딥 트랜스퍼 러닝 기반의 아기 울음소리 식별

박철*, 이종욱**†, 오스만*, 박대희**, 정용화**
*고려대학교 컴퓨터정보학과, **고려대학교 컴퓨터융합소프트웨어학과
e-mail: ku_bozhao@korea.ac.kr

Infant cry recognition using a deep transfer learning method

Zhao Bo*, Jonguk Lee**†, Othmane Atif*, Daihee Park**, Yongwha Chung**
*Dept. of Computer Information Science, Korea University
**Dept. of Computer Convergence Software, Korea University

Abstract

Infants express their physical and emotional needs to the outside world mainly through crying. However, most of parents find it challenging to understand the reason behind their babies' cries. Failure to correctly understand the cause of a baby' cry and take appropriate actions can affect the cognitive and motor development of newborns undergoing rapid brain development. In this paper, we propose an infant cry recognition system based on deep transfer learning to help parents identify crying babies' needs the same way a specialist would. The proposed system works by transforming the waveform of the cry signal into log-mel spectrogram, then uses the VGGish model pre-trained on AudioSet to extract a 128-dimensional feature vector from the spectrogram. Finally, a softmax function is used to classify the extracted feature vector and recognize the corresponding type of cry. The experimental results show that our method achieves a good performance exceeding 0.96 in precision and recall, and f1-score.

1. INTRODUCTION

Crying is the main way of communication that infants rely on to express their current states and needs to the outside world. Babies cry for different reasons such as hunger, sleepiness, pain, etc. While trained professionals, like maternity matrons and pediatric nurses, can understand the physical and psychological state of a baby from his cry, for parents who lack experience, distinguishing a baby's different cries remains a big challenge [1]. If parents cannot correctly understand the cause of their baby's cry and take appropriate actions, this might affect their cognitive and motor development, especially when they are still in a stage of rapid brain development [2]. Hence, it is necessary to build a system that helps parents understand the meaning of their babies' cries, making sure it is harmless to the babies. In this study, we propose an infants' cry recognition system that leverages the hidden acoustic information behind babies' crying sounds to help inexperienced parents recognize the psychological and physiological state of their babies. This system is non-invasive and harmless as it only relies on sound data.

Since the early '90s, many studies have analyzed the acoustic properties of babies' cries and since then, researches using statistical methods to recognize different types of cries were reported. For example, Baeck and Souza [3] built a Bayesian classifier for babies' cries to detect whether a baby is in pain or not. Bănică et al. [4] applied the Gaussian Mixture Models-Universal Background Model (GMM-UBM) and an I-vectors-based method, which had proved to be successful in speech and language recognition, to the task of babies' cries recognition. Recently, with the success of deep learning in speech recognition, research that utilizes deep learning methods to recognize different types of infants' cries has been increasing. For example, Chang and Li [5] presented a work where the audio data was first converted into a spectrogram using Fast Fourier transform (FFT) and then classified into hungry, in pain, and sleepy using a convolutional neural network (CNN) as a classifier. Later, the same authors extended their work by using a 2D CNN model to detect the cry signal and then used a 1D CNN model to recognize the reason behind the detected cry signal [6]. Turan and Erzin [7] proposed a special kind of CNN model known as Capsule Network Architecture to recognize infants' cries and achieved results that outperformed traditional CNN models. After training classifiers using different acoustic features of babies' cries, such as Bark Frequency Cepstral Coefficients (BFCC) and Mel Frequency Cepstral Coefficients (MFCC), Liu et al. [8] concluded that the artificial neural network (ANN) classifier trained on BFCC had the best recognition performance results.

Most of the work previously mentioned built their own deep learning architectures and trained them on babies' cries datasets directly. However, those datasets tend to be small because the collection of such data is expensive due to the ethical and legal issues involved. Besides, the data collected at home or in hospitals is very likely to contain background noise such as the one caused by people talking [6]. Moreover, the cry sounds, pause durations and frequency differ from one baby to another [9]. Thus, deep learning models, which are

† Corresponding author

data-hungry methods, end up suffering from over-fitting and low validation accuracy because they are trained directly on small babies' cries datasets.

