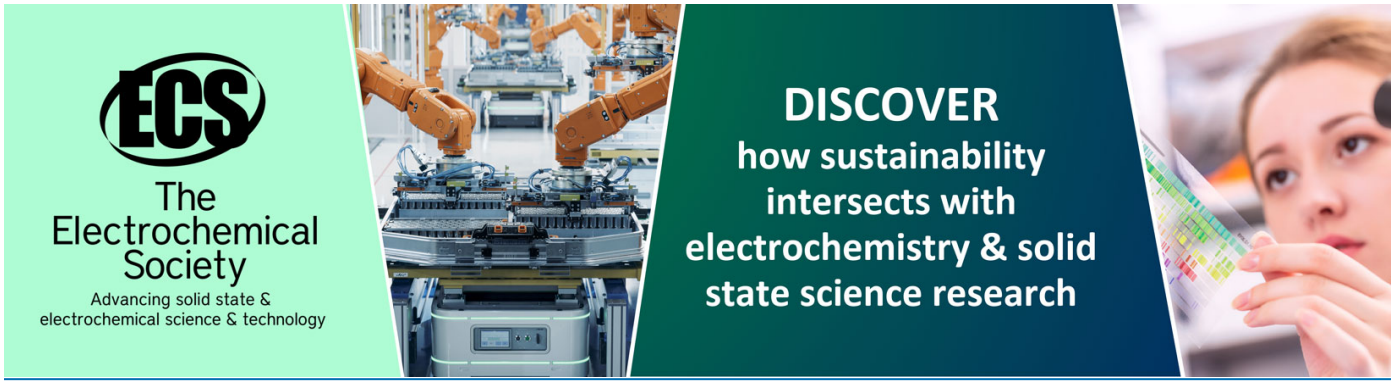
Infant cry classification using CNN – RNN



Journal of Physics: Conference Series **1528** (2020) 012019 doi:10.1088/1742-6596/1528/1/012019

Infant cry classification using CNN – RNN

Tusty Nadia Maghfira, T. Basaruddin* and Adila Krisnadhi

Faculty of Computer Science, Universitas Indonesia, Indonesia

*chan@cs.ui.ac.id

Abstract. The study of infant cry recognition aims to identify what an infant needs through her cry. Different crying sound can give a clue to caregivers about how to response to the infant's needs. Appropriate responses on infant cry may influence emotional, behavioral, and relational development of infant while growing up. From a pattern recognition perspective, recognizing particular needs or emotions from an infant cry is much more difficult than recognizing emotions from an adult's speech because infant cry usually does not contain verbal information. In this paper, we study the problem of classifying five different types emotion or needs expressed by infant cry, namely hunger, sleepiness, discomfort, stomachache, and indications that the infant wants to burp. We propose a novel approach using a combination of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) that acts as feature extraction and classifier method at once. Particularly, CNN learns salient features from raw spectrogram information and RNN learns temporal information of CNN obtained features. We also apply 5-folds cross-validation on 200 training data set and 50 validation data set. The model with the best weight is tested on 65 test set. Evaluation in Dunstan Baby Language dataset shows that our CNN-RNN model outperforms the previous method by average classification accuracy up to 94.97%. The encouraging result demonstrates that the application of CNN-RNN and 5-folds cross-validation offers accurate and robust result.

1. Introduction

Crying is a nonverbal communication signal that is commonly shown by newborns since they still cannot speak a word. It can be considered as the natural behavior of an infant in building social interaction with their caregiver. Infant cry can be caused by many different reasons, which each type of them represents different state and urgent needs. Hopefully, by this signal, it can motivate parents or caregivers to alleviate distress and give affection, safety, and protection to their infants [1]–[3]. Adult's responsiveness to infant's cues may influence an infant's emotional development that acts as building blocks of their future learning [4]. However, parents and inexperienced caregivers mostly find it difficult to recognize what infants try to convey through their cry because it does not contain any verbal information, and it sounds irregularly compare to adults [5]. Developing an infant cry classification system can be addressed as a helpful solution towards early diagnosis of the psychological and physical condition of an infant, which leads to better childcare.

Research development on infant cry recognition tends to be slower compared to the study of speech emotion recognition of adult voices since it is hard to differentiate emotions from nonverbal sounds, and there is still no infant cry data set with benchmark annotation [6]. Previous studies mostly extract low-level features and train them by using machine learning algorithms [7]–[16]. However, there are some drawbacks of these approaches such as on their development, validation, and optimization of the results

may take too much time and leads to an increase in computing costs [17], [18]. At the same time, deep neural networks have been adapted into speech emotion recognition [17]–[21]. Regarding these studies, CNN and RNN are proven can be used to recognize paralinguistic information contained in the speech signal and obtain better results than conventional method approaches.

