Recognition of a Baby's Emotional Cry Towards Robotics Baby Caregiver

Regular Paper

Shota Yamamoto¹, Yasunari Yoshitomi², Masayoshi Tabuse², Kou Kushida³ and Taro Asada²

¹ Kumon Educational Japan Co., Ltd., Japan
² Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, Japan
³ Kyoto Prefectural Tanabe Senior High School, Japan
* Corresponding author E-mail: yoshitomi@kpu.ac.jp

Received 12 Jun 2012; Accepted 5 Dec 2012

DOI: 10.5772/55406

© 2013 Yamamoto et al.; licensee InTech. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract We developed a method for pattern recognition of baby's emotions (discomfortable, hungry, or sleepy) expressed in the baby’s cries. A 32-dimensional fast Fourier transform is performed for sound form clips, detected by our reported method and used as training data. The power of the sound form judged as a silent region is subtracted from each power of the frequency element. The power of each frequency element after the subtraction is treated as one of the elements of the feature vector. We perform principal component analysis (PCA) for the feature vectors of the training data. The emotion of the baby is recognized by the nearest neighbor criterion applied to the feature vector obtained from the test data of sound form clips after projecting the feature vector on the PCA space from the training data. Then, the emotion with the highest frequency among the recognition results for a sound form clip is judged as the emotion expressed by the baby’s cry. We successfully applied the proposed method to pattern recognition of baby’s emotions. The present investigation concerns the first stage of the development of a robotics baby caregiver that has the ability to detect baby's emotions. In this first stage, we have developed a method for detecting baby's emotions. We expect that the proposed method could be used in robots that can help take care of babies.

Keywords Baby’s Cry, Speech Recognition System, Fundamental Frequency, Emotion, Pattern Recognition

1. Introduction

In Japan, the number of births has shown a downward trend. This trend is considered a serious problem. One of the reasons for the decrease in births may be the lack of childcare support as a social system. In many cases, the mother works outside the home and needs childcare to take care of her baby.

The number of people living together has also shown a downward trend. As a result, a young mother often takes care of her baby by herself. However, she might not have the opportunity to learn appropriate childcare practices.

As is well known, it can be extremely dangerous to leave a baby alone. Nevertheless, a mother typically must do much housework and may want to enjoy another activity. In these cases, raising a child safely becomes more difficult. Unfortunately, baby abuse and neglect have been increasing because of the stress of mothers taking care of their babies. Therefore, we think that decreasing
the mother’s stress may slow the decreasing number of births.

Investigations on a baby’s cry have primarily used frequency analysis [1]-[9]. However, clipping of the sound form of a baby’s cry from wave data has been performed by manual operations. Because manual operation is time consuming, automatic detection of a baby’s cry and automatic recognition of a baby’s emotion have not been resolved.

We have been researching a system for improving baby care by recognizing a baby’s emotion from the baby’s cry. As stated above, clipping of the sound form of a baby’s cry from wave data has typically been performed manually [9]. We previously proposed a method for detecting a baby’s cry by using a speech recognition system [10] and fundamental frequency analysis [11]. In this paper, we propose a method for pattern recognition of baby’s emotions (“discomfortable,” “hungry”, or “sleepy”) by using the sound form clips, which are detected by our reported method [11].

To our best knowledge, there are no published papers on robotics baby caregiver. The present investigation concerns the first stage of the development of a robotics baby caregiver that has the ability to detect baby’s emotions. In this first stage, we have developed a method for detecting baby’s emotions. We expect that the proposed method could be used in robots that can help take care of babies.

2. Method for Detection of a Baby’s Cry

Fig. 1 shows the flow chart of our reported method [11] for detection of a baby’s cry. We explain the procedure in the following.

2.1 Word Recognition

We use a speech recognition system named Julius [10] for word recognition [11]. When Julius recognizes a sound form segment as expressing a certain word, the sound form segment is used in the next processing. However, in our reported method [11], when Julius recognizes a sound form segment as silent, the sound form segment is omitted.

2.2 Threshold Treatment of the Sound Form

The sound form segment recognized as not silent by Julius might contain sound mostly composed of noise. We neglect the sound form segment having an amplitude below a threshold decided experimentally beforehand, because the sound form segment may be composed of noise [11]. The threshold was experimentally decided as the value of \( m + 2\sigma \), where \( m \) and \( \sigma \) are the average and the standard deviation, respectively, of the sound amplitude when neither voice nor unusual sound is present in the recorded sound form [11].

2.3 Detection of a Baby’s Cry

We proposed the following two conditions for recognizing a sound form segment as coming from a baby’s cry [11]:

Condition 1: The word reliability for a sound form segment obtained using Julius is under a specified threshold.

Condition 2: For a certain time period, the change of the fundamental frequency of the sound form segment is more than another specified threshold.

