81% for LDA, 77.20 ± 0.23% and 42.72% ± 0.90% for 2-layer FCN, which are indicated in Fig. 2. Compared with the 50-dimensional feature vectors selected by individual feature algorithm in Fig. 2, this overlap feature set had the best performance for both LDA and FCN except a slightly lower accuracy (74.49%) compared with MRMR (74.65%) for LDA, even though it consists of only 46 features.

Table 17 shows that most of the top selected features for CRY and LAU are used to distinguish those two categories from BAB and FUS. Top selected features for FUS are used to mainly distinguish from BAB. Apparently, most categories are hard to distinguish from BAB and FUS. The set of overlap features for SCR are highly similar across the columns of the table. This is possibly because SCR is an acoustically unique class, distinguished from all other categories by features related to the LspFreq[0] LLD. Further detailed analysis of the features shown in Table 17 is provided in Section 6.

We used t-SNE to visualize the overlap among the five vocalization types. For each feature selection algorithm, we created a union feature vector from the top 5 features from all one-vs-one classifications. There are 10 one-vs-one classifiers, but many of these classifiers select similar features, therefore the union vectors are always much smaller than $10 \times 5 = 50$ dimensions: Fisher-selection has a 35-dimensional union vector, Extratree has 40, MRMR has 47, and Chi-square has 28. LDA is then used to compute a 4-dimensional subspace of each union vector, then we feed the reduced feature vectors to t-SNE, implemented in sklearn with the package default hyperparameters. Fig. 3 shows the resulting clusters, plotted for a uniform subsample of the test dataset of one fold (600 samples for BAB, 120 samples for each other class). A general pattern of those clusters is that CRY, LAU, and SCR form their own clusters at the edge of BAB while BAB forms the majority of the dataset. FUS does not have a clear cluster boundary and lies between BAB and the other classes.

### 5.4. Results for mother vocalizations

Table 18 shows mean and standard deviations of accuracy, weighted F1 score, and macro F1 score for three-fold cross-validation tests for mother vocalizations. Table 19 shows statistical
significance tests results for mother vocalizations. Results for mother vocalizations have patterns similar to those for infant vocalizations. NN based models are better than LDA. Among NN based models, the FCN has a non-significant tendency to perform better than the CNSA; the CNSA achieves comparable accuracy and weighted F1 scores but lower macro F1 score compared with the FCN under the same experimental conditions. With augmented data, the 2-layer FCN marginally improves its macro F1 score, by 0.92%.

5.5. Emotional valence classification of infant-directed speech

Infant-directed speech was classified into POSITIVE vs. NEUTRAL emotional valence, using associated video labels as described in Section 4.3. Table 20 gives results for this binary emotional valence labeling task. The two-layer FCN outperformed LDA in all three metrics. Accuracy of the FCN is not as high as it would be for acted speech (accuracy of the full eight-class emotion classifier, using the RAVDESS corpus of acted emotion, was 71.88%), but the accuracy achieved by the FCN and CNSA in Table 20 is significantly better than the LDA, as shown in statistical significance test results in Table 21, indicating that the acoustic features in infant-directed speech contain sufficient information to classify POSITIVE vs. NEUTRAL emotional valence with significantly different accuracies for neural network vs. LDA classifiers.

6. Discussion

This section presents analysis of models trained under various settings, comparisons of classification methods between various feature selection methods and human coders, and detailed visualizations, describing the pairwise differences between vocalization types for infants and adults.

6.1. Overall analysis of models

In this study, we compared three models on automatic infant/mother vocalization classifications. Because of different learning capabilities of these models, the accuracies and F1 scores diverge as shown in Tables 15 and 18. Table 22 includes accuracy, weighted F1 score, and macro F1 score separately for all three folds of cross-validation for mother vocalizations, for the 2-layer FCN model with augmented data. Comparing the accuracy across folds, it is possible to observe that fold 2 has the lowest training-corpus accuracy (and weighted and macro F1), but the highest test-corpus accuracy (and weighted and macro F1), suggesting that the network is overfitting the training corpus. The FCN and CNSA show similar trends across all experimental conditions, therefore we can perhaps infer that both of these classifiers are overfitting the training data. Since they are overfitting, we would normally expect to see improvements on these classifiers by applying web data augmentation (Lamel et al., 2002; Gorin et al., 2016). However, we barely gained any benefit from data augmentation, apparently because the in-domain and out-domain datasets had different acoustic feature distributions. The design of complementary acoustic features, on the other hand, provided small benefits in almost all cases. Feature selection methods using Fisher scores further improve macro F1 scores as it alleviates overfitting problems of the training data because of the reduction of number parameters of the features. Future
work can be done by exploring advanced data augmentation and complementary feature
techniques for further improving these classifiers.

6.2. Comparisons of classification methods between feature selection and human coders

Table 5 provides descriptions of each infant vocalization code, as it was used by human coders. Table 17 lists top weighted features for each pairwise infant vocalization classification overlapped among at least three feature selection methods.

Based on Table 5, we notice that cry and fus are possibly the two most confusable classes for human coders, because descriptions for other classes are very different from those of cry and fus. According to the descriptions, fus has lower intensity but more continuous vocalizations than cry. For feature selection methods, logMelFreqBand[4] and logMelFreqBand[5] were selected as the most discriminative LLDs for cry vs. fus. Unlike human coders, who are instructed to differentiate cry vs. fus using primarily intensity/energy, feature selection methods were able to capture the difference using the middle logMelFreqBands.

For human coders, descriptions in Table 5 suggest that lau can be easily discriminated from other classes. For feature selection methods, F0 and the first three logMelFreqBand were selected. F0 was selected by most classes, not just lau. For lau-related classifications, variations among the lower logMelFreqBand are also useful.

For bab, human coders can easily identify it as speech-like vocalizations. For feature selections, classifiers aimed to find the difference between bab vs. fus as the hidden cluster showed that fus were mostly overlapped with bab. Feature selection methods use F0 and two high-frequency MFCC (mfcc[11] and mfcc[12]) to classify those two classes. Bab vs. fus has one of the lowest F1 scores.

For scr, human coders discriminate it depending on its high pitch. Surprisingly, feature selection methods do not use F0 to distinguish this category, apparently because its typical pitch frequency is too high to be tracked by any standard pitch tracker. Instead, feature selection methods identified the screeching pitch as a high-frequency narrowband peak, well-tracked by the first two LSPFreq features and the standard first formant F1 feature.

6.3. Fundamental frequency

A large number of features in the overlap set (Table 17) are based on F0. Fig. 4 shows the fundamental frequency contour, overlaid on the spectrogram, for two examples of each of the five infant vocalization types. For cry, the F0 contour is generally continuous, with occasional pitch-doubling discontinuities, as shown in the example on the left. The F0 of fus, on the other hand, shows more extended and less broken up vocalizations but with generally lower F0. Lau has F0 bouncing up and down across all of the frequency range. For bab, babies are trying to produce speech-like syllables: we observe that most of the energy is concentrated in lower-frequency bands, and the F0 contour looks speech-like. For scr, by contrast, F0 is mostly around 1500–3000 Hz, and this is also the most obvious characteristic that human coders used to identify scr as described in Table 5.
6.4. Log Mel frequency band energies

The logMelFreqBand LLDs are bandpass energies, computed in 8 frequency bands sampled uniformly on the Mel scale between 0 Hz and 8000 Hz.