Motivated by the success of deep neural networks on Speech Emotion Recognition (SER), in this paper, we propose CNN-RNN for infant cry recognition, which work together sequentially to cover both feature extraction and classification steps (Section 2). CNN plays a role in extracting salient best emotion representation features from spectral properties of cry sounds. Then the obtained features are passed to RNN to be learned sequentially at each time step so that its values can be mapped to the associated target emotion. In order to overcome the data limitation problem and ensure the effectiveness of our model, we implement 5-folds cross-validation and test the model with the best weight to the test set. Experimental results in Section 3 show that our proposed model outperforms the CNN system [22] on classifying among five different types of universal infant cries.

2. Convolution Neural Network – Recurrent Neural Network (CNN-RNN)

CNN is one of the deep learning models that specially designed for analyzing visual imagery data. Common practical applications that suitably analyzed by CNN are images, videos, and time-frequency representation of audio. Essential features of data are extracted by the use of convolutional and pooling operations consecutively layer by layer. Every layer level increases, the spatial resolution of the layer is decreased, but we can obtain deeper abstract features. The main idea of CNN is to take advantage of local connectivity, weight sharing, pooling, and the use of many layers [20], [23]. In contrast to conventional ANN, the neurons in every layer are only connected to a small region of the previous layer known as the receptive field.

Learning the meaning behind infant cries needs analysis from the beginning to the end of the audio. Loss of information in the middle of the sound signal may affect in misclassification. Because feature properties in successive time steps can form the decision of the classification result for each cry signal. Therefore, RNN becomes the right and popular option for dealing with speech recognition problems because it can explore temporal information sequentially. Unlike other neural network methods that each input is independent to each other, RNN can make an accurate prediction based on previously shared information that stored and passed through memory cells.

In this study, we propose a combination of CNN and RNN as shown in Fig. 1. Once we get the time-frequency mapping from spectrogram of the audio signal, we pass it to CNN, which contains three consecutive convolutions with max-pooling following each of them. There are 16 filters on the first layer of convolution and 32 filters on the rest layer with the same size 3 x 3. Each of them, along with max-pooling layers, gradually convolves and extracts salient spatial features. Furthermore, the three-dimensional feature tensor output from the last layer of CNN is flattened into one-dimensional feature vector. Then it is reshaped into two-dimensional space (64 x 64) to adjust it as the input of RNN. Our proposed architecture includes two stacked RNN layers with 64 memory cells for each layer. The outputs of the last RNN layer are passed into a fully connected layer that classifies extracted spatial-temporal information into one of five possible classes.

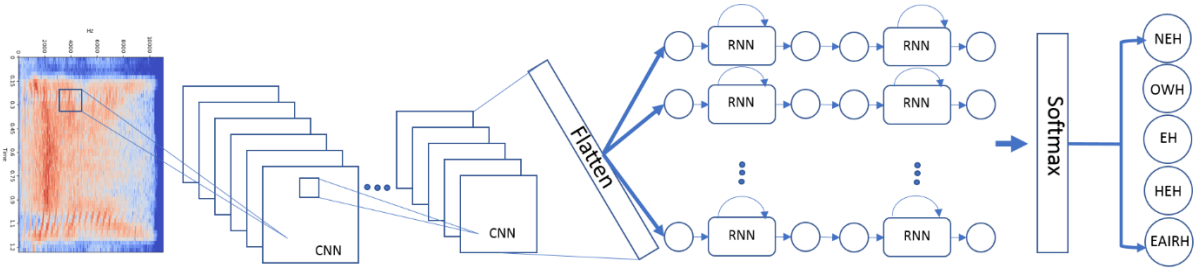


Figure 1. The Proposed CNN-RNN Model for Infant Cry Recognition

3. Experiments

3.1. Data

The data used for this study is obtained from Dunstan Baby Language that was discovered by Priscilla Dunstan and her research team in 2006. First, Priscilla tried to listen to and observe some sounds that his son, Tomas, produced before he cried. In order to validate their idea, she and her team done worldwide research and concluded a statement that these five words used universally by all infants in their first three months regardless of different language, culture, and race. The five universal words of Dunstan are: "neh" means hungry, "owh" means sleepy, "eh" means infant wants to burp, "eairh" means infant has stomach cramp because they cannot release the gas in stomach, and "heh" means infant feels uncomfortable internally.

