Interspeech 2018
2-6 September 2018, Hyderabad, India

Infant emotional outbursts detection in infant-parent spoken interactions

Xu Yijia[1,2], *Hasegawa-Johnson Mark Allan*[1,2], *McElwain Nancy L*[2,3]

[1]Department of Electrical and Computer Engineering
[2]Beckman Institute for Advanced Science and Technology
[3]Department of Human Development and Family Studies
University of Illinois at Urbana Champaign, IL, USA

yijiaxu3@illinois.edu, jhasegaw@illinois.edu, mcelwn@illinois.edu

Abstract

Detection of infant emotional outbursts, such as crying, in large corpora of recorded infant speech, is essential to the study of dyadic social process, by which infants learn to identify and regulate their own emotions. Such large corpora now exist with the advent of LENA speech monitoring systems, but are not labeled for emotional outbursts. This paper reports on our efforts to manually code child utterances as being of type "laugh", "cry", "fuss", "babble" and "hiccup", and to develop algorithms capable of performing the same task automatically. Human labelers achieve much higher rates of inter-coder agreement for some of these categories than for others. Linear discriminant analysis (LDA) achieves better accuracy on tokens that have been coded by two human labelers than on tokens that have been coded by only one labeler, but the difference is not as much as we expected, suggesting that the acoustic and contextual features being used by human labelers are not yet available to the LDA. Convolutional neural network and hidden markov model achieve better accuracy than LDA, but worse F-score, because they over-weight the prior. Discounting the transition probability does not solve the problem.

Index Terms: infant vocalizations, infant emotional outbursts, convolutional neural network, linear discriminant analysis, hidden markov model

1. Introduction

We are interested in studying the dyadic social processes by which infants learn to express and regulate their own emotions. An infant may cry, fuss, laugh, babble or hiccup spontaneously, but she may also produce signals of this kind as part of a dialog, in which she seeks to evoke confirmation or comfort from a nearby adult caregiver. It is possible that some fraction of emotional outbursts are monologues (instinctive outbursts produced with no consideration of an intended audience), and some fraction are intended to be part of a dialog, and it is possible that these fractions change over developmental time scales.

In order to study the dynamic changes in intent, it is necessary to detect emotional outbursts (cry, fuss, laugh, hiccup and babble) in a very large corpus of recorded infant speech. Such large corpora do exist, but are not labeled with the level of detail we require. Previous work has focused on infant cry detection, or infant laugh detection, for applications like remote infant monitoring or purposes of infant clinical psychology [1], [2]. There is no such corpus or automatic detection algorithm for this task of detecting infant emotional outbursts.

This paper reports on our novel infant-parent spoken interaction corpus collected by the Language Environment Analysis (LENA) system, and our efforts to manually code child utterances as being of type "laugh", "cry", "fuss", "babble" and "hiccup", as well as to develop algorithms capable of performing the same task automatically.

When two human labelers independently annotate a set of child vocalizations, they achieve much higher rates of inter-coder agreement for some of the five categories than for others. These differences suggest the ambiguity of the sounds between the five categories as perceived by the human ear, and helps explain the possible errors in the machine classifier. We therefore explore 3-way, 4-way and 5-way classifiers, by eliminating the hiccup category, which does not have explicit implications for the child's emotional regulation, or combining the two classes that are easily confused with each other, i.e. fuss and cry, to eliminate the ambiguity.

In order to automate the annotating process of child utterances, we explore the linear discriminant analysis (LDA) classifier on selected prosodic and spectral features of child utterances, as well as the convolutional neural network (CNN) on filter bank features of child utterances, following with a hidden markov model (HMM) to learn the pattern of child utterance sequences.

Linear discriminant analysis achieves better accuracy on tokens that have been coded by two human labelers than on tokens that have been coded by only one labeler, suggesting that the acoustic and contextual features being used by human labelers are not yet available to the LDA. A 5-way LDA classifier achieves much higher accuracy on tokens that have been coded by two human labelers (69.33%) than tokens that have been coded by one one labeler (55.68%). Yet, a 3-way LDA classifier achieves similar accuracy on tokens coded by two labelers(73.89%) versus one labeler(72.73%).

Convolutional neural nets and hidden markov models achieve better accuracy than LDA, but worse F-score, apparently because they over-weight the prior. Discounting the transition probability does not solve the problem; no stream weight has been found that causes the HMM to produce an F-score better than LDA.