Table 17 shows that the class LAU was discriminated from the classes CRY, FUS, and BAB by having less variability, than other classes, in its low-frequency logMelFreqBands values. As Fig. 5 shows, LAU, BAB, and CRY all have relatively high energy at low frequencies. However, LAU shows less variability of energy in the lower Mel frequency bands, so linregerrA and iqr2–3 of logMelFreqBand(d)[0] + linregerrA, iqr1–3 and iqr2–3 of logMelFreqBand(d)[1] were used for discriminating between LAU versus CRY, FUS and BAB.

Relatively higher log Mel Frequency bands (logMelFreqBand[4–5]) were used to discriminate CRY vs. FUS and CRY vs. BAB. Table 5 shows that the definition of the CRY/FUS distinction used by human coders emphasizes the sustained high-frequency energy of the CRY class, so it is not surprising that functionals based on logMelFreqBand[4–5] are selected to discriminate CRY vs. FUS/BAB. Fig. 6 provides spectrograms of examples of logMelFreqBand[4–5] of CRY, FUS and BAB.

6.5. LSP frequency and first formant frequency

The SCR category is defined by an extremely high-pitched screeching noise: although F0 is typically 500 Hz for most infant vocalizations, the SCR category is defined almost uniformly by F0 above 1000 Hz. Table 17 shows that features based on F0 were sometimes chosen to distinguish SCR, but that features based on F1 and lspFreq[0–1] were selected more often. We speculate that F1 and lspFreq[0–1] are serving as proxy measures for the very high pitch, and are being selected because the F0 tracker often fails to correctly track the very high pitch frequencies of SCR. Figs. 7 and 8 presents plots of lspFreq[0–1] and F1 value, overlaid on the spectrogram, for selected samples for all classes. During all other classes, the F1 measure tracks the true first formant frequency of an infant at around 600–1000 Hz (Kent and Murray, 1982), the lspFreq[1] measures tracks the second formant at around 2000–3000 Hz, and the lspFreq[0] measure is somewhere in between. During SCR, the F1 measure tracks the pitch frequency, which is in the range of 1500–2000 Hz, while both lspFreq[0] and lspFreq[1] track its second harmonic, at 3000–4000 Hz.

7. Conclusions

In this study, three types of classifiers were evaluated on a benchmark dataset. CNSA-based classifier achieved the best UAR compared with previous studies. These classifiers were also trained and tested to classify vocalizations produced by infants and mothers. Utterances were acquired from in-home recordings, and from dyadic social interaction experiments in the laboratory. Neural net based models achieved significantly better performance than either LDA for classifying both mother and infant vocalizations; the 2-layer FCN showed a non-significant tendency to outperform the CNSA. Augmenting our spoken corpus by adding relevant infant vocalization and adult speech samples from external corpora produces small improvements in the macro F1, but small decreases in accuracy; these changes were not tested for statistical significance. Employing top 1000 features selected by Fisher scores
yields larger macro F1 improvement without sacrificing too much accuracy. The creation of complementary acoustic features, including extra prosodic and phonetic features, improves accuracy, the weighted F1 score and the macro F1 score. By cross-examining multiple feature selection algorithms for each pairwise classifications among the infant vocalization types, we found that F0 is by far the most frequently selected LLD for discriminating these classes, but that useful complementary information is provided by the dynamics of low-frequency logMelFreqBands (for LAU specifically), the average levels of high-frequency logMelFreqBands (for CRY specifically), and the features lspFreq[0–1], and F1 (which are apparently used as proxy measures to detect the extremely high pitch frequencies produced during SCR). By plotting those important feature contours overlaid on spectrograms, it is possible, in some cases, to visualize the differences among vocalization categories.

Acknowledgments

This work was supported by funding from the National Institute on Drug Abuse, United States (R34DA050256), the National Institute of Mental Health, United States (R21MH112578) and the National Institute of Food and Agriculture, U.S. Department of Agriculture (ILLU-793-339).

We also thank Dr. Katrin D. Bartl-Pokorny, Univ. Prof. Dr. Christa Einspieler, in Assoc. Prof. Dr. Peter B. Marschik, Dr. Florian B. Pokorny, Dr. Dajie Zhang at the interdisciplinary Developmental Neuroscience, Division of Phoniatrics, Medical University of Graz, Austria, and any other contributors for their input of study design and data development of the CRIED database.

References

Aldeneh Z, Mower-Provost E, 2017. Using regional saliency for speech emotion recognition. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2741–2745. 10.1109/ICASSP.2017.7952655.

An N, Thanh N, Liu Y, 2019. Deep CNNs with self-attention for speaker identification.. IEEE Access 7, 85327–85337. 10.1109/ACCESS.2019.2917470.

Anders F, Hlawitschka M, Fuchs M, 2020. Comparison of artificial neural network types for infant vocalization classification. IEEE/ACM Trans. Audio Speech Lang. Process 1. 10.1109/taasl.2020.3037414.

Boersma P, 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. IFA Proc. 17, 97–110.

Boersma P, Weenink D, 2018. Praat: doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 3 February 2018 http://www.praat.org/.

Boersma P, Weenink D, 2020. Praat: doing phonetics by computer. URL: from http://www.praat.org/.

Burkhardt F, Paeschke A, Rolfes M, Sendlmeyer WF, Weiss B, 2005. A database of german emotional speech. In: Interspeech. pp. 1517–1520.

Busso C, Bulut M, Lee C-C, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS, 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Lang. Resour. Eval 42 (4), 335–359.

Dallaert F, Polzin T, Waibel A, 1996. Recognizing emotion in speech. In: Interspeech. pp. 1970–1973.

Dongxin X, Richards JA, Gilkerson J, 2014. Automated analysis of child phonetic production using naturalistic recordings. J. Speech Lang. Hear. Res 57 (5), 1638–1650. 10.1044/2014_JSLHR-S-13-0037.

Dunstan P, 2012. Calm the Crying: Using the Dunstan Baby Language. Penguin Books Ltd., London.

Ebrahimpour MK, Schneider S, Noelle DC, Kello CT, 2020. InfantNet: A deep neural network for analyzing infant vocalizations. arXiv e-prints, arXiv:2005.12412, arXiv:2005.12412
El Ayadi M, Kamel MS, Karray F, 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognit. 44 (3), 572–587.

Eyben F, Wöllmer M, Schuller B, 2010. OpenSMILE - The Munich versatile and fast open-source audio feature extractor. In: MM'10 - Proceedings of the ACM Multimedia 2010 International Conference. Institute for Human-Machine Communication, Technische Universität München, pp. 1459–1462.

Fernald A, 1985. Four-month-old infants prefer to listen to motherese. Infant Behav. Dev 8, 181–195.

Font F, Roma G, Serra X, 2013. Freesound technical demo. In: ACM International Conference on Multimedia (MM’13). ACM, Barcelona, Spain, pp. 411–412. 10.1145/2502081.2502245.

Garnica O, 1977. Some prosodic and paralinguistic features of speech to very young children. In: Snow CE, Ferguson CA (Eds.), Talking to Children: Language Input and Acquisition. Cambridge University Press, Cambridge, England, pp. 63–88.

Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Lawrence W, Moore RC, Plakal M, Ritter M, 2017. Audio Set: An ontology and human-labeled dataset for audio events. In: Proc. IEEE ICASSP 2017 New Orleans, LA. pp. 776–780.