While all the three issues need to be eventually addressed, this paper focuses mainly on dealing with the problem of low availability of infants' cries data, and in order to solve this issue and guarantee a good recognition performance, we considered some strategies that have been proposed for similar cases. These include methods such as data argumentation, deep transfer learning [10], and semi-supervised learning. Among those methods, one that particularly stood out as an important tool to solve the issue of insufficient effective training data is deep transfer learning. Transfer learning has been widely applied in various research related to recognition and it showed superior results compared to other strategies in audio classification with a small number of data [11].

In order to eliminate the negative impact of insufficient baby cry data and improve the babies' cries recognition, in this paper, we take advantage of deep transfer learning to propose a simple and easy-to-use infant cry recognition system that can help first-time parents recognize four types of babies' cries, i.e. hunger, tiredness, boredom, and discomfort with a high accuracy. The proposed system first extracts the log-mel spectrogram from the cry signal, and then gives it as input to a fine-tuned CNN model named VGGish [12] that was pre-trained on AudioSet [13], giving it the advantage of having learnt on a large and diverse dataset. The VGGish model then classifies the data into one of the four classes and informs the parents of the result.

2. PROPOSED METHOD

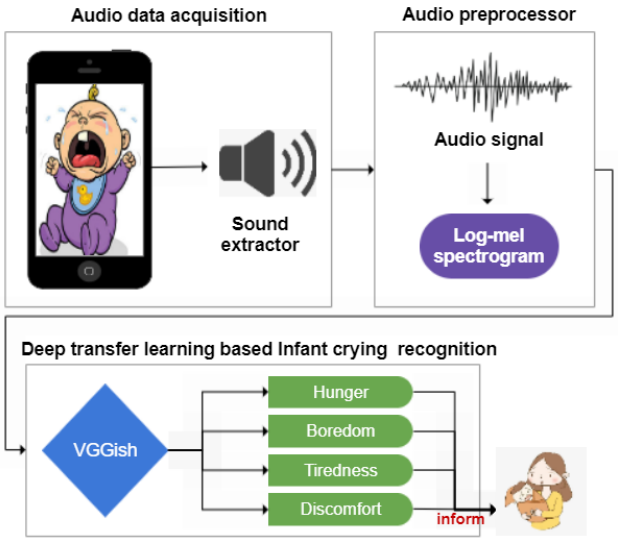
The overall architecture of the proposed system is shown in Figure 1. The system is composed of three modules: *audio data acquisition module*, *audio preprocessing module*, and *deep transfer learning based infant cry recognition module*.

2.1 Audio Data Acquisition Module

This module receives video data from several smartphones, then extracts 0.18~2.5 seconds long audio data with a mono channel, using a 16kHz sampling rate and 16-bit resolution, and forwards the sound signals to the preprocessor.

2.2 Audio Preprocessing Module

The higher correlation between pre-training and transfer learning tasks, the better performance can be obtained on the transfer task [10]. The VGGsih architecture uses 96×64 log-mel spectrogram as its input, thus, for a better feature transfer, in this module, after pre-emphasis and z-score normalization, the audio data is converted into log-mel spectrogram with 64 mel-bins and 96 time-bins by applying the public VGGish spectrogram feature extractor.

