Sensing with Contexts: Crying Reason Classification for Infant Care Center with Environmental Fusion

Chun-Min Chang*‡, Huan Yu Chen*‡, Hsiang-Chun Chen†, Chi-Chun Lee*‡
* Department of Electrical Engineering, National Tsing Hua University, Taiwan
E-mail: cmchang@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw
† Department of Early Childhood Education, National Tsing Hua University, Taiwan
E-mail: hcchen@mail.nd.nthu.edu.tw
‡ MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

***Abstract*—Crying is the only communication way from infants due to immature larynx and pharynx. It usually represents the urgent demand from infants. Caregivers and parents need to solve the urgent demand as soon as possible. Unfortunately, they are suffered from not knowing why infants cried and try many methods to comfort infants, in a result of that, there are a lot of material to teach parents how to understand infants' cry sound. However, it didn't work in the baby care center. The law in Taiwan allows caregiver taking care at most 5 infants in a time. If one of infants cried and caregiver could not solve immediately, chain problem will happen such as all infants crying simultaneously. Therefore, cry reason classification are important for the baby care center. Furthermore, event in the center may induce infants cry due to sensitivity to environment of infants. In this paper, we proposed a deep network that learns context information from the cry sound and leverages the event and environment factor to increase the performances of reason classification.**

I. Introduction

Since newborns are not capable of uttering semantically meaningful words due to immature larynx and pharynx, crying is the most common way to communicate with caregivers [1]. Infant cries for many reasons though those reasons usually represent specific urgent demands that caregivers or parents need to resolve. However, caregivers are not usually able to be around babies 24/7, which creates a growing demand of continuous infant monitoring solutions. Existing solutions are often camera based with *passive* monitoring capability, i.e., while the system is always on, the data is only available whenever caregivers have time to access it. In order to proactively react to baby's need, automatically infer the *need* and the *crying reasons*, beyond simply crying event detection, is a critical technological component to be developed. Specifically, recognizing the reasons of baby cries can be considered as a sophisticated acoustic event detection (AED) work when using audio signals.

Research in AED has been developed for many years. The core idea of AED is to understand and recognize the real world scenes by modeling the captured audio signals therein. Applications already exist in contexts such as health care [2], [3], smart home [4], [5], security surveillance [6], [7], [8]. The increasing availability of datasets on baby sounds, such as AUDIOSET [9], CRIED database [10], SPLANN database [11], has enabled development of AED algorithms for detecting baby cries/sounds reliably.

While most of past AED research works have worked on developing systems that would alert the caregivers when baby cries [12], [13], [14], this is often inadequate since the real difficulties faced by the parents/caregivers is the confusion about the reasons of baby cries instead of simply spotting baby cries. In fact, there exists many teaching materials attempting to characterize baby cries to help caregivers "understand" the meaning these infant sounds. For example, Dunstan Baby Language is a recent system developed that uses a few words for explaining the reasons of baby cries; specifically, they create five sounds, each with a distinct intent, that carry a consistent meaning by infants across cultures and linguistics groups prior to the language acquisition period. Several algorithms have been proposed to implement this language system in automatically characterizing infant sounds [15], [16].

Training automatic infant classification is a difficult issue because cry sound is nonverbal sound to define the real reason and build the connection from voice to crying reason and it is hard to access benchmark annotation infant cry database. Traditionally, researchers calculate few representations of spectrum such as MFCC, LPCC, spectrogram or low-level feature like pitch and intensity to analyze differences from different cry [1]. Rodriguez et al. used MFCC, LPCC to classify an infants psycho-physiological state such as hunger, pain or discomfort [17]. Bnic et al. use GMM-UBM framework to classify infant cry into 5 categories [18]. In recent years, researchers tend to solve this issue with network. Ji et al. use vanilla-DNN to classify asphyxiated infants or not [19]. Maghfira et al. classify infant cry into 5 categories using CRNN [16]. Turan et al. employed Capsule Network to classify emotional cry in domestic environments. It is proven that the reason information may be calculated in the cry sound.