Originally, Dunstan Baby Language is a video that available online and in DVD format. This 60-minutes Dunstan Baby Language video consists of several examples of 5 types of infants cries along with explanation for each of their characteristics and some tips about how to soothe babies cry based on their types. There are also completed with cut of dialogues and testimonies from parents who have learned infant's cry sound from it. All babies' cries sound along the video were recorded in studio conditions so that there will be no noises and resonances. As the purpose of this research, we only remove that irrelevant information and only keep babies' cries sounds. In order to obtain only the clear sound of babies' cry, the video file is cut and converted into some audio files with "wav" format, 16000 Hz sample rate, mono audio channel, and maximum duration-1 second length for each cry sound. The extracted infant's cry sound audio consists of 315 in total, each of which represents 56 "neh" sounds, 106 "owh" sounds, 55 "eh" sounds, 61 "heh" sounds, and 37 "eairh" sounds.

3.2. Experimental Setup

First of all, the experiment is done by preparing input data which is a spectrogram of each audio sound. We can obtain it by transforming audio signals into its time-frequency maps using Short Time Fourier Transform (STFT). We divide each audio file into 2048 FFT bins frame, and each of them is Fourier transformed over time. Every time instance, an input signal is multiplied with an analysis window that has the same size as FFT bins. How much we go through time is affected by hop size 512 bins. The result of each audio file will be a 1025 x 18 two-dimensional time-frequency array.

In this research, all the data will be divided into training, validation, and testing. Since the data is in small size and imbalanced for each class, experiments will be conducted using 5-fold cross-validation procedure. The purpose of its application is to evaluate the model on limited data so that we can obtain a model with the best representation. Testing data is taken with a percentage of 20% for each class with total of them are 65 files. We manually select a testing dataset with the hope that its variations are evenly distributed for each class. While validation and training data consecutively are 50 files and 200 files with random selection data distribution. Training set with the best weight from all folds will be used to train the test set.

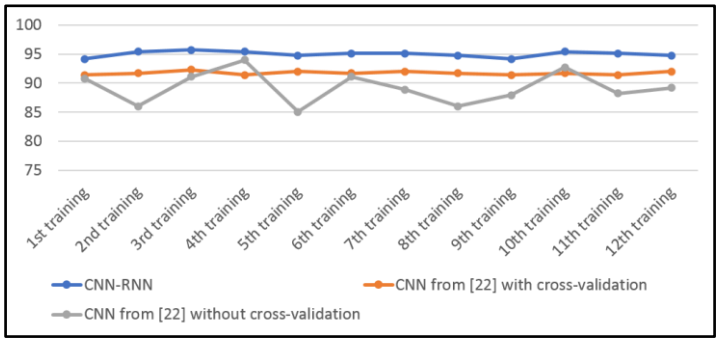
Implementation of the proposed method is done by using Python with the use of Keras library and TensorFlow backend. In order to evaluate the performance of our proposed method, we compare it with the result of [22]. The experiment is done with the same data and treatment. The loss of the model is measured with binary cross-entropy based on Franti et al. experiment. However, as additional analysis, we also try to evaluate our proposed method using categorical cross-entropy since infant cries classification is categorized as a multiclass problem. Table 1 shows the hyperparameters settings used in our experiments.

Table 1. The hyperparameters Settings of Proposed Method

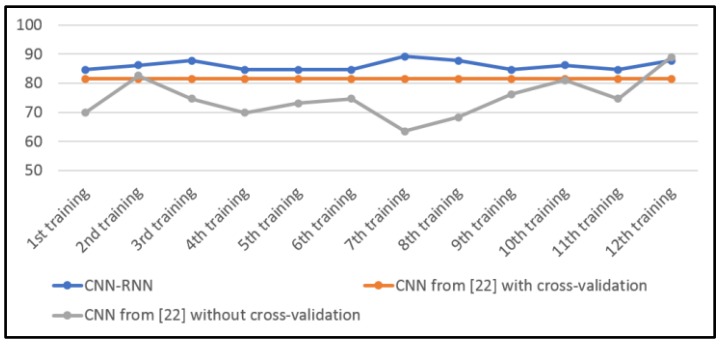
Parameter	Value	Parameter	Value
Activation function	Relu	**Number of epochs**	10
Optimizer	Adam	**Number of each experiments**	12
Dropout	0.2	**Batch size**	8

3.3. Experimental Results

All experiment results of our proposed method are listed in Fig. 2. Fig. 2(a) shows a comparative result of our proposed method with Franti et al. approach with and without cross-validation in binary cross-entropy loss function scheme. While Fig. 2(b) shows the same experiment result by using categorical cross-entropy. The average of all experimental results are shown in Table 2.