The two thresholds and the time period are experimentally decided. When at least one of the above two conditions is met, the sound form segment is judged as coming from a baby’s cry. We do not need a special template to detect baby’s cry.

2.3.1 Judgment by Word Reliability

Julius outputs a word reliability of 0 to 1 for a recognized word. High word reliability occurs when the difference of the likelihood of other candidate words is large. Accordingly, when the value of the word reliability is small, the speech recognition result may be doubtful. In the case of a baby’s cry, the word reliability given by Julius is assumed to be small.

We decided a threshold for distinguishing the word reliability of a baby’s cry from that of an adult’s voice [11].
2.3.2 Fundamental Frequency Analysis for a Short Time Period

In this study, we found that the fundamental frequency of a baby’s cry suddenly changes, for example, in the time ranges of 1.5 to 1.8 s and 6.6 to 6.8 s, as shown in Fig. 2 [11].

Therefore, we decided a certain time period and another threshold of change in the fundamental frequency for distinguishing a baby’s cry from other sounds [11].

![Sound form](image)

*Figure 2. Large change of fundamental frequency of baby’s cry for a short period [11].*

3. Pattern Recognition of a Baby’s Cry

Fig. 3 shows a flow chart for emotion recognition of a baby by using the baby’s cry. The 32-dimensional fast Fourier transform (FFT) is performed for the sound form clips to be used as training data. Then, the power of the sound form judged as a silent region is subtracted from each power of the frequency element. The power of each frequency element after the subtraction is treated as one of the elements of the feature vector. We perform principal component analysis (PCA) for the feature vectors of the training data under the condition that the main components are assembled up to the lowest dimension, at which the cumulative contribution ratio of PCA exceeds 80%. Then, the feature vectors for the sound form clips of the test data are calculated in the same way as those for the training data. The emotion of the baby is recognized by the nearest neighbor criterion applied to the feature vector obtained from the test data of the sound form clips after projecting the feature vector on the PCA space set from the training data. We use the Euclidean distance as the distance metric between the feature vectors. Then, the emotion with the highest frequency among the recognition results for the sound form clip is judged as the emotion expressed by the baby’s cry. When the recognition results of two or more emotions have the highest frequency among the three emotions, the second nearest neighbor feature vector for each sound form clip in the PCA space is treated in the same manner as that of the nearest one for judging the emotion expressed by the baby’s cry.

![Flow chart](image)

*Figure 3. Flow chart for recognizing a baby’s emotion from his or her cry.*

| Sound form clip No. | Discomfortable | Hungry | Sleepy | Judgment |
|---------------------|----------------|--------|--------|----------|
| 1                   | ○              | ○      |        | Sleepy   |
| 2                   | ○              | ○      | ○      | Sleepy   |
| 3                   | ○              | ○      | ○      | Sleepy   |
| 4                   | ○              | ○      | ○      | Sleepy   |
| 5                   | ○              | ○      | ○      | Sleepy   |
| 6                   | ○              | ○      | ○      | Sleepy   |
| 7                   | ○              | ○      | ○      | Sleepy   |
| 8                   | ○              | ○      | ○      | Sleepy   |

*Table 1. Example of emotion recognition using a continuous cry form.*

Table 1 shows an example that explains the judgment of the emotion expressed by the baby’s cry. In this case, “sleepy” is judged as the emotion expressed by the baby’s cry because the emotion with the highest frequency is “sleepy” among the three emotions.

4. Performance Evaluation

4.1 Experimental Conditions

4.1.1 Voice Recording

We used a wireless or wired microphone and notebook computer for recording sound. The sound forms are
We used three kinds of sound forms for a total of four experiments: (1) the cry of a male baby (age: one and one-half months), which was recorded in his home and was composed of 57 pieces of continuous sound forms having 10 to 12 cries for each piece; (2) a mixture (hereinafter referred to as mixture 1) of baby, adult voices, car and toy noises, which were recorded in a day nursery and composed of three pieces of continuous sound forms having 8 to 21 cry regions for each piece; (3) a mixture (hereinafter referred to as mixture 2) of baby and adult voices, which were also recorded in the day nursery and were composed of five pieces of continuous sound forms having 8 to 21 cry regions for each piece; (3) voices of five male and two female adults, which were recorded in our laboratory at Kyoto Prefectural University and were composed of seven pieces of continuous sound forms having 14 to 22 pronunciations of “taro” for each piece.

The sound data described above as (3) were the same neutral parts used in our previously reported study [12]. The experiment was performed in the following computational environment: the personal computer was a DELL OPTIPLEX GX260 (CPU: Pentium IV 2.4 GHz; main memory: 512 MB); the OS was Microsoft Windows XP; the development language was Microsoft Visual C++ 6.0.