There are a lot of voice on the baby care center. It relatively matters to derive the voice source with event prediction in baby care center then other public places due to the influence of environment of the care baby center. Nevertheless, the dilemma of baby care center is that understand-infants-crying-reason situation is more important than event sensing and recognition. In Taiwan, the law allows caregivers to take care at most 5 babies in a time for the baby care center;

Fig. 1: the scene of the baby care center

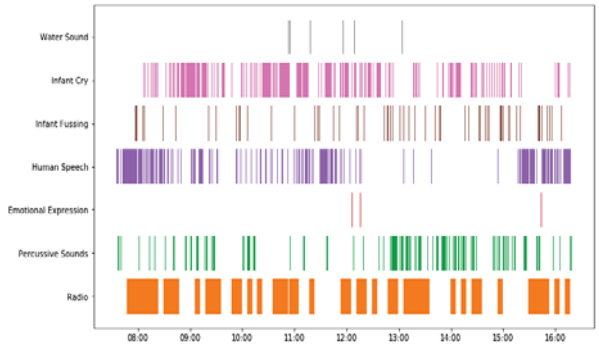

Fig. 2: Status of event for specific day in the baby care center. It shows routine annotation of baby care center in a day from around 8 a.m. to 4 p.m.

there will be 6-15 babies to take care in a room of baby care center. Once more than one infants cry, caregivers need to be aware of why infants cry to prevent affecting other babies as soon as possible. Vallotton et al. have studied how infants are influenced by their quality of care. The adequate and appropriate reaction of caregivers can influence babies' development robustly [20]. Therefore, cry reason classifier are necessary; it helps interaction from caregivers to babies and tends to reduce the stress of caregivers.

The infants can sense the context and start to cry if something happened. The sensitivity of infants' hearing is more than we thought. Furthermore, the environment of baby care center is more complicated than that of home. Infants in the baby care center may notice more various voice than home infants. In this paper, we proposed a structure which merge information of the environments. Audion information are gathered as a context box which contain audio information lasting a series of time. The environment annotation which are used for additional information are divided into two parts. The first is inter-environment information which is the annotation of event per second. The second is intra-environment information which finds the relationship of specific timing respectively by multiplication each other. After that, we proposed an algorithm taking those information to be trained a model to predict the reason of infants' crying.

Rest of this paper are as follows, Section 2 introduce database, pre-processing, and proposed method, Section 3 is the experimental setup, results and analyse, and Section 4 is conclusion.

TABLE I: *Summary of numbers of crying data we use*

Baby	M1	M2	M3	F1	F2	F3	F4
Love	251	147	77	217	530	521	2214
Sleepy	158	124	196	256	202	123	1002
Hungry	23	28	332	100	0	113	1439
Angry	630	78	0	79	39	73	0
Tot.	1062	377	605	652	771	830	4655

II. Research Methodology

A. the Baby Care Center database

The Baby Care Center database is a newly collectedly database with the annotation of baby cry reason and event. We collaborate with a baby care center in Taiwan. It is set with two cameras in different side of the room on a babysitting room in the center. Figure 1 shows the real scene of that room and workplace for crying acquisition in the center. The database consists of 7 babies (3 male and 4 female) whose age locates between on 2 months and 6 months and 3 caregivers to take care of those babies. Recording starts on the first baby taking place in the room in the morning and continue shooting until all the babies are picked by their parents. The recording almost lasts 8 hours per day. Each recording includes two kinds of annotations. One is the event annotations. There are 7 different kinds of label, water sound, human speech, human emotional expression, percussive sound, infant cry, infant fussing sound, radio sound. Another is cry reason annotations. There are 6 people majored in early childhood education responsible for labeling the cry reason. The cry reason are defined strictly into 5 class, diaper, love, hungry, angry, sleepy. Each label annotation is at least 2 people labeling and confirm with kappa-test which are more than 70%. In this experiment, we utilize 4 days and take out data that only include cry reason annotations for 4 different label, love, sleepy, hungry, angry. Diaper is too less to train. All in all, there is total of 8952 samples (Love: 3957, Sleepy: 2061, Hungry: 2035, Angry: 899) with environment information.