Geurts P, Ernst D, Wehenkel L, 2006. Extremely randomized trees. Mach. Learn 63 (1), 3–42.

Gillick L, Cox SJ, 1989. Some statistical issues in the comparison of speech recognition algorithms. In: International Conference on Acoustics, Speech, and Signal Processing. Vol. 1. pp. 532–535. 10.1109/ICASSP.1989.266481.

Gorin A, Lileikyté R, Huang G, Lamel L, Gauvain J-L, Laurent A, 2016.

Gozsilovszky G, 2019. Using Fisher vector and bag-of-audio-words representations to identify styrian dialects, sleepiness, baby & orca sounds. In: Proc. Interspeech pp. 2413–2417.

Gujral A, Feng K, Mandhyan G, Snehil N, Chaspari T, 2019. Leveraging transfer learning techniques for classifying infant vocalizations. In: 2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI). pp. 1–4. 10.1109/BHI.2019.8834666. [PubMed: 32864624]

Hanson HM, 1997. Glottal characteristics of female speakers: Acoustic correlates. J. Acoust. Soc. Am 101, 466–481. [PubMed: 9000737]

He D, Cheng Z, Hasegawa-Johnson M, Chen D, 2017. Using approximated auditory roughness as a pre-filtering feature for human screaming and affective speech AED. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Vol. 2017-August, pp. 1914–1918. 10.21437/Interspeech.2017-593.

He K, Zhang X, Ren S, Sun J, 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hinton GE, Krizhevsky A, Wang SD, 2011. Transforming auto-encoders. In: International Conference on Artificial Neural Networks. Springer, pp. 44–51.

Homola D, 2016. A python package for parallelized mutual information based feature selection. URL: https://danielhomola.com/feature%20selection/phd/mifs-parallelized-mutual-information-based-feature-selection-module/.

Huckvale D, 2018. Neural network architecture that combines temporal and summative features for infant cry classification in the interspeech 2018 computational paralinguistics challenge. In: Proc. Interspeech 2018. pp. 137–141. 10.21437/Interspeech.2018-1859.

Inoue T, Nakagawa R, Kondou M, Koga T, Shinohara K, 2011. Discrimination between mothers’ infant- and adult-directed speech using hidden Markov models. Neurosci. Res 70 (1), 62–70. 10.1016/j.neures.2011.01.010. [PubMed: 21256898]

Ji C, Basodi S, Xiao X, Pan Y, 2020. Infant sound classification on multi-stage CNNs with hybrid features and prior knowledge. In: International Conference on AI and Mobile Services. Springer, pp. 3–16.

Ji C, Chen M, Li B, Pan Y, 2021a. Infant cry classification with graph convolutional networks. arXiv preprint arXiv:2102.02909

Ji C, Mudiyanselage TB, Gao Y, Pan Y, 2021b. A review of infant cry analysis and classification. Eurasip J. Audio Speech Music Process 2021 (1), 1–17.

Johnson WF, Emde RN, Scherer KR, Klinnert MD, 1986. Recognition of emotion from vocal cues. Arch. Gen. Psychiatry 43, 280–283. [PubMed: 3954549]
Kalashnikova M, Burnham D. 2018. Infant-directed speech from seven to nineteen months has similar acoustic properties but different functions. J. Child Lang 45, 1035–1053. 10.1017/S0305000917000629. [PubMed: 29502549]

Kent RD, Murray AD. 1982. Acoustic features of infant vocalic utterances at 3, 6, and 9 months. J. Acoust. Soc. Am 72, 353–365. 10.1121/1.388089. [PubMed: 719278]

Knox M, Mirghafori N. 2007. Automatic laughter detection using neural networks. In: International Speech Communication Association - 8th Annual Conference of the International Speech Communication Association, Interspeech 2007, Vol. 4, fourth ed. pp. 2364–2367.

Lamel L, Gauvain J-L, Adda G. 2002. Lightly supervised and unsupervised acoustic model training. Comput. Speech Lang 16, 115–129.

Li J, Cheng K, Wang S, Morstatter F, Trevino R, Tang J, Liu H. 2017. Feature selection: A data perspective. ACM Comput. Surv (ISSN: 0360–0300) 50 (6), 10.1145/3136625.

Lim BP. 2011. Computational Differences between Whispered and Non-Whispered Speech. (Ph.D. thesis) University of Illinois at Urbana-Champaign, USA.

Liu H, Setiono R. 1995. Chi2: Feature selection and discretization of numeric attributes. In: Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence. IEEE, pp. 388–391.

Livingstone SR, Russo FA. 2018. The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE (ISSN: 19326203) 13 (5), 1–35.

Van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. J. Mach. Learn. Res 9 (11).

Maghfira TN, Basarudtin T, Krisnadhi A. 2020. Infant cry classification using CNN – RNN. J. Phys. Conf. Ser 1528, 012019. 10.1088/1742-6596/1528/1/012019.

Mahdhaoui A, Chetouani M. 2011. Supervised and semi-supervised infant-directed speech classification for parent-child interaction analysis. Speech Commun. 53 (9), 1149–1161. 10.1016/j.specom.2011.05.005.

Marschik P, Pokorny F, Peharz R, Zhang D, Krieber M, Tomantschger I, Bartl-Pokorny K, Einspieler C, Bölte S, Tamminies K, O’Muircheartaigh J, Roeyers H, Spitte A, Urlesberger B, Schuller B, Poustka L, Ozonoff S, Pernkopf F, Pock T, Enzinger C, Sigalsagas J, Roche L, Enzinger C, Gugatschka M, Nielsen-Saines K, Kaufmann W. 2017. A novel way to measure and predict development: A heuristic approach to facilitate the early detection of neurodevelopmental disorders. Curr. Neurol. Neurosci. Rep 17 (5).

Mccowan I, Lathoud G, Lincoln M, Lisowska A, Post W, Reidsma D, Wellner P. 2005. The AMI meeting corpus. In: Noldus LPJJ and Grieco F and Loijens LWS and Zimmerman PH (Eds.), 5th International Conference on Methods and Techniques in Behavioral Research. Wageningen, pp. 137–140.

Moore R, Lewis W. 2010. Intelligent selection of language model training data. In: ACL 2010 – 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. Microsoft Research, pp. 220–224.

Papaeliou C, Minadakis G, Cavouras D. 2002. Acoustic patterns of infant vocalizations expressing emotions and communicative functions. J. Speech Lang. Hear. Res 45, 311–317. 10.1044/1092-4388(2002/024), URL: http://pubs.asha.org/doi/10.1044/1092-4388. [PubMed: 12003513]

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. 2019. Pytorch: An imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d’Alché Buc F, Fox E, Garnett R (Eds.), Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res 12, 2825–2830.
Peng H, Long F, Ding C, 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell 27 (8), 1226–1238. 10.1109/TPAMI.2005.159. [PubMed: 16119262]

Petroni M, Malowany AS, Johnston CC, Stevens BJ, 1995. Classification of infant cry vocalizations using artificial neural networks (ANNs). In: 1995 International Conference on Acoustics, Speech, and Signal Processing, Vol. 5. pp. 3475–3478. 10.1109/ICASSP.1995.479734.