(a)



(b)

Figure 2. Comparison of Accuracy Result Between Proposed Method and [22] Works : (a). Binary Cross-Entropy Loss Function; (b). Categorical Cross-Entropy Loss Function

Based on all the results, we can assume that the proposed CNN-RNN structure gives better results than the previous method. We also can assume that experimental results with binary are better than categorical cross-entropy. Actually, the binary problem is a special case of multiclass that has two classes. The difference between those two is the probability of each class in categorical cross-entropy is dependent, there is only one hot positive class among all negative classes. While binary cross-entropy works on each individual output independently, and each case can belong to more than one class. Therefore, its probability score is obtained from the sum of all positive class scores. It can be applied to the multiclass problem, but it may cause the determination of ground truth to become ambiguous. This is the reason why binary cross-entropy gives a higher result than categorical though it is not a valid result. Both binary and categorical cross-entropy overall results show that our proposed method exceeds the performance of Franti et al. method. Moreover, model validation using 5-folds cross-validation can also improve the performance of our proposed method.

Table 2. Comparison of Average Accuracy Result Between Proposed Method and [22] Works

	Franti et al. without Cross-Validation	**Franti et al. with Cross-Validation**	**CNN – RNN**
Binary Cross Entropy	89.26	91.71	**94.97**
Categorical Cross Entropy	74.73	81.54	**86.03**

4. Related Work

Previous infant cry recognition approaches typically extract low-level features and let the machine learning method map them into some physical and emotion class. Among some recent researches, there is an experiment done by [14] that uses MFCC in extracting features and KNN for classification stage, and its accuracy is higher than Naïve Bayes, Neural Network, and SVM by 75.95%. Another work was proposed by [16], which compares some feature extraction methods, including LPC, LPCC, MFCC, and BFCC. They also proposed a compressed sensing method as a classification method with KNN and ANN as a comparison. The best classification rate is obtained from a combination of BFCC and ANN with 76.47%, while their proposed classification method gave the best result up to 71.05% when combined with MFCC. It is clear that a different combination of feature extraction and classification method can give different performance results. Moreover, approaches of machine learning algorithms on low-level descriptors' features still have not obtained optimal results.

In recent years, deep learning algorithms have become a highlight in several studies of computer vision and speech processing since they can produce high-level salient features from raw data and also act as a classifier for them. Among many different approaches, CNN is one of deep learning algorithms that is widely applied as it is known for its outstanding performance in image processing. Related to them, CNN starts to be an alternative over traditional approaches in SER by extracting representative features through spectrogram. For example, a study by [22] proposed a CNN-based network for infant pre-cry utterances recognition, and they achieved better accuracy results compared to low-level features based approaches. Besides CNN, the RNN method which is known to have excellent performance for sequential data such as speech signals, is also widely applied by some researchers. A few recent studies have developed experiments by combining CNN and RNN with end-to-end model architecture, and their performance exceeds plain CNN or RNN [18], [20], [21].

5. Conclusion

In this paper, we proposed a novel approach to infant cry classification which based on the combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) with cross-validation. We evaluate our proposed method by comparing its performance with the basic CNN-based method on Dunstan Baby Language. Experimental results show that our proposed method outperforms the previous method by accuracy up to 86.03% in categorical cross-entropy and 94.97% in binary cross-entropy. The results indicate spatial and temporal features obtained by CNN-RNN can efficiently extract the representation of the physical and emotional needs of infants. We also prove that the use of cross-validation can be the factor to improve method performance. In the future, we will continue our work by learning the joint representation of CNN-RNN and handcrafted features and collecting more credible infant cry data for a better generalization.

6. Acknowledgement

This study was financially supported by Hibah Riset Publikasi Internasional Terindeks Tugas Akhir (PITTA B) Universitas Indonesia 2019 (NKB-0517/UN2.R3.1/HKP.05.00/2019).