B. Audio Feature representations

The audio slices which contain cry reason label are used due to too less data with annotation of cry reason in four days, almost 32 hours. Each slice involves a whole 20 second audio data as a context box which denoted as X. Spectrogram representation are extracted from the input of these context box. It is crucial to visualizing the time-frequency energy on the feature because differences of distinguish crying sound often are presented at the energy of each frequency band. Frequency energy in spectrogram is shown in the colormap in which different color shows different intensity of energy.

Every audio data are divided into 20 second clip which processed with 10 ms hop length, windows of 250 ms, 40 mel-frequency filter becoming 40×101 using Librosa library. In total, there are $20 \times 101 \times 40$ dimension features for each context box.

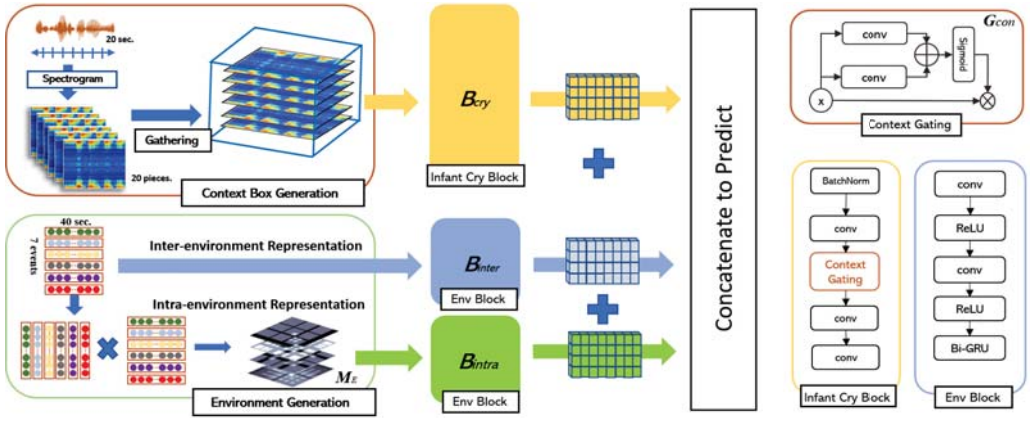

Fig. 3: Figure shows our proposed framework we proposed. Context box generate the acoustic feature with spectrogram with 20 seconds as well as environmental information are presented with inter-environmental vector and intra-environmental vector, then fusion with simple concatenate to predict the result.

C. Environment representations

Environment representations are divided into two set, inter-environment representations and intra-environment representations. Infants' cry sound are unique; event occurrence can be considered as a factor of cry reason because it is possible that some cry sound are triggered by the similar pattern event occurrence. Connecting alternative moment for event occurrence as an additional information make proposed framework able to grasp the information when cry happens if there is similar situation or not. Environment vector as inter-environment representations are defined as E_T^C, C denoted as event binary labels, containing $1, 2, c, ..., 7$ within T denoted as time series, containing $0, ..., t, ..., 40$ for every context box, so It will be 7×40. Then connecting alternative moment for event occurrence, E_T^C will be transposed as $E_T'^C$. Then we multiply both get an environment matrix, M_E which is defined as :

$$M_E^{KH} = \begin{bmatrix} E_1'^1 & E_2'^1 & E_3'^1 & \dots & E_7'^1 \\ E_1'^2 & E_2'^2 & E_3'^2 & \dots & E_7'^2 \\ \vdots & & \vdots & \vdots & \vdots \\ E_1'^T & E_2'^T & E_3'^T & \dots & E_7'^T \end{bmatrix} \begin{bmatrix} E_1^T \\ \vdots \\ E_3^t \\ \vdots \\ E_7^T \end{bmatrix}$$

Environment matrix represent that how event on k^{th} similar to the event on h^{th}. K and H are the same length as T.

D. Architecture of the network

Our proposed method consists of three part, infant cry block, inter-env block, intra-env block.