2. Infant-parent spoken interaction corpus

The participants in our corpus are drawn from a sample of fifteen families (9 female children; 6 male children). Children averaged 17.67 months of age (SD = 3.5 months; range = 13 to 24 months). Families are recruited via distribution of study fliers to local child care centers. Families are eligible to participate if parents are native English speakers, only English is spoken in the home, and children do not have any known hearing loss or difficulties.

Protocols for the participation of human subjects in this research were approved by the University of Illinois Institutional Review Board.

10.21437/Interspeech.2018-2429

The LENA system, developed on over 18,000 hours of naturalistic in-home recordings, has been validated for use with children between 2 and 48 months of age [3]. LENA includes a light-weight digital recorder that is securely placed into specially-designed child clothing and records the focal child's vocalizations as well as speech by family members for up to 16 hours, to capture a wide variety of parent-child interactions. Audio data from the digital recorder are processed in the lab by LENA software to automatically segment instances of the focal child's vocalizations, adult female speech, adult male speech, and other child speech. The sampling rate of the LENA recordings is 16kHz.

3. Annotation

Five labelers are asked to annotate each of five families' 16-hour LENA recordings. Each audio recording is automatically segmented by LENA system into instances of focal child's vocalizations. The labeler is asked to annotate each child's vocalization segments into one of the five categories: cry, fuss, laugh, babble and hiccup. "Fuss" is defined as whining or fussing that does not reach a full-blown cry. "Babble" is defined as non-intelligible speech that included consonant and/or vowel sounds (e.g. baba, dada, oaahh). "Hiccup" is a catch-all category that included reflexive sounds (e.g. hiccup, cough, yawn) or sounds that do not fall within one of the other categories. The labeler is also responsible for adjusting the LENA segmentation as needed, by deleting the incorrect segments, if the segment is not the target child vocalizing, or modifying the boundaries of the segment if the segment is either too long and contains other speakers or noises, or so short that child speech is cutoff.

A limited number of segments from two of the recordings are chosen to be annotated by two pairs of labelers. Each of the labelers in a pair annotates the same segments independently, and the result is used for the annotation reliability check.

Table 1 shows the annotations by labeler pairs, in terms of the count of the annotation classes.

Table 1: *Cross tabulation between annotations by labeler pairs*

	babble	cry	fuss	hiccup	laugh	total
babble	143	1	3	25	0	172
cry	0	92	21	0	0	113
fuss	20	38	115	69	2	244
hiccup	8	0	0	61	3	72
laugh	2	2	10	15	81	110
total	173	133	149	170	86	711

Table 1 shows that labelers achieve much higher rates of agreement for some of the categories than for others.

3.1. Balanced corpus

There are a total of 12768 child vocalization segments in five 16-hour LENA recordings, including 803 cry, 681 laugh, 2356 fuss, 1326 hiccup and 7602 babble annotations, which leads to a highly unbalanced corpus. In order to create a balanced training corpus with the same number of segments in each emotional outburst class, while maximizing the total number of segments, we keep the laugh class segments unchanged, which have the fewest examples in the corpus. We then randomly select the same number of examples from each of the other classes, to make up our balanced corpus, consisting of 3405 examples in total.

The smaller set of segments annotated by two labelers is also balanced among the five classes by using the same technique, resulting into 97 segments for each of the classes, and 485 segments in total.

We have explored the 5-way, 4-way and 3-way classifiers. A 4-way classifier is tested by eliminating the "hiccup" category, in order to focus only on the sound categories that have implications for the child's emotional expression and regulation. When we explore the 3-way classifier, we further combine the cry and fuss classes because they are easily confused with each other and overlap conceptually. While making these modifications, we keep the corpus balanced using the same technique described above.

4. Method

4.1. Linear discriminative analysis

4.1.1. Feature selection

We define 64 prosodic and spectral features to represent each child vocalization segment [4]. The open-source audio feature extractor, openSMILE [5], is used to extract the 64 spectral and prosodic features using a 30 ms Hamming window with 10 ms overlap, with the emobase configurations. Table 2 shows the features we extracted, and their statistical measurements or type of descriptors.