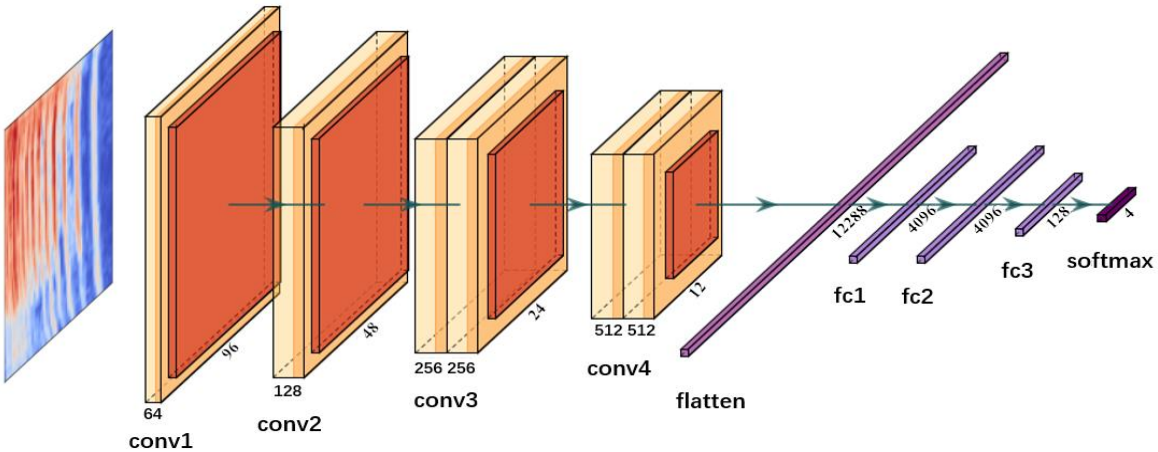


(Figure 1) General architecture of the infant cry recognition system.

2.3 Deep Transfer Learning Based Infant Cry Recognition Module

Deep neural networks often contain millions of parameters. The initialization of the weights of these parameters can have a significant effect on the performance results and it has been a subject of continuous research. Instead of choosing random initialization, initialization-based transfer learning can offer better initialization by using weights from a neural network trained on other data or tasks [10].

To take advantage of the deep transfer learning, in this module, the log-mel spectrogram feature is fed to the VGGish model, which is a variant of the VGG model specifically built for audio classification. The VGGish model was trained on AudioSet which contains over 2 million human-labeled 10-



(Figure 2) VGGish architecture used for infant cry recognition.

second YouTube soundtracks and is much larger compared to our collected babies' cries dataset. In this paper, we fine-tuned and trained the collected baby cry data on the pre-trained VGGish model and, since the number of features (128 embedded spaces) is small enough, we added a softmax layer directly after the last fully connected layer to perform the final step of classification. The VGGish architecture is provided in figure 2.

3. EXPERIMENTAL RESULTS

3.1 Data Collection Experiments

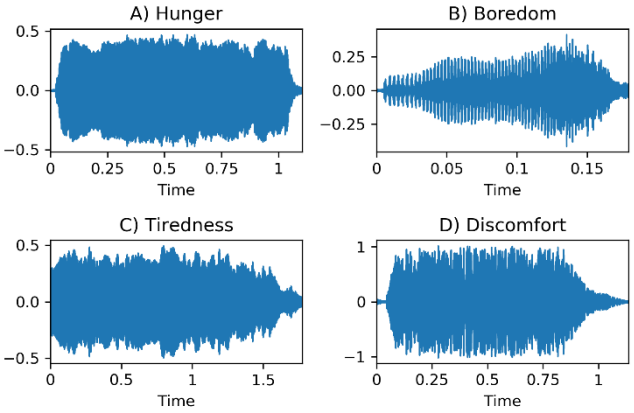
The infant cry data used in this paper was collected and annotated by professional baby caregivers and experienced nurses using their own smartphones at home and in a hospital at Hebei, China. Permission to collect data and consent to use it in our research was obtained from the parents of each baby. The collected data was stored in mp4 format files 10~15 seconds long each, then 0.18~2.5 seconds long audio clips were manually extracted from the videos and edited. In total, 852 audio clips contain 4 kinds of cry signals, namely hunger, boredom, tiredness, and discomfort, that we collected from 10 babies including 5 boys and 5 girls aged between 1 and 4 months old. The dataset is described in detail in table 1 and the number in each field represents the number of audio clips of the different cries of different babies. The hunger cry was collected when the baby hadn't been fed in a long time. The boredom cry was collected when the baby was trying to get a hug or trying to get attention. The tiredness cry was collected when the baby was sleepy, and the discomfort cry was collected when the baby was getting an injection or having colic. The signal waveforms of different cries are provided in figure 3. As seen in the figure, it is not easy to differentiate between different cries just by looking at their waveforms.