Pretzer GM, Lopez LD, Walle EA, Warlaumont AS, 2019. Infant-adult vocal interaction dynamics depend on infant vocal type, child-directedness of adult speech, and timeframe. Infant Behav. Dev (ISSN: 0163–6383) 57, 101325. 10.1016/j.infbeh.2019.04.007. [PubMed: 31100586]

Reggiannini B, Sheinkopf SJ, Silverman HF, Li X, Lester BM, 2013. A flexible analysis tool for the quantitative acoustic assessment of infant cry. J. Speech Lang. Hear. Res 56, 1416–1428. [PubMed: 23785178]

Reyes-Galaviz OF, Cano-Ortiz SD, Reyes-García CA, 2008. Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies. In: 2008 Seventh Mexican International Conference on Artificial Intelligence. IEEE, pp. 330–335.

Sachs J, 1977. The adaptive significance of linguistic input to prelinguistic infants. In: Snow CE, Ferguson CA (Eds.), Talking to Children: Language Input and Acquisition. Cambridge University Press, Cambridge, England, pp. 51–62.

Scherer KR, 1996. Adding the affective dimension: A new look in speech analysis and synthesis. In: Interspeech. pp. 1811–1814.

Schuller BW, Batliner A, Bergler C, Pokorny FB, Krajewski J, Cychosz M, Vollmann R, Roelen SD, Schnieder S, Bergelson E, Cristia A, Steidl A, Warlaumont AS, Yankowitz L, Nöth E, Amiriparian S, Hantke S, Schmitt M, 2019. The INTERSPEECH 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTER-SPEECH, Vol. 2019-September. International Speech Communication Association, pp. 2378–2382. 10.21437/Interspeech.2019-1122.

Schuller B, Steidl S, Batliner A, 2009. The INTERSPEECH 2009 emotion challenge. In: INTERSPEECH. pp. 312–315.

Schuller B, Steidl S, Batliner A, Burkhart F, Devillers L, Müller C, Narayanan S, 2010. The INTERSPEECH 2010 paralinguistic challenge. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010. pp. 2794–2797.

Schuller B, Steidl S, Batliner A, Marschik P, Baumeister H, Dong F, Hantke S, Pokorny FB, Rathner E, Bartl-Pokorny KD, Einspieler C, Zhang D, Baird A, Amiriparian S, Qian K, Ren Z, Schmitt M, Tzirakis P, Zafeiriou S, 2018. The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats. In: INTERSPEECH. pp. 122–126.

Szekely C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Van-houcke V, Rabinovich A, 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9.

Thiessen ED, Hill EA, Saffran JR, 2005. Infant-directed speech facilitates word segmentation. Infancy 7 (1), 53–71. 10.1207/s15327078in0701_5. [PubMed: 33430544]

Tronick E, Als H, Adamson L, Wise S, Brazelton T, 1978. The infant’s response to entrapment between contradictory messages in face-to-face interaction. J. Amer. Acad. Child Psychiatry 17, 1–13. [PubMed: 632477]

Turan MAT, Erzin E, 2018. Monitoring infant’s emotional cry in domestic environments using the capsule network architecture. In: Proc. Interspeech 2018. pp. 132–136. 10.21437/Interspeech.2018-2187.

Tzinis E, Potamianos A, 2017. Segment-based speech emotion recognition using recurrent neural networks. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, pp. 190–195.

Van Bezooijen R, Otto SA, Heenan TA, 1983. Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics. J. Cross-Cult. Psychol 14 (4), 387–406.

Vassilakis P, 2005. Auditory roughness as means of musical expression. Sel. Rep. Ethnomusicol 12, 119–144.

Speech Commun. Author manuscript; available in PMC 2022 September 01.
Warlaumont AS, Oller DK, Buder EH, Dale R, Kozma R, 2010. Data-driven automated acoustic analysis of human infant vocalizations using neural network tools. J. Acoust. Soc. Am (ISSN: 0001–4966) 127, 2563–2577. 10.1121/1.3327460. [PubMed: 20370038]

Van de Weijer J, 2001. Vowels in Infant- and Adult-Directed Speech, Vol. 49. Lund University Working Papers in Linguistics.

Williams CE, Stevens KN, 1972. Emotions and speech: Some acoustical correlates. J. Acoust. Soc. Am 52 (48), 1238–1250. [PubMed: 4638039]

Xu Y, Hasegawa-Johnson M, McElwain N, 2018. Infant emotional outbursts detection in infant-parent spoken interactions. In: Proc. Interspeech 2018. pp. 242–246. 10.21437/Interspeech.2018-2429.

Yamamoto S, Yoshitomi Y, Tabuse M, Kushida K, Asada T, 2013. Recognition of a baby’s emotional cry towards robotics baby caregiver. Int. J. Adv. Robot. Syst 10 (86), 1–7.

Yeh S-L, Chao G-Y, Su B-H, Huang Y-L, Lin M-H, Tsai Y-C, Tai Y-W, Lu Z-C, Chen C-Y, Tai T-M, Tseng C-W, Lee C-K, Lee C-C, 2019. Attention networks and adversarial augmentation for styrian dialect contiuous sleepiness and baby sound recognition. In: Proc. Interspeech pp. 2398–2402.

Yumoto E, Gould WJ, 1982. Harmonics-to-noise ratio as an index of the degree of hoarseness. J. Acoust. Soc. Am 71, 1544–1550. [PubMed: 7108029]

Zhang L, Wang L, Dang J, Guo L, Guan H, 2018. Convolutional neural network with spectrogram and perceptual features for speech emotion recognition. In: Cheng L, Leung ACS, Ozawa S (Eds.), Neural Information Processing. Springer International Publishing, Cham, pp. 62–71.
Fig. 1.
Overview of the CNSA.
Fig. 2.
Number of top selected features vs. average accuracy and macro F1 score for three-fold cross validation tests for FCN (left column) and LDA (right column). Four feature selection methods (Fisher, Exratree, MRMR, and Chi square) are included. Dashed lines indicate accuracy and F1 of a 46-feature overlap set, composed of features selected by all four algorithms.
Fig. 3. Clusters computed by selecting top 5 features from each one-vs-one classifier, merging the resulting feature vector, then projecting using LDA followed by t-SNE.
Fig. 4.
Fundamental frequency contour overlaid on the spectrogram for two selected audio samples in each class. Blue dots with white outline are detected F0 values.
Fig. 5.
Linear-frequency spectrogram and logMelFreqBand(d)[0–1] for CRY, FUS, LAU and BAB.
Fig. 6.
Linear-frequency spectrogram and logMelFreqBand[4–5] for CRY, FUS, and BAB.
Fig. 7.
Linear-frequency spectrogram overlaid with lspFreq[0–1].
Fig. 8.
Linear-frequency spectrogram overlaid with first formant frequency F1.
Table 1

Data sources, infant ages, models, features, target classes and best accuracy recorded in published infant vocalization experiments.