Table 2: *Spectral and prosodic acoustic features extracted using openSMILE*

Feature	Descriptors
previous vocalization class	class number
time duration of segments	duration
f0	slope, offset, mean, max, zero-crossing rate of log f0, inter-quartile difference
loudness	mean, max/min, inter-quartile difference
probability of voicing	probability
12 mel-frequency cepstral coefficients (a range from 0 to 8kHz)	mean, max/min, inter-quartile difference
signal zero-crossing rate	mean, max/min, inter-quartile difference

In order to maximize the power of features that are able to discriminate between different emotional outbursts, we apply feature selection algorithms to select the most discriminative features. Sequential forward selection (SFS), sequential backward selection (SBS), sequential floating forward selection (SFFS) and sequential floating backward selection (SFBS) [6], with LDA classifier measuring 5-fold accuracy upon balanced dataset, are used to select the features.

The subset of 23 features obtained from SFBS algorithm result in the highest 5-fold LDA accuracy on the balanced dataset. SFBS algorithm starts from the full set of features, and sequentially removes the feature that least reduces the value of objective function. After each backward step of removing the fea-

tures, SFBS performs forward steps by adding features from the set of features previously removed, as long as the objective function value increases.

Therefore, we define our prosodic and spectral features as these 23 features: previous vocalization class, time duration of segment, max value of f0, mean value of f0, slope of f0, zero-crossing rate of log f0, mean value of loudness, max value of loudness, probability of voicing, 4^{th} MFCC mean value, 7^{th} MFCC mean value, 11^{th} MFCC mean value, 3^{rd} MFCC min value, 7^{th} MFCC min value, 3^{rd} MFCC max value, 7^{th} MFCC max value, 1^{st} MFCC inter-quartile difference value, 6^{th} MFCC inter-quartile difference value, 7^{th} MFCC inter-quartile difference value, 9^{th} MFCC inter-quartile difference value, signal zero-crossing rate mean value, signal zero-crossing rate min value and signal zero-crossing rate max value.

Different feature selection algorithms produced completely different selected feature sets, but often with similar resulting classification accuracies. We speculate that the variability among selected feature sets may indicate that different features carry redundant information. If a feature selection algorithm selects one of the features in a redundant set, then it does not need to select any of the others; in this way it would be possible for different feature selection algorithms to select non-overlapping feature sets, yet achieve comparable accuracy.

4.1.2. Training and evaluation

We use 5-fold cross validation for the experiment. We randomly split the balanced corpus, consisting of 3405 child vocalization segments, into 5 folds, and consider each fold as test examples once and the rest of the 4 folds as training examples. An LDA classifier is applied to the 4-fold training examples each time, to generate a linear decision boundary. The LDA model fits a Gaussian density to each class, assuming that all classes share the same co-variance matrix. The fitted model is then used to predict the 1-fold test examples. For evaluation metrics, we measure the average accuracy and F-score values between the ground truth and predictions of vocalization segments across the five 1-fold test examples.

4.2. Convolutional neural network

4.2.1. Training

A child vocalization audio segment is divided into non-overlapping 500 ms frames. Each frame inherits all the labels of its parent audio. The 500 ms frames are decomposed with a short-time Fourier Transform applying 25 ms windows every 10 ms. The resulting spectrogram is integrated into 64 mel-spaced frequency bins, and the magnitude of each bin is log-transformed after adding a small offset to avoid numerical issues. This gives log-mel spectrogram patches of 50 x 64 bins that form the input to the convolutional neural network. During training, we fetch mini-batches of 16 input examples by randomly sampling from all patches.

4.2.2. Evaluation

We use the 5-fold cross validation to evaluate our detection task. We divide the balanced corpus, consisting of 3405 examples, into 5-folds randomly trained on 4-fold data, and tested on the rest 1-fold data. For our metrics, we calculate the averaged accuracy and F-score values across the five 1-fold test data.

In the evaluation process, each 500 ms frame from each child vocalization audio segment is passed into the model, and we average the classifier output scores across the frames in an audio segment.

4.2.3. Architecture

Because our balanced dataset is relatively small, we apply a shallow convolutional neural network to avoid the overfitting issue. The 50 x 64 filter bank frame is passed through a stack of convolutional layers, where we use filters with a receptive field of 3 x 3, to capture local spatio temporal patterns in the filterbank features. The convolutional stride is fixed to 1; the spatial padding is carried out by max-pooling after each convolutional layers, with kernel size of 4x4 and stride of 4. A stack of convolutional layers is followed by a fully-connected layer with 64 neurons; the final layer is the soft-max layer connected to the class labels. All hidden layers use RELU non-linearities [7],[8].