4.1.2 Method for Evaluation

First, we obtained the threshold for the amplitude of the sound form and the threshold for the word reliability using sound data for the first half of (1) the baby’s cry, and (3) the adult voices, as described in Section 4.1.1. Fig. 4 shows the distribution of the word reliability of a baby’s cry and an adult voice. Then, using the sound data of (1), we experimentally decided the condition of a sudden change in the fundamental frequency as 0.1 s of the time for checking and 150 Hz as the threshold.

Moreover, to judge whether the sound form segment came from the baby’s cry, we checked all sound form segments that were recognized as not silent by Julius, in the following two ways: the sound form segments were checked by eye, and the real sounds were checked by ear. Then, we compared the sound form segments, judged as coming from the baby’s cry by the eye and the ear checking, with the segments recognized by our reported method [11]. Here, we used two criteria, the detection and mis-detection rates defined below, for evaluation of our reported method [11].

The detection rate (%) is defined as \( \frac{(N(A\cap B)/N(A)) \times 100}{N(X)} \), where \( N(X) \) is the number of elements of set \( X \), \( A \) is the set of sound segments of the baby’s cry, and \( B \) is the set of sound segments detected by the proposed method. The mis-detection rate (%) is defined as \( \frac{(N(B) - N(A\cap B))}{N(A\setminus B)} \times 100 \), where \( N(X) \), \( A \), and \( B \) are described above, and \( J \) is the set of all sound form segments judged as not silent by Julius. Fig. 5 shows the relationship among the sets of \( A, B, \) and \( J \). In this study, the accuracy of the distinguishing of baby’s cry from other sounds is measured as the above detection rate.

![Figure 4. Word reliability distributions of a baby and an adult](image1.png)

![Figure 5. Schematic diagram showing sets and their relationship used for explaining detection and mis-detection rates](image2.png)
reliability, resulting again in a high mis-detection rate. As shown in Fig. 4, word reliability for the adult voice of “taro” is sometimes low. Therefore, the adult voice may lead to mis-detection because word reliability is one of the judgment criteria for a baby’s cry.

The main goal of this study is to apply the proposed method to improve baby care in the baby’s home. Therefore, the most important result among those shown in Table 2 is the column of “Baby”. In this case, the detection rate is 69.4% and the mis-detection rate is 2.0%. Accordingly, our reported method [11] can be useful as a pre-processing module of the proposed method.

Table 2. Results of a baby’s cry detection [11].

|                | Baby | Adult | Mixture 1 | Mixture 2 |
|----------------|------|-------|-----------|-----------|
| Number of sound clips detected by Julius | 1172 | 333   | 186       | 239       |
| Number of baby’s cry clips | 569  | 0     | 60        | 126       |
| Number of sound clips detected by proposed method (Number of baby’s cry clips in the above number) | 407 (395) | 66 (0) | 82 (39) | 97 (94) |
| Detection rate | 69.4% | ---   | 65.0%    | 66.2%     |
| Mis-detection rate | 2.0%  | 19.8% | 34.1%    | 2.7%      |

Table 3. Accuracy (%) of recognition of baby’s cry and other sounds.

Table 3 shows accuracy (%) of recognition of baby’s cry and other sounds. Table 3 is obtained using detection rate and mis-detection rate shown in Table 2. The average accuracy of recognition of baby’s cry and other sounds is 83.7 %, 80.2 %, 65.5 % and 81.8 % in the case of baby, adult, mixture 1 and mixture 2, respectively. In the case of adult, there were no sounds of baby’s cry. Therefore, the average accuracy of recognition of baby’s cry and other sounds corresponds to the accuracy of recognition of other sounds. As described above, the most important result among those shown in Table 3 is the case of “Baby”. In this case, the average accuracy of recognition of baby’s cry and other sounds is 83.7 %.

4.2.2 Application to Emotion Recognition

Table 4 shows examples that explain the judgment of the emotion expressed by the male baby’s cry. The emotion of “hungry” was easier to be recognized by the proposed method than those of “discomfortable” and “sleepy”.

Table 5 shows the emotion recognition results. The average accuracies of emotion recognition are 62.1%. This result suggests that the proposed method is applicable to emotion recognition.

Table 4. Example of emotion recognition using continuous cry form of a male baby.

The accuracy of “discomfortable” is lower than that of other types of emotion. This is because the sound expressing “discomfortable” has significant variations,
resulting in difficulties of pattern recognition when the nearest neighbor criterion in the PCA space is used. The accuracy of “discomfortable” might be improved when considering the second nearest neighbor in the PCA space.

| Emotion     | Recognition accuracy |
|-------------|----------------------|
| Discomfortable | 3/10 (30%)           |
| Hungry      | 13/14 (92.9%)        |
| Sleepy      | 2/5 (40%)            |
| Total       | 18/29 (62.1%)        |

Table 5. Results of emotion recognition rates using a continuous cry form.