<Table 1> Summary of the babies' cries dataset

Infant	Hunger	Boredom	Tiredness	Discomfort
1 month boy	35	0	0	9
1 month boy	30	6	4	12
1 month girl	14	144	24	0
1 month girl	20	1	6	2
2 months boy	56	76	38	14
2 months boy	14	28	2	106
3 months boy	14	12	28	17
3 months girl	20	31	30	20
3 months girl	0	14	9	0
4 months girl	2	0	11	3
Total	**205**	**312**	**152**	**183**

3.2 VGGish for Infant Cry Recognition

We used a VGGish model pre-trained on AudioSet following the same architecture that is shown in figure 2. The input is a log-mel spectrogram of size 96×64, followed by four groups of convolution/maxpool layers. The filter size used in each convolution layer is 3×3 and in each pooling layer, the filter size is 2×2. After that, there are three fully connected layers with the dimensionality of 4096, 4096 and 128 consecutively. Since VGGish can embed the 96×64 size log-mel spectrogram into 128-dimensional highly represented features which can be easily distinguished, we directly added a SoftMax layer at the end without any need to use a downstream model to perform classification.



(Figure 3) Audio signal waveform of different baby cries.

3.3 Implementation Details

The network's fine-tuning and training were conducted in a computer running Windows 10, with an Intel i7-6700K CPU, 32GB of RAM, and a GTX 1080 graphic card.

The VGGish network was implemented using the Keras library [14] with TensorFlow as backend. We used both of the convolution layer weights and fully connected layer weights that were pre-trained on AudioSet as the initial weights and then fine-tuned on our babies' cries dataset using the Adam optimizer with a 0.0001 learning rate, a beta1 of value 0.9, beta2 of value 0.999, ReLU as the activation function and we trained it for 30 epochs. Five-fold cross-validation in the scikit-learn library was used to evaluate the performance of the VGGish model on the babies' cries dataset.

3.4 Results & Comparison

The result of the 5-fold cross-validation experiment using our proposed method is shown in table 2 below. Precision, recall, and f1-score metrics for each class are provided in table 3. As seen in table 2 and table 3, the average accuracy of the five folds is 96.3%, the weighted average precision, recall, and f1-score are 0.964, 0.963 and 0.963 respectively. A comparison of the most recent research on infants' cries recognition is provided in table 4. As seen in the table, it is a trend to combine spectrograms with CNN to recognize infants' cries reasons, and our proposed system achieves the best recognition results, confirming that transfer learning helps improve the performance of infant cry recognition. Also, the results show that the proposed system can be used in real applications to help inexperienced first-time parents understand the needs of their crying babies immediately.

<Table 2> Five-fold cross-validation results of infant cry recognition

Fold	Accuracy (%)
1	96.5
2	95.4
3	96.0
4	95.9
5	97.7
Average	96.3

<Table 3> Precision, recall, and f1-score metrics results of infant cry recognition

Class	Precision	Recall	F1-Score
Hunger	0.963	0.976	0.969
Boredom	0.981	0.994	0.987
Tiredness	0.935	0.929	0.932
Discomfort	0.956	0.925	0.940
Weighted Avg.	0.964	0.963	0.963

<Table 4> Comparison of the most recent research on infant cry recognition

Dataset	Feature Used	Classifier	Class Num.	Acc. (%)	Ref.
DBL	MFCC	GMM	5	70.0	[4]
Private	Spectro-gram	CNN	3	78.5	[5]
Private	Wave-form	CNN	4	78.3	[6]
CRIED	Spectro-gram	Capsule Network	3	86.1	[7]
Private	BFCC	ANN	6	76.5	[8]
DBL	Spectro-gram	CNN	5	89.0	[15]
Private	MFCC	RBM-CNN	5	78.6	[16]
Private	Spectro-gram	VGGish	4	96.3	Proposed