1) Infant Cry Block: Infant cry block are based on the paper [21]. Yan et al. employed typical context-gating module to parse the data information [22]. The module follows this equation:

$$Y = \sigma(\mathcal{N}_1(X) + \mathcal{N}_2(X)) \odot X$$

where X is the input data, sigmoid activation function denoted as σ. $\mathcal{N}$ are two similar network with different learned weighted, vanilla CNN, we used in this paper. $\odot$ means element-wise multiplication. Besides, we add 2 more CNN with ReLu after it and 1 more CNN with ReLU before it.

2) Env Block: The only differece of these two block,inter-env block and intra-env block, are the input form from Session 2.2. We employed CRNN to become major net on blocks. The module follows this equation:

$$z_1 = ReLU(\mathcal{N}_1(X))$$
$$z_2 = ReLU(\mathcal{N}_2(z_1))$$
$$z_3 = \mathcal{N}_{gru}(z_2)$$

where z_1,z_2,z_3 are the output of CNN $\mathcal{N}_1$, CNN $\mathcal{N}_2$, bi-GRU $\mathcal{N}_{GRU}$. X is the input data.

III. Experimental Setup and Results

A. Experimental setup

In this paper, we use imbalanced learning library on python due to bias dataset. Learning rate is set as 0.001. Epoch is 40 and batch size is defined as 64. Leave one day out is employed for cross validation. Unweighted average recall is used for calculation of model performance.

We compare our model with the following architecture in this paper:

- CNN: it is the most common method for the proposed cry reason method in recent years no matter for pathology or physical [19].
- CRNN: it is taking more information of audio to calculate and predict cry reason than ANN method which is less complicated [16].
- Capsule Net: Capsule Net have been proposed by Hinton et al. as a method for learning robust unsupervised representation of images [23]. It is proven being useful on cry reason classification on the CRIED database which is the cry reason database containing cry reason such as

TABLE II: *Summary of the experiments. Results are presented as UAR(%)*

	Proposed method				Caps net.				ANN	
	proposed.	w/ B_{inter}	w/ B_{intra}	B_{cry}	w/ both	w/ B_{inter}	w/ B_{intra}	w/o env	CRNN	CNN
Love	63.1	47.7	46.4	31.8	41.7	39.0	38.0	35.7	27.0	46.4
Sleepy	34.2	25.7	29.7	38.6	18.6	16.4	16.2	15.7	24.1	15.6
Hungry	30.8	41.9	47.2	39.0	37.4	37.5	38.5	37.7	43.5	26.9
Angry	58.3	45.6	34.8	41.9	42.4	42.5	37.8	44.1	47.6	12.0
UAR	46.6	40.2	39.5	37.8	35.0	33.8	32.6	33.3	35.5	25.2

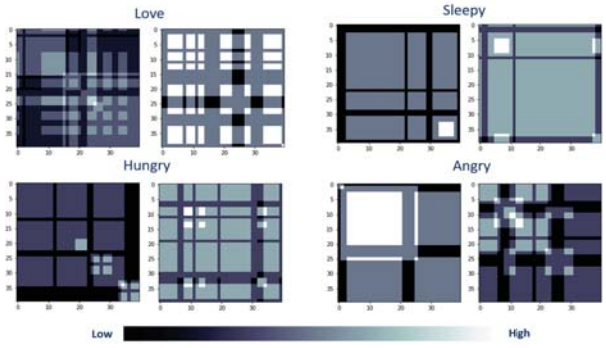

Fig. 4: The visualization of intra-environment factor of 4 reason respectively. There are 40 $\times$ 40 with value ranging from 0 (darker color) to 7 (lighter color). High value implies high correlation between times.

neutral/positive mood vocalisations, fussing vocalisations and crying vocalisation [24].

- B_{cry}: Proposed method only using context box information

Furthermore, performing of the environment factor should be checked. We further divide our proposed model into 4 situation with or without environment representations to compare the performing of each architecture.