| Studies                | Data sources                                         | Infant ages | Models                   | Features                                      | Target classes            | Best accuracy |
|------------------------|------------------------------------------------------|-------------|--------------------------|-----------------------------------------------|---------------------------|---------------|
| Petroni et al. (1995)  | 230 crying episodes recorded in hospital             | 1–2 m       | FF, FT, RNN, TDNN, CC   | Mel-cepstrum Mel filterbank                   | ANGER, FEAR, PAIN         | 79.4%         |
| Warlaumont et al. (2010)| 20-min sessions of Infant/caregiver joint play       | 3, 6, 9 m   | SOM+ single layer perceptron | Spectrogram                                   | VOCANT, GROWL, SQUEAL    | 46.6%         |
| Yamamoto et al. (2013) | In-home recordings of one infant                     | 1.5 m       | ASR KNN                 | ASR+F0 PCA of FFT                            | CRY, NON-CRY:             | 67.4%         |
|                        |                                                      |             |                          |                                               | DIS/HUN/SLE:              | 62.1%         |
| Schuller et al. (2019) | BabySounds: 12445 infant vocalizations, in-home recordings | 2–36 m     | SVM                      | ComParE+BoAW+ auto-encoder based features    | BabySounds: CAN,NON-CAN, CRY, LAU, OTH | 58.7%         |
| Goxztolya (2019)       | BabySounds                                           | 2–36 m      | SVM                      | ComParE+ Fisher vector                        | BabySounds                | 59.5%         |
| Yeh et al. (2019)      | BabySounds                                           | 2–36 m      | SVM                      | ComParE+BoAW+ auto-encoder based              | BabySounds                | 62.39%        |
| Anders et al. (2020)   | Vocalizations from online sound libraries            | <9 m        | CNN                      | Log mel-scaled spectrograms                   | CRY, FUS, BAB, LAU, OTH   | 72.0%         |
| Ebrahimpour et al. (2020)| Infant/parent vocalizations, in-home recordings     | 3, 6, 9, 18 m| CNN +inception          | Audio waveform                               | INFANT/ADULT: VOC/NON-VOC: | 97.01%        |
|                        |                                                      |             |                          |                                               |                           | 90.49%        |
|                        |                                                      |             |                          |                                               |                           | 77.03%        |
|                        |                                                      |             |                          |                                               |                           | 60.28%        |
| Turan and Erzin (2018) | CRIED: 5587 infant vocalizations                     | 1–4 m       | CNN +Capsule             | Spectrogram                                  | NEUTRAL, FUS CRY         | 86.1%         |
| Ji et al. (2021a)      | Baby Chillanto (2267 vocalizations)/Baby2020 (5540 vocalizations) | 0–9 m/0–3 m | CNN+GCN                 | Spectrogram                                  | Chillanto: ASPHYXIA, DEAF,HUNGER, NORMAL,PAIN/2020,HUNGER SLEEPY,WAKEUP | 94.39%/74.37% |
| Maghfira et al. (2020) | Dunstan Baby Language: 315 crying sounds            | 0–3 m       | CNN-RNN                  | Spectrogram                                  | Dunstan: NEH,OW,H,EH EAR/H,HEH | 86.03%        |
| Ji et al. (2020)       | Dunstan Baby Language/Baby Chillanto                 | 0–3 m/0–9 m | Multistage CNN           | Hybrid feature                               | Dunstan/Chillanto         | 88.22%/95.10% |

Note: except as noted, each study uses a different dataset, therefore the percentage accuracies listed in the last column are generally not comparable across studies. Though not comparable, accuracies are reported here to give the reader a sense for the state of the art.
### Table 2

Number of families participating in the TDP and IDP recordings, grouped by infant’s age.

| Age           | TDP LENA | IDP LENA | IDP LAB | Total |
|---------------|----------|----------|---------|-------|
| 3 months      | –        | 7        | 79      | 86    |
| 6 months      | –        | 6        | 70      | 76    |
| 9 months      | –        | 5        | 53      | 58    |
| 12 months     | –        | 6        | 0       | 6     |
| 13 to 24 months | 10      | –        | –       | 10    |
Table 3

Descriptions of the adult facial expression codes used in this study.

| Class       | Descriptions                                                                                                                                 |
|-------------|----------------------------------------------------------------------------------------------------------------------------------------------|
| FLAT        | Neutral flat: displaying a very plain or uninterested facial expression, can be regardless of the vocalization or body movement. Could also be bored, tired, disengaged, uninterested, (almost like a depressed facial expression, uninterested, tired, do not want to be there). |
| INTERESTED | Neutral interested: looks attentively at infant. Bright eyes. There is a hint of positivity in some cases but with no visible smile. During free play, this is usually the “default” if nothing else is obvious. |
| SIMPLE SMILE| Mild/low positive: simple smile; bright; animated, on face.                                                                                   |
| BROAD SMILE | Strong/high positive: broad full smile (might be accompanied by laughter); exaggerated play face; mock face.                                   |
Table 4

Number of infant vocalizations of each class in the training and testing set for each fold of validation test.

| Number of fold | Dataset split | CRY  | FUS  | LAU  | BAB  | SCR  |
|----------------|---------------|------|------|------|------|------|
| 1              | Training      | 1145 | 4642 | 899  | 16 954 | 145  |
|                | Testing       | 621  | 2304 | 417  | 8493 | 58   |
| 2              | Training      | 1192 | 4612 | 859  | 16 991 | 131  |
|                | Testing       | 574  | 2334 | 457  | 8456 | 72   |
| 3              | Training      | 1195 | 4638 | 874  | 16 949 | 130  |
|                | Testing       | 571  | 2308 | 442  | 8498 | 73   |
### Table 5

Descriptions of infant vocalization codes.

| Class | Descriptions                                                                                                                                 |
|-------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| CRY   | Crying/screaming; crying is higher intensity than fussing and often includes gasps for air in between cries                                    |
| FUS   | Fussing/whining; lower intensity than crying; fussing sounds will often be less broken up and more extended than crying                    |
| LAU   | Laughter, chuckles, giggling                                                                                                                   |
| BAB   | Babbling/cooing; Babbling is non-intelligible speech, such as babababa, dada, oaaah, rrr, that typically involves a combination of consonant and vowel sounds. Cooing involves vowel sounds only. BAB can also include sounds that are only consonants (e.g., “mmmm”) as long as it is not fussy sound. Intelligible speech (e.g., mama, papa, etc.) are also annotated as BAB. |
| SCR   | High-pitched screeching                                                                                                                       |
Table 6
Number of adult female vocalizations of each class in the training and testing set for each fold of validation test.

| Number of fold | Dataset split | IDS  | ADS  | PLA  | RHY  | LAU  | WHI  |
|----------------|---------------|------|------|------|------|------|------|
| 1              | Training      | 4078 | 201  | 2253 | 533  | 1087 | 506  |
|                | Testing       | 2015 | 107  | 1148 | 262  | 541  | 256  |
| 2              | Training      | 3993 | 221  | 2255 | 531  | 1134 | 524  |
|                | Testing       | 2100 | 87   | 1146 | 264  | 494  | 238  |
| 3              | Training      | 4115 | 194  | 2294 | 526  | 1035 | 494  |
|                | Testing       | 1978 | 114  | 1107 | 269  | 593  | 268  |
### Table 7

Descriptions of the mother vocalizations.

| Class | Descriptions |
|-------|---------------|
| IDS | Infant-directed speech (or motherese), with much change of modulation and bigger jump in pitch. Elongated vowels. Could be delivered with a “cooing” pattern of intonation different from that of normal adult speech. With many glissando variations that are more pronounced than those of normal speech. |
| ADS | Adult-directed speech: speaking normally as if mother is talking to an adult (matter of fact) with “regular” rhythm and intonation. |
| PLA | Non-speech playful noises, including “raspberries,” kissing sounds, and animal sounds that occur alone (i.e., not as a part of a sentence); includes using a word to represent a playful noise (e.g., “bounce-bounce-bounce”, “shake-shake-shake”). |
| RHY | Rhythmic sounds (e.g., singing, sing-songy or cartoon-like voice). Repeating similar sounds at regular intervals without changing the manner of speaking. |
| LAU | Showing positive emotion with laughter, chuckle or explosive vocal sound (e.g., high-pitch positive vocalization sounds that does not contain speech). |
| WHI | Whispering; speech is low and gaspy, including “shhhh.” |
Functionals used in all experiments.