References

[1] J. Bowlby, *Attachment and Loss: Attachment Vol 1*, vol. I. 1969.
[2] S. M. Bell and M. D. S. Ainsworth, "Infant Crying and Maternal," vol. 43, no. 4, pp. 1171–1190, 1972.
[3] A. D. Murray, "Infant crying as an elicitor of parental behavior: An examination of two models," *Psychol. Bull.*, vol. 86, no. 1, pp. 191–215, 1979.
[4] R. Caulfield, "Social and emotional development in the first two years," *Early Child. Educ. J.*, vol. 24, no. 1, pp. 55–58, 1996.
[5] G. Lei, Y. Hongzhi, L. Yonghong, and M. Ning, "Pitch Analysis of Infant Crying," *Int. J. Digit. Content Technol. its Appl.*, vol. 7, no. 6, pp. 1072–1079, 2013.
[6] S. Jeyaraman, H. Muthusamy, W. Khairunizam, S. Jeyaraman, T. Nadarajaw, S. Yaacob, and

S. Nisha, "A review: survey on automatic infant cry analysis and classification," *Health Technol. (Berl).*, vol. 8, no. 5, pp. 391–404, 2018.
[7] Y. Abdulaziz and S. M. S. Ahmad, "An accurate infant cry classification system based on continuos hidden Markov model," *Proc. 2010 Int. Symp. Inf. Technol. - Syst. Dev. Appl. Knowl. Soc. ITSim'10*, vol. 3, pp. 1648–1652, 2010.
[8] M. Dewi Renanti, A. Buono, and W. Ananta Kusuma, "Infant cries identification by using codebook as feature matching, and MFCC as feature extraction," *J. Theor. Appl. Inf. Technol.*, vol. 56, no. 3, pp. 437–442, 2013.
[9] V. V Bhagatpatil and P. V. M. Sardar, "An Automatic Infant's Cry Detection Using Linear Frequency Cepstrum Coefficients (LFCC)," vol. 5, no. 12, pp. 1379–1383, 2014.
[10] S. Bano, "Decoding Baby Talk : A Novel Approach for Normal Infant Cry Signal Classification," *2015 Int. Conf. Soft-Computing Networks Secur.*, pp. 1–4, 2015.
[11] K. Srijiranon and N. Eiamkanitchat, "Application of neuro-fuzzy approaches to recognition and classification of infant cry," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2015-Janua, pp. 1–6, 2015.
[12] S. S. Jagtap, P. K. Kadbe, and P. N. Arotale, "System propose for Be acquainted with newborn cry emotion using linear frequency cepstral coefficient," *Int. Conf. Electr. Electron. Optim. Tech. ICEEOT 2016*, pp. 238–242, 2016.
[13] A. Chaiwachiragompol and N. Suwannata, "The Features Extraction of Infants Cries by Using Discrete Wavelet Transform Techniques," *Procedia Comput. Sci.*, vol. 86, pp. 285–288, Jan. 2016.
[14] W. S. Limantoro, C. Fatichah, and U. L. Yuhana, "Application development for recognizing type of infant's cry sound," *Proc. 2016 Int. Conf. Inf. Commun. Technol. Syst. ICTS 2016*, pp. 157–161, 2017.
[15] L. Liu, Y. Li, and K. Kuo, "Infant cry signal detection, pattern extraction and recognition," *2018 Int. Conf. Inf. Comput. Technol. ICICT 2018*, no. 2, pp. 159–163, 2018.
[16] L. Liu, W. Li, X. Wu, and B. X. Zhou, "Infant cry language analysis and recognition: An experimental approach," *IEEE/CAA J. Autom. Sin.*, vol. 6, no. 3, pp. 778–788, 2019.
[17] M. Papakostas, E. Spyrou, T. Giannakopoulos, G. Siantikos, D. Sgouropoulos, P. Mylonas, and F. Makedon, "Deep visual attributes vs. hand-crafted audio features on Multidomain Speech Emotion recognition," *Computation*, vol. 5, no. 2, pp. 1–15, 2017.
[18] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," *2016 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA 2016*, 2017.
[19] C. W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," *Proc. - IEEE Int. Conf. Multimed. Expo*, no. July, pp. 583–588, 2017.
[20] Y. Mu, L. A. Hernández Gómez, A. C. Montes, C. A. Martínez, X. Wang, and H. Gao, "Speech Emotion Recognition Using Convolutional- Recurrent Neural Networks with Attention Model," *DEStech Trans. Comput. Sci. Eng.*, no. cii, pp. 341–350, 2017.
[21] D. Luo, Y. Zou, and D. Huang, "Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition," no. September, pp. 152–156, 2018.
[22] E. Franti, I. Ispas, and M. Dascalu, "Testing the Universal Baby Language Hypothesis - Automatic Infant Speech Recognition with CNNs," *2018 41st Int. Conf. Telecommun. Signal Process. TSP 2018*, pp. 1–4, 2018.
[23] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.