- $B_{cry} + B_{intra}$
- $B_{cry} + B_{inter}$
- $B_{cry} + B_{intra} + B_{inter}$
- Caps net $+B_{intra}$
- Caps net $+B_{inter}$
- Caps net $+B_{intra} + B_{inter}$

B. Results

Table II lists the summary of our experiment results. There are several interesting observations from table. Our proposed method without environment information (B_{cry}) are better than the vanilla CNN, CRNN, and Capsule Net. Performances of it outperforms three of them by +12.3%, 2.3%, 4.5%. It implies the information of audio feature from proposed Net are digged out more than that of other models. It is worth noticing that performance from CRNN even outperforms Capsule Net by 2.2%. The mainly difference may be that input feature has been kept to next layer and provides original time-frequency relationship to the final vector, on the other hand, time-frequency relationship have been disappeared on the Capsule Net. We further observe that it is slightly increasing by adding additional environmental information for Capsule Net from 33.3% to 33.8% with additional inter-environmental factor(w/ B_{inter}), to 35% with additional both environmental factor(w/ both), and slightly decreased with additional intra-environmental factor (w/ B_{intra})from 33.3% to 32.6%. This result verify our inference. Capsule Net have intermingled the information from original input, in a result, additional intra-environmental factor instead confuse the net. On the contrary, performances gain better when our proposed method adding environmental information. It is achieved a boost of 2.4% and 1.7% by adding inter-environmental factor (w/ B_{inter})and intra-environmental factor (w/ B_{intra}) respectively. By applying both environmental information to the net, proposed. obtains 8.8% improvement from 37.8% to 46.6%. Then, we look details for those 4 reason performance. It is obvious to observe that proposed method gain better performances on Love and Angry, both of which increase the most by adding environmental information. These two reason are classified into psychogenic cry reason and the others are physical cry reason. It may be possible that event happened and it is usually annoying to infants. Annoyed is the possible reason that cause the psychogenic cry. We further examine the effect of environment factor.

Fig 4 shows the visualization of intra-environment factor of four reason respectively. Each graph contains values of 40 $\times$ 40. Each value ranges from 0 (darker color) to 7 (lighter color). High value implies high similarity between times. The graph are picked by the samples which is predicted rightly in Proposed Net with environmental factor but wrongly in Proposed Net.There are few interesting point by directly observation. Two distinct patterns exists above these graph. One is full of small block on the graph. The other is bigger block on the graph, and quantity are less than the first pattern. For hungry and sleepy, there are usually the bigger block on the graph. Hungry and sleepy are the cry reason that infants can not be satisfied by physical demand. It seems that environment or event are not the mainly concerned for the cry reason, therefore we can find the confusing of angry with two of them by comparing Table II and Fig 4. On the contrary, both environment graph for love filled with small and slight box is different from graph from sleepy and hungry obviously. slight box on the graph may imply that happening of event are highly concentrated and repeatedly so fast. This characteristic may

cause infants be panic to find love demand further. There are similar situation on graph for angry cry too. It is more likely the same phenomenon but not such stable like the graph of love. The long duration event may be also annoying to infants. These two observation may imply that environment of the baby care center may lead to the infants' cry in some aspect.

IV. Conclusions

In this work, we proposed a deep network that learns event of context information embedded into environment information and cry from infants to predict the demand and reason of infants. Specifically, the network leverages the event information to the audio context box to increase the performance of the results. Furthermore, environmental information are more sensitive to love and angry due to the characteristic of reason inducement.

There are few future directions. An immediate future work is to employ the event detection for this net. True annotation of event are used by this research at the present stage. If this system would be used by the baby care center, environment information should be derived automatically. Furthermore, the habit of each infants matters too. Each infants have different sensitivity. We then will employ temperament to make further studies [25].

Acknowledgment

The research is supported by the Ministry of Science and Technology Taiwan (MOST-109-2634-F-007-012-), NOVATEK Fellowship, and Taiwan Biotech Co., LTD..

References

[1] L. L. LaGasse, A. R. Neal, and B. M. Lester, "Assessment of infant cry: acoustic cry analysis and parental perception," *Mental retardation and developmental disabilities research reviews*, vol. 11, no. 1, pp. 83–93, 2005.

[2] K. Lee, A. Qadir, S. A. Hasan, V. Datla, A. Prakash, J. Liu, and O. Farri, "Adverse drug event detection in tweets with semi-supervised convolutional neural networks," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 705–714.