4. CONCLUSION

In order to address the necessity of infants' cries recognition, this paper proposed a deep transfer learning-based infant cry recognition system to help inexperienced parents recognize the cause of their babies' cries and take better care of them. We took advantage of the VGGish model pre-trained on AudioSet and fine-turned it on our collected dataset. Our results are promising and confirm that transfer learning helps secure a good performance in infants cries' recognition. In our next step, we will focus on making the proposed system noise-robust so that it can be applied in real life situations where background noise normally occurs.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3B07044938 and NRF-2020R1I1A3070835).

References

[1] P. High, "The Happiest Baby on the Block: The New Way to Calm Crying and Help Your Newborn Baby Sleep Longer," Journal of Developmental & Behavioral Pediatrics, Vol. 26, No. 1, pp. 68-69, 2005.

[2] R.E. Grunau, M.F. Whitfield, J.P. Thomas, A.R. Synnes, I.L. Cepeda, A. Keidar, M. Rogers, M. MacKay, P.H. Richard, and D. Johannesen, "Neonatal Pain, Parenting Stress and Interaction, in Relation to Cognitive and Motor Development at 8 and 18 Months in Preterm Infants," Pain, Vol. 143, No. 1-2, pp. 138-146, 2009.

[3] H.E. Baeck and M.N. Souza, "A Bayesian Classifier for Baby's Cry in Pain and Non-Pain Contexts," Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 3, pp. 2944-2946, 2003.

[4] I.A. Bănică, H. Cucu, A. Buzo, D. Burileanu, and C. Burileanu, "Automatic Methods for Infant Cry Classification," International Conference on Communications, pp. 51-54, 2016.

[5] C.Y. Chang and J.J. Li, "Application of deep learning for recognizing infant cries," IEEE International Conference on Consumer Electronics-Taiwan, pp. 1-2, 2016.

[6] C.Y. Chang and L.Y. Tsai, "A CNN-Based Method for Infant Cry Detection and Recognition," Workshops of the International Conference on Advanced Information Networking and Applications, pp. 786-792, 2019.

[7] M.A.T. Turan and E. Erzin, "Monitoring Infant's Emotional Cry in Domestic Environments Using the Capsule Network Architecture," Interspeech, pp. 132-136, 2018.

[8] L.C. Liu, W. Li, X.W. Wu, and B.X. Zhou, "Infant Cry Language Analysis and Recognition: an experimental approach," IEEE/CAA Journal of Automatica Sinica, Vol. 6, No. 3, pp. 778-788, 2019.

[9] P.S. Zeskind and B.M. Lester, "Analysis of Infant Crying," Biobehavioral Assessment of the Infant, pp. 149-166, 2001.

[10] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How Transferable are Features in Deep Neural Networks?" Advances in Neural Information Processing Systems, pp. 3320-3328, 2014.

[11] J. Pons, J. Serrà, and X. Serra, "Training Neural Audio Classifiers with Few Data," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 16-20, 2019.

[12] 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. Wilson, "CNN Architectures for Large-Scale Audio Classification," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 131-135, 2017.

[13] J.F. Gemmeke, D.P.W. 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," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 776-780, 2017.

[14] Keras, ver.2.4.0, Available Online: https://github.com/keras-team/keras (accessed on 15 September 2020).

[15] E. Franti, I. Ispas, and M. Dascalu, "Testing the Universal Baby Language Hypothesis - Automatic Infant Speech Recognition with CNNs," 2018 41st International Conference on Telecommunications and Signal Processing, pp. 1-4, 2018.

[16] B.F. Yong, H.N. Ting, and K.H. Ng, "Baby cry recognition using deep neural networks," World Congress on Medical Physics and Biomedical Engineering 2018, pp. 809-813, 2019.