| Name      | Description                                      | D   | 9   | C   |
|-----------|--------------------------------------------------|-----|-----|-----|
| max       | The maximum value of the contour                |     |     |     |
| min       | The minimum value of the contour                |     |     |     |
| maxPos    | The absolute temporal offset of the maximum value |     |     |     |
| minPos    | The absolute temporal offset of the minimum value |     |     |     |
| range     | max-min                                          |     |     |     |
| mean      | Arithmetic mean                                  |     |     |     |
| stddev    | Standard deviation                               |     |     |     |
| skewness  | 3rd order central moment                         |     |     |     |
| kurtosis  | 4th order central moment                         |     |     |     |
| quartile1 | The first quartile                               |     |     |     |
| quartile2 | The second quartile                              |     |     |     |
| quartile3 | The third quartile                               |     |     |     |
| iqr1–2    | The inter-quartile range: quartile2–quartile1   |     |     |     |
| iqr2–3    | The inter-quartile range: quartile3–quartile2   |     |     |     |
| iqr1–3    | The inter-quartile range: quartile3–quartile1   |     |     |     |
| percentile1.0 | Outlier-robust minimum value (1st percentile) |     |     |     |
| percentile99.0 | Outlier-robust maximum value (99th percentile) |     |     |     |
| pctlrange0–1 | percentile99.0-percentile1.0                  |     |     |     |
| upleveltime75 | The percentage of time the signal is above (0.75 × range+min) |     |     |     |
| upleveltime90 | The percentage of time the signal is above (0.90 × range+min) |     |     |     |
| linregc1  | The slope of a linear approximation of the contour |     |     |     |
| linregc2  | The offset of a linear approximation of the contour |     |     |     |
| linregerrA | The linear error (difference between the actual feature contour, as a function of time, and its linear approximation) |     |     |     |
| linregerrQ | The quadratic error (difference between the actual feature contour, as a function of time, and its quadratic polynomial approximation) |     |     |     |

Column 1: name; column 2: description. Columns 3–5 contain * if the functional was computed for, respectively, the default feature set (emobase2010.conf)(D), the IS09_emotion.conf portion of the complementary set (9), or the rest of the complementary feature set (C).
### Table 9

Number of infant vocalizations for each class from external corpora selected in three-fold cross validation tests of different models for default and default+complementary features.

| Model | Features       | Fold | CRY | FUS | LAU | BAB |
|-------|----------------|------|-----|-----|-----|-----|
| FCN   | Default        | 1    | 34  | 23  | 139 | 905 |
|       |                | 2    | 40  | 21  | 90  | 950 |
|       |                | 2    | 41  | 15  | 224 | 846 |
|       | Default+complementary | 1  | 23  | 22  | 149 | 1359 |
|       |                | 2    | 27  | 32  | 97  | 1394 |
|       |                | 3    | 17  | 26  | 180 | 1374 |
| CNSA  | Default        | 1    | 32  | 4   | 56  | 2441 |
|       |                | 2    | 10  | 4   | 56  | 2499 |
|       |                | 3    | 4   | 6   | 119 | 3309 |
|       | Default+complementary | 1 | 11  | 5   | 33  | 3418 |
|       |                | 2    | 9   | 25  | 33  | 3006 |
|       |                | 3    | 6   | 19  | 57  | 2735 |

FCN = Fully-connected network, CNSA = convolutional neural net with self-attention.
Table 10

Number of adult female vocalizations of each emotion class (NEUTRAL vs. HAPPY) in the training and testing set for each fold of validation test.

| Number of fold | Dataset split | NEUTRAL | HAPPY |
|----------------|---------------|---------|-------|
| 1              | Training      | 4188    | 3714  |
|                | Testing       | 1881    | 2382  |
| 2              | Training      | 3772    | 4753  |
|                | Testing       | 1823    | 2440  |
| 3              | Training      | 3704    | 4822  |
|                | Testing       | 1891    | 2371  |
Table 11
Vocalization type classification scores for infant and mother vocalizations trained on the default feature set using the CNN and CNSA (CNN with self-attention).

| Model | Settings | Accuracy      | Weighted F1  | Macro F1     |
|-------|----------|---------------|--------------|--------------|
| CNN   | Infant   | 80.97% ± 0.07%| 79.24% ± 0.39%| 45.85% ± 1.69%|
|       | Mother   | 67.08% ± 0.56%| 64.78% ± 1.33%| 47.68% ± 3.86%|
| CNSA  | Infant   | 81.07% ± 0.13%| 79.69% ± 0.28%| 52.16% ± 0.78%|
|       | Mother   | 67.12% ± 0.53%| 65.93% ± 0.68%| 51.34% ± 0.70%|

Mean and standard deviations of accuracy, weighted F1 score, and macro F1 score for three-fold cross-validation tests are shown below.
### Table 12

**CNSA Hyper-parameters.**

| Name                              | Settings                       |
|-----------------------------------|--------------------------------|
| Input feature dimension           | $40 \times T$                  |
| CNN kernel sizes                  | $40 \times 8, 40 \times 16$    |
|                                  | $40 \times 32, \text{ and } 40 \times 64$ |
| Number of filters ($n_f$)         | 384                            |
| Max pooling kernel size           | 7                              |
| Max pooling stride                | 7                              |
| Attention hidden dimension ($n_h$)| 1024                           |
| Attention hops ($n_a$)            | 20                             |
| $H$ dimension                     | $T \times n_f$                 |
| $W_1$ dimension                   | $n_f \times n_0$               |
| $W_2$ dimension                   | $n_0 \times n_a$               |
| $E$ dimension                     | $n_f \times n_a$               |
| Dense layer hidden dimension      | 1024                           |
| Dropout                           | 0.2                            |

*T* is the number of rows of the feature matrix, thus for default features, $T = 40$; for default+complementary, $T = 47$; for raw Mel spectrogram, $T$ = the number of 10 ms audio frames in the segment.
Table 13
Optimization hyperparameters for FCN and CNSA training.

| Name           | Settings |
|----------------|----------|
| Number of epochs | 60       |
| Optimizer      | RMSprop  |
| Learning rate  | $10^{-4}$|
| Batch size     | 128      |
### Table 14

Vocalization type classification scores for infant vocalizations of CRIED dataset: accuracy, macro F1 score, and UAR for LOSO training/testing.