[3] S. N. Musy, D. Ausserhofer, R. Schwendimann, H. U. Rothen, M.-M. Jeitziner, A. W. Rutjes, and M. Simon, "Trigger tool–based automated adverse event detection in electronic health records: systematic review," *Journal of medical Internet research*, vol. 20, no. 5, p. e198, 2018.

[4] H.-Y. Lin, Y.-L. Hsueh, and W.-N. Lie, "Abnormal event detection using microsoft kinect in a smart home," in *2016 International Computer Symposium (ICS)*. IEEE, 2016, pp. 285–289.

[5] R. M. Alsina-Pagès, J. Navarro, F. Alías, and M. Hervás, "homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring," *Sensors*, vol. 17, no. 4, p. 854, 2017.

[6] S. Helmer and F. Persia, "High-level surveillance event detection using an interval-based query language," in *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*. IEEE, 2016, pp. 39–46.

[7] J. Sun, J. Shao, and C. He, "Abnormal event detection for video surveillance using deep one-class learning," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3633–3647, 2019.

[8] R. Yu, H. Wang, and L. S. Davis, "Remotenet: Efficient relevant motion event detection for large-scale home surveillance videos," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1642–1651.

[9] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[10] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny *et al.*, "The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats." in *Interspeech*, 2018, pp. 122–126.

[11] R. I. Tuduce, M. S. Rusu, C. Horia, and C. Burileanu, "Automated baby cry classification on a hospital-acquired baby cry database," in *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2019, pp. 343–346.

[12] P. K. Atrey, N. C. Maddage, and M. S. Kankanhalli, "Audio based event detection for multimedia surveillance," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 5. IEEE, 2006, pp. V–V.

[13] H. Phan, M. Krawczyk-Becker, T. Gerkmann, and A. Mertins, "Dnn and cnn with weighted and multi-task loss functions for audio event detection," *arXiv preprint arXiv:1708.03211*, 2017.

[14] 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)*. IEEE, 2016, pp. 1–5.

[15] E. Franti, I. Ispas, and M. Dascalu, "Testing the universal baby language hypothesis-automatic infant speech recognition with cnns," in *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2018, pp. 1–4.

[16] T. N. Maghfira, T. Basaruddin, and A. Krisnadhi, "Infant cry classification using cnn–rnn," in *Journal of Physics: Conference Series*, vol. 1528, no. 1. IOP Publishing, 2020, p. 012019.

[17] R. L. Rodriguez and S. S. Caluya, "Waah: Infants cry classification of physiological state based on audio features," in *2017 International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT)*, 2017, pp. 7–10.

[18] I.-A. Bănică, H. Cucu, A. Buzo, D. Burileanu, and C. Burileanu, "Automatic methods for infant cry classification," in *2016 International Conference on Communications (COMM)*. IEEE, 2016, pp. 51–54.

[19] C. Ji, X. Xiao, S. Basodi, and Y. Pan, "Deep learning for asphyxiated infant cry classification based on acoustic features and weighted prosodic features," in *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, 2019, pp. 1233–1240.

[20] C. D. Vallotton, "Do infants influence their quality of care? infants communicative gestures predict caregivers responsiveness," *Infant Behavior and Development*, vol. 32, no. 4, pp. 351–365, 2009.

[21] J. Yan and Y. Song, "Weakly labeled sound event detection with resdual crnn using semi-supervised method."

[22] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," *arXiv preprint arXiv:1706.06905*, 2017.

[23] J. T. Geiger and K. Helwani, "Improving event detection for audio surveillance using gabor filterbank features," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 714–718.

[24] M. A. T. Turan and E. Erzin, "Monitoring infant's emotional cry in domestic environments using the capsule network architecture." in *Interspeech*, 2018, pp. 132–136.

[25] S. B. Crockenberg, "Are temperamental differences in babies associated with predictable differences in care giving?" *New directions for child and adolescent development*, vol. 1986, no. 31, pp. 53–73, 1986.