4.3. Hidden markov model

We believe that an infant is more likely to cry if her previous emotion state is cry, but less likely to cry if her previous emotion is laugh. Therefore, we propose to use HMM to capture this pattern of the vocalization sequences. An HMM has the ability to correct some of the predictions made by CNN model, by explicitly representing the higher probability of consistent label sequences [9],[10].

Because HMM works with sequential data in nature, we no longer use the balanced corpus consisting of randomly sampled child vocalization segments from LENA recordings. Instead, we split the five 16-hour LENA recordings into four training data sequences and one testing sequence. This division is repeated five times in a cross-validation sequence, so that each recording is a test recording once.

The transition probability of the HMM is obtained from the four manually labeled LENA recordings in the training fold, capturing the probability of transitioning from one category to another. The initial state probability is uniform.

The emission probability of each child vocalization observation given its emotion state label is obtained from the CNN outputs. The CNN model trained by 4-fold examples from the balanced corpus is applied to the 16-hour LENA testing recording, to get the emission probability.

We use the Viterbi algorithm to generate the most likely sequence of hidden emotion states, which is our classification prediction of the 16-hour testing LENA recording.

For our metrics, we use accuracy and F-score to measure the agreement between predicted labels and true labels for each segment in the testing LENA recordings, and average the results across the five LENA recordings.

5. Results

5.1. Linear discriminant analysis

Table 3 shows the averaged 5-fold cross validation F-score and accuracy results by the LDA classifier on the balanced corpus. Table 4 shows the averaged 5-fold cross validation F-score and accuracy results by LDA classifier on the smaller balanced set of waveform segments for annotations on which two labelers agreed.

The result shows that LDA achieves better accuracy on tokens that have been coded by two human labelers than on tokens that have been coded by only one labeler, suggesting that the acoustic and contextual features being used by human la-

Table 3: *Classification accuracy and F-score achieved by LDA classifiers on balanced LENA corpus*

	Accuracy	F-score
5-way classifier	55.68%	55.23%
4-way classifier	61.90%	61.27%
3-way classifier	72.73%	72.73%

Table 4: *Classification accuracy and F-score achieved by LDA classifiers on smaller balanced set of waveform segments with annotations agreed on by two labelers*

	Accuracy	F-score
5-way classifier	69.33%	69.23%
4-way classifier	75.42%	73.51%
3-way classifier	73.89%	73.40%

belers are not yet available to the LDA. The 5-way LDA classifier achieves much higher accuracy on tokens that have been coded by two human labelers(69.33%) than tokens that have been coded by one labeler(55.68%), whereas the 3-way LDA classifier achieves similar accuracy on tokens coded by two labelers(73.89%) versus one labeler(72.73%).

Focusing on vocalizations that are clear to the human ear (i.e., on which two labelers agree) helps to improve the accuracy of the 5-way classifier, but not of the 3-way classifier. It may be that this difference between the 5-way classifier and the 3-way classifier is the result of the much larger number of tokens omitted (due to labeler disagreement) for 5-way versus 3-way classification. The 5-way classification task is more difficult for human labelers, in the sense that there are a larger number of tokens on which labelers disagreed.

5.2. Convolutional neural network and hidden markov model

Table 5 shows the average 5-fold cross validation classification accuracy and F-score values achieved by the 4-way and 5-way CNN classifier on the balanced corpus. CNN on filterbank features result in worse classification accuracy and F-score than simple LDA on prosodic and spectral features.

Table 5: *Classification accuracy and F-score achieved by CNN classifiers on the balanced LENA corpus*

	Accuracy	F-score
5-way classifier	45.36%	43.95%
4-way classifier	51.59%	49.94%

In order to add the contextual information about the pattern of vocalization sequences into the CNN model, we explore the CNN-HMM by taking the CNN probability outputs as the emission probability for an HMM. Because we have five CNN models for 5-fold cross validation, we apply each of them as the emission probability model for an HMM on the 16-hour testing LENA recording, and averaged five accuracy and F-score values for each testing sequence. The reported general accuracy and F-score for classifiers with classifier cardinalities is measured as the average of five testing sequences. Table 6 and 7 show the classification accuracy and F-score values of the CNN model only and CNN-HMM model on all LENA recordings.