One of the reasons why the emotion recognition rate of the proposed method for “sleepy” is low is the sound form clip expressing “sleepy” may not be detected by our reported method [11], which uses a threshold for the amplitude of the sound form to neglect noise. Because the sound form clip expressing “sleepy” tends to have a small amplitude, this emotion is sometimes neglected.

We think that the stress of baby care while doing activities away from the baby and waiting for notification of the baby’s cry and corresponding emotion recognized by the proposed method would be less than the stress while doing activities and always paying attention to the sound sent through a wireless microphone beside the baby. The system can have the function of sending an image of the baby to the person taking care of the baby only when the baby expresses an emotion. Consequently, the system enables the person to stay apart from the baby when the baby is not in need of immediate care. We expect that the system can be implemented in robots helping to take care of babies.

Since there are no published papers on robotics baby caregiver to our best knowledge, to develop a robotics baby caregiver might be one of the most challenging targets in the field of robotics research. The present investigation concerns the first stage of the development of a robotics baby caregiver that has the ability to detect baby’s emotions. In this first stage, we have developed a method for detecting baby’s emotions. The second stage is to develop an automatic, real-time system using the proposed method. The third stage is to develop a robotics baby caregiver that has functions developed at the first and second stages and functions for appropriately reacting according to the detected baby’s emotions for taking care of the baby.

5. Conclusion

The present investigation concerns the first stage of the development of a robotics baby caregiver that has the ability to detect baby’s emotions. In this first stage, we developed a method for pattern recognition of baby’s emotions expressed in the baby’s cries. The main goal of this study is to apply the proposed method to improve baby care in the baby’s home. We successfully applied the proposed method for pattern recognition of baby’s emotions. The average accuracies of emotion recognition were 62.1% when assuming that the emotion was one of “discomfortable”, “hungry”, or “sleepy”. In future work, we will apply the proposed method to more babies because the emotions expressed in their cries might have variations. Moreover, we will investigate the influence of noise on the accuracy of the proposed method. We expect that the proposed method could be used in robots that can help take care of babies.

6. Acknowledgments

We would like to thank all the participants who cooperated with us in the experiments.

7. References

[1] Truby H. M, Lind J (1965) Cry Sounds of the Newborn Infant. Acta Paediatr. Scand. Suppl., 163 : 8-59.
[2] Wolff P. H (1969) The Natural History of Crying and Other Vocalizations in Early Infancy. in B. W. Foss, (ed.), Determinants of Infant Behavior IV. London: Methuen and Co, 81-111.
[3] Fuller B. F, Horii Y (1986) Differences in Fundamental Frequency, Jitter, and Shimmer among Four Types of Infant Vocalizations. J. of Communication Disorders, 19(6) : 441-447.
[4] Fuller B. F, Horii Y (1988) Spectral Energy Distribution in Four Types of Infant Vocalizations. J. of Communication Disorders, 21(3): 251-261.
[5] Xie Q, Ward R. K, Laszlo C. A (1996) Automatic Assessment of Infants’ Levels-of-distress from the Cry Signals. IEEE Trans. Speech and Audio Processing, 4(4): 253-265.
[6] Kikuchi K, Arakawa K (2006) Estimation of the Cause of Younger Babies’ Cries by Frequency Analyses of Their Voice. IEICE Technical Report, SIS2005-69 : 55-60 (in Japanese).
[7] Mima Y, Arakawa K (2006) Cause Estimation of Younger Babies’ Cries from the Frequency Analyses of the Voice – Classification of Hunger, Sleepiness, and Discomfort. IEICE Technical Report, SIP2006-79 : 43-46 (in Japanese).
[8] Nishimura K, Mima Y, Arakawa K (2007) Estimation of the Cause of Cries for One-Month Old Babies. IEICE Technical Report, SIS2006-87: 35-40 (in Japanese).
[9] Kushida K, Tabuse M, Yoshitomi Y (2008) Pattern Recognition of Emotional States in Baby Voice. Proc. of Human Interface Symposium, 25-28 (in Japanese).
[10] Open-Source Large Vocabulary CSR Engine Julius. http://julius.sourceforge.jp/en_index.php?q=index-en.html. Accessed 2012 June 3.
[11] Yamamoto S, Yoshitomi Y, Tabuse M, Kushida K, Asada T (2010) Detection of Baby Voice and its Application Using Speech Recognition System and Fundamental Frequency Analysis. Proc. of 10th WSEAS Int. Conf. on Applied Computer Science, 341-345.

[12] Nakano M, Ikezoe F, Tabuse M, Yoshitomi Y. (2009) A Study on the Efficient Facial Expression Using Thermal Face Image in Speaking and the Influence of Individual Variations on Its Performance. J. IEEJ, 38(2): 156-163 (in Japanese).