| Authors     | Model      | Features           | Settings | Accuracy | Macro F1 | UAR     |
|-------------|------------|--------------------|----------|----------|----------|---------|
| Ours        | LDA        | Default features   | U        | 80.17%   | 62.29%   | 64.27%  |
| Ours        | FCN        | Default features   | U        | 85.38%   | 68.16%   | 68.29%  |
| Ours        | CNSA       | Default features   | U        | 86.25%   | 68.77%   | 67.09%  |
|            |            | Default features   | W        | 81.22%   | 65.72%   | 72.44%  |
|            |            | Default features   | U+Fisher | 85.70%   | 68.28%   | 69.82%  |
|            |            | Default features   | W+Fisher | 80.83%   | 66.22%   | 74.17%  |
|            |            | Default+comp       | U        | 85.49%   | 67.56%   | 66.35%  |
|            |            | Default+comp       | W+Fisher | 81.78%   | 67.91%   | 76.15%  |
| COMPARE     | W+Fisher   |                     |          | 80.94%   | 66.63%   | 75.84%  |
| Ours        | CNSA       | Default features   | U        | 86.72%   | 68.63%   | 64.92%  |
| Ours        | CNSA       | Default features   | W        | 78.25%   | 65.28%   | 75.04%  |
| Schuller et al. (2018) | END2YOU | CNN-based          | CNN+LSTM | 70.8%    | –        | –       |
| Baseline    | OPENSMILE  | COMPARE            | SVM      | 82.6%    | –        | 75.6%   |
|            | OPENXBOW   | COMPARE            | BoAW+SVM | 84.2%    | –        | 76.9%   |
|            | AUDEEP     | AE-based           | RNN+SVM  | 83.5%    | –        | 74.4%   |
| Turan and Erzin (2018) | CapsNet | Spectrogram        | CNN      | 86.1%    | –        | 71.6%   |
| Huckvale (2018) | Combined NN | COMPARE            | LSTM+dense | –        | –        | 68.72%  |

“U/W” indicates training with unweighted/weighted sampler respectively. “Fisher” implies top 1000 features are selected based on Fisher scores. Best result for each metric is bolded.
Table 15

Vocalization type classification scores for infant vocalizations in our TDP and IDP datasets.

| Model | Features | Settings   | Accuracy     | Weighted F1   | Macro F1   |
|-------|----------|------------|--------------|---------------|------------|
|      |          |            |              |               |            |
| LDA   | Default  | U          | 78.84% ± 0.19% | 78.20% ± 0.15% | 53.70% ± 0.31% |
|       | Default+comp | U         | 79.55% ± 0.29% | 78.61% ± 0.31% | 54.80% ± 0.60% |
| FCN   | Default  | U          | 81.17% ± 0.02% | 80.48% ± 0.004% | 55.69% ± 0.32% |
|       | Default  | U+augmented | 81.11% ± 0.12% | 80.36% ± 0.11% | 56.71% ± 0.52% |
|       | Default+comp | U       | 81.48% ± 0.11% | 80.70% ± 0.19% | 56.54% ± 0.53% |
|       | Default+comp | U+Fisher  | 81.52% ± 0.17% | 80.91% ± 0.05% | 58.04% ± 0.91% |
|       | Default+comp | W+Fisher  | 76.71% ± 1.39% | 77.95% ± 1.02% | 56.69% ± 0.28% |
|       | Default+comp | U+augmented | 81.38% ± 0.25% | 80.73% ± 0.21% | 56.56% ± 0.42% |
| CNSA  | Default  | U          | 81.07% ± 0.13% | 79.69% ± 0.28% | 52.16% ± 0.78% |
|       | Default  | U+augmented | 80.96% ± 0.04% | 79.23% ± 0.13% | 50.51% ± 1.43% |
|       | Default+comp | U       | 81.43% ± 0.13% | 80.22% ± 0.43% | 55.23% ± 0.74% |
|       | Default+comp | U+Fisher  | 80.74% ± 0.51% | 80.41% ± 0.38% | 56.87% ± 1.29% |
|       | Default+comp | W+Fisher  | 81.06% ± 0.27% | 80.40% ± 0.25% | 56.63% ± 0.85% |
|       | Default+comp | U+augmented | 81.54% ± 0.21% | 80.19% ± 0.31% | 55.09% ± 1.21% |
| Pretrained | U               | 81.41% ± 0.11% | 80.00% ± 0.40% | 54.95% ± 1.38% |

Mean and standard deviations of accuracy, weighted F1 score and macro F1 score for three-fold cross-validation tests are shown below. “U/W” indicates training with unweighted/weighted sampler respectively. “Fisher” implies top 1000 features are selected based on Fisher scores. Best result for each metric is bolded.
### Table 16

Statistical significance tests results for the mean and standard deviation of three validation sets of infant vocalizations trained under unweighted sampler and various settings are shown below.

| Model       | Feature set        | p-value          |
|-------------|--------------------|------------------|
| FCN         | Default vs. default+comp | $1.17 \times 10^{-1} \pm 0.14$ |
| LDA vs. FCN | Default+comp       | $1.14 \times 10^{-13} \pm 8.07 \times 10^{-14}$ |
| LDA vs. CNSA | Default+comp      | $8.59 \times 10^{-10} \pm 1.09 \times 10^{-9}$ |
| CNSA vs. FCN | Default+comp      | $5.31 \times 10^{-1} \pm 6.93 \times 10^{-1}$ |
Table 17

Overlap feature set. Features shown in each square were among the top 30 features for that square, according to at least three of the four feature selection algorithms.