We explore different stream weights λ as multipliers of the emission log probability to weight the relative classification importance of the emission probability and transition probability. In this table, the setting $\lambda = 0$ is the setting in which the CNN is completely ignored; under this setting, the HMM simply generates the sequence with the highest *a priori* probability, which is the sequence that labels every segment as babble.

The results show that, for both the 4-way and 5-way classifier, CNN-HMM achieves higher accuracy than CNN and, indeed, approaches the accuracy of LDA, but achieves considerably worse F-score than LDA, because they over-weight the prior. However, discounting the transition probability by adjusting the λ value does not solve the problem. CNN-HMM F-score values are always worse than the LDA F-score values.

Table 6: *Classification accuracy and F-score achieved by 4 way CNN-HMM on LENA recordings*

	CNN	CNN-HMM				
λ		0	0.3	0.5	1	1.2
Accuracy (%)	59.31	65.06	69.75	67.43	64.20	63.44
F-score (%)	50.61	20.13	54.09	54.15	52.88	52.44

Table 7: *Classification accuracy and F-score achieved by 5-way CNN-HMM on LENA recordings*

	CNN	CNN-HMM				
λ		0	0.3	0.5	1	1.2
Accuracy (%)	52.82	58.30	63.55	61.24	57.82	57.13
F-score (%)	46.86	14.86	46.38	48.50	48.82	48.75

6. Conclusion

In this paper, we have reported on our infant-parent spoken interaction corpus with manual coding for infant emotional outbursts. We developed the algorithms, including LDA on prosodic and spectral features, as well as CNN-HMM on filterbank features, to automatically code the infant emotional outbursts. Human labelers achieve much higher rates of intercoder agreement for some of these categories than for others. Eliminating the "hiccup" category, which does not have explicit implications for the childs emotional regulation, or combining "fuss" and "cry" categories, which are easily confused with each other, helps to reduce the classification errors. LDA achieves better accuracy on tokens that have been coded by two human labelers than on tokens that have been coded by only one labeler, suggesting that the acoustic and contextual features being used by human labelers are not yet available to the LDA. CNN-HMM achieve better accuracy than LDA, but worse F-score, because they over-weight the prior.

7. Acknowledgements

We thank Macie Berg, Rachel Diechstetter, Emily Flammersfeld, Elizabeth Mooney, and Shreya Patel who manually annotating the LENA files. This study was supported by a seed grant from the Social and Behavioral Sciences Research Initiative at the University of Illinois at Urbana-Champaign.

8. References

[1] H. Rao, J. C. Kim, A. Rozga, and M. A. Clements, "Detection of laughter in childrens speech using spectral and prosodic acoustic features." *Proceedings of the 14th Annual Conference of the International Speech Communication Association, Interspeech*, pp. 1399–1403, 2013.

[2] Y. Lavner, R. Cohen, D. Ruinskiy, and H. Ijzerman, "Baby cry detection in domestic environment using deep learning," in *2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, Nov 2016, pp. 1–5.

[3] D. Xu, J. A. Richards, and J. Gilkerson, "Automated analysis of child phonetic production using naturalistic recordings," in *Journal of Speech, Language, and Hearing Research*, 2014, pp. 1638–1650.

[4] M. Pervaiz and T. A. Khan, "Emotion recognition from speech using prosodic and linguistic features," *IJACSA*, vol. 7, no. 8, 2016.

[5] F. Eyben, M. Wollmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," *Proceedings of the international conference on Multimedia*, pp. 1459–1462, 2010.

[6] S. Raschka, "Mlxtend," Apr. 2016. [Online]. Available: http://dx.doi.org/10.5281/zenodo.594432

[7] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. W. Wilson, "Cnn architectures for large-scale audio classification," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135, 2017.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[9] L. R. Rabiner, "Readings in speech recognition," A. Waibel and K.-F. Lee, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, ch. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pp. 267–296. [Online]. Available: http://dl.acm.org/citation.cfm?id=108235.108253

[10] Y.-L. Lin and G. Wei, "Speech emotion recognition based on hmm and svm," *Machine Learning and Cybernetics*, vol. 8, 2005.