|         | CRY                        | Functional Feature | FUS                        | Functional Feature | LAU                        | Functional Feature | BAB                        | Functional Feature | SCR                        | Functional Feature |
|---------|----------------------------|--------------------|----------------------------|--------------------|----------------------------|--------------------|----------------------------|--------------------|----------------------------|--------------------|
| CRY     | logMelFreqBand[4] quartile1 | logMelFreqBand[5] quartile1 | logMelFreqBand[5] quartile1 | logMelFreqBand[5] quartile1 |
|         | logMelFreqBand[5] quartile1 | logMelFreqBand[5] quartile1 | logMelFreqBand[5] quartile2 | signal energy              |                           |
|         | F0finEnv                   | logMelFreqBand[4] quartile1 | logMelFreqBand[5] quartile1 | logMelFreqBand[5] quartile2 |
|         |                             | F0final             | percentile99.0            | percentile99.0          |
|         |                             | F0final             | linregerrQ                | linregerrA              |
|         |                             | F0final             | linregerrA                | linregerrA              |
|         |                             | F0final             | stddev                     | percentile99.0          |
|         |                             | F0finEnv            | percentile99.0            | percentile99.0          |
|         |                             | mfcc[11] percentile99.0 | mfcc[12] percentile99.0   | pitch range              |
|         |                             | pitch max           | pitch                      | pitch percentile99.0    |
|         |                             | pitch range         | pitch percentile99.0      | pitch pctlrnge0–1       |
|         |                             | pitch quartile3     |                           |                           |
|         | logMelFreqBand[0] linregerrA | logMelFreqBand[0] iqr2–3 | logMelFreqBand[1] iqr1–3  | logMelFreqBand[1] iqr1–3 |
|         | logMelFreqBand[1] iqr2–3   | logMelFreqBand[0] linregerrA | logMelFreqBand[1] iqr1–3  | logMelFreqBand[1] iqr1–3 |
|         | logMelFreqBand[2] iqr1–3   | logMelFreqBand[1] linregerrA | logMelFreqBand[1] iqr2–3  | logMelFreqBand[1] iqr2–3 |
|         | logMelFreqBand[2] iqr2–3   | logMelFreqBand[2] linregerrA | logMelFreqBand[2] iqr2–3  | logMelFreqBand[2] iqr2–3 |
|         |                             | F0final             | linregerrA                |                           |
|         |                             | F0final             | iqr 2–3                   |                           |
|         |                             | F0final             | iqr 1–3                   |                           |
|         |                             | F0final             | linregerrQ                |                           |
|         |                             | F0final             | linregerrA                |                           |
|         |                             | F0final             | stddev                     |                           |
|         |                             | F0final             | iqr 1–2                   |                           |
|         |                             | F0final             | iqr 2–3                   |                           |
|         |                             | logMelFreqBand[0] iqr2–3 | logMelFreqBand[1] iqr1–3  |                           |
|         |                             | logMelFreqBand[1] iqr1–3 | logMelFreqBand[1] iqr2–3  |                           |
|       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|
| CRY   | FUS   | LAU   | BAB   | SCR   |       |       |       |
|       | Feature | Functional |       | Feature | Functional |       | Feature | Functional |       |
| BAB   | pitch   | percentile99.0 | F0final | percentile99.0 | F0final | linregerrQ | – | – | F0final(d) | linregerrQ |
|       | mfcc[10] | percentile99.0 | F0final | linregerrQ | F0final | linregerrA | – | – | F0final(d) | linregerrQ |
|       |         |         | F0final | stddev | F0final(d) | linregerrA | – | – | – | – |
|       |         |         |         |         |         | linregerrA | – | – | – | – |
|       |         |         |         |         |         | linregerrQ | – | – | – | – |
|       |         |         |         |         |         | percentile99.0 | – | – | – | – |
|       |         |         |         |         |         | percentile99.0 | – | – | – | – |
|       |         |         |         |         |         | percentile99.0 | – | – | – | – |
|       |         |         |         |         |         | percentile99.0 | – | – | – | – |
|       |         |         |         |         |         | percentile99.0 | – | – | – | – |
| SCR   | lspFreq[0] | quartile2 | lspFreq[0] | quartile2 | lspFreq[0] | quartile2 | lspFreq[0] | quartile2 | – | – |
|       | lspFreq[0] | quartile3 | lspFreq[0] | quartile3 | lspFreq[0] | quartile3 | lspFreq[0] | quartile3 | – | – |
|       | lspFreq[0] | mean    | lspFreq[0] | mean    | lspFreq[0] | mean    | lspFreq[0] | mean    | – | – |
|       | lspFreq[1] | percentile99.0 | lspFreq[0] | percentile99.0 | lspFreq[0] | percentile99.0 | – | – | – | – |
|       | lspFreq[1] | quartile2 | lspFreq[0] | quartile1 | lspFreq[0] | quartile1 | lspFreq[0] | quartile1 | – | – |
|       | lspFreq[1] | quartile3 | lspFreq[0](d) | linregerrQ | lspFreq[1] | quartile1 | lspFreq[1] | quartile1 | – | – |
|       | lspFreq[1] | mean    | lspFreq[1] | quartile1 | lspFreq[1] | quartile2 | lspFreq[1] | quartile2 | – | – |
|       | F1     | quartile1 | lspFreq[1] | quartile2 | lspFreq[1] | quartile3 | lspFreq[1] | quartile3 | – | – |
|       | F1     | quartile2 | lspFreq[1] | quartile3 | lspFreq[1] | mean    | lspFreq[1] | mean    | – | – |
|       | mfcc[1] | quartile2 | lspFreq[1] | mean    | F1     | quartile1 | F1     | quartile1 | – | – |
|       |         |         |         |         |         |         |         |         | – | – |

A square represents overlap between the top 30 features selected for a one-vs-one classification problem (row-vs-column), and the row’s corresponding one-vs-all classification problem. For example, the first row shows overlapped features selected by CRY-vs-all and CRY-vs-FUS, CRY-vs-LAU, CRY-vs-BAB, and CRY-vs-SCR. Empty cell implies no overlapped features found.
| Model   | Setting          | Accuracy       | Weighted F1    | Macro F1     |
|---------|------------------|----------------|----------------|--------------|
|         |                  | 62.58% ± 1.08% | 62.20% ± 0.13% | 50.21% ± 0.74% |
| LDA     | Default          | 68.11% ± 0.45% | 67.13% ± 0.67% | 54.29% ± 0.58% |
|         | Augmented data   | 66.19% ± 0.42% | 66.20% ± 0.48% | 55.21% ± 0.38% |
| FCN     | Default          | 67.12% ± 0.53% | 65.93% ± 0.68% | 51.34% ± 0.70% |
|         | Augmented data   | 66.11% ± 0.55% | 65.11% ± 0.56% | 51.87% ± 0.32% |
| CNSA    | Default          | 67.06% ± 0.05% | 65.71% ± 0.07% | 51.04% ± 1.46% |

Mean and standard deviations of accuracy, weighted F1 score, and macro F1 score for three-fold cross-validation tests are shown below. Best result for each metric is bolded.
### Table 19

Significance test results for the mean and standard deviation of three validation sets, comparing accuracy of three classifiers of mother vocalizations.

| Model          | Feature set | $p$-value          |
|----------------|-------------|--------------------|
| LDA vs. FCN    | Default     | $7.40 \times 10^{-15} \pm 5.39 \times 10^{-15}$ |
| LDA vs. CNSA   | Default     | $8.08 \times 10^{-9} \pm 6.87 \times 10^{-9}$    |
| CNSA vs. FCN   | Default     | $1.04 \times 10^{-1} \pm 3.49 \times 10^{-2}$    |
Table 20

Emotion classification results for infant-directed speech data.

| Model | Setting                | Accuracy    | Weighted F1 | Macro F1   |
|-------|------------------------|-------------|-------------|------------|
| LDA   | Default                | 60.93% ± 0.23% | 60.82% ± 0.25% | 60.10% ± 0.22% |
| FCN   | Default                | 64.68% ± 0.06% | **63.52% ± 0.14%** | **62.31% ± 0.15%** |
| CNSA  | Default                | 64.16% ± 0.40% | 61.34% ± 1.43% | 59.64% ± 2.06% |
| CNSA  | Pretrained embeddings  | **65.10% ± 2.62%** | 62.53% ± 2.41% | 60.97% ± 2.83% |

Mean and standard deviations of accuracy, weighted F1 score, and macro F1 score for three-fold cross-validation tests are shown below. Best result is bold.
Table 21

Significance tests results or the mean and standard deviation of three validation sets, comparing three emotion classifiers for infant-directed speech.

| Model       | Feature set | p-value          |
|-------------|-------------|------------------|
| LDA vs. FCN | Default     | $1.76 \times 10^{-6} \pm 2.10 \times 10^{-6}$ |
| LDA vs. CNSA| Default     | $1.42 \times 10^{-3} \pm 1.99 \times 10^{-3}$ |
| CNSA vs. FCN| Default     | $2.46 \times 10^{-1} \pm 3.20 \times 10^{-1}$ |
Table 22

Accuracy, weighted F1 score and macro F1 score for three-fold cross validation tests on both training and testing data for 2-layer FCN trained with augmented data for mother vocalization.

| Fold | Accuracy | Weighted F1 | Macro F1 |
|------|----------|-------------|----------|
| Training |
| 1    | 84.37%   | 84.09%      | 81.39%   |
| 2    | 82.35%   | 81.95%      | 79.39%   |
| 3    | 85.81%   | 85.60%      | 83.34%   |
| Testing |
| 1    | 65.74%   | 65.87%      | 55.27%   |
| 2    | 66.76%   | 66.88%      | 55.64%   |
| 3    | 66.07%   | 65.85%      | 54.72%   |