Emotion Clustering Using the Results of Subjective Opinion Tests for Emotion Recognition in Infants' Cries

N. Satoh[1], *K. Yamauchi*[1], *S. Matsunaga*[1], *M. Yamashita*[1], *R. Nakagawa*[2] *and K. Shinohara*[2]

[1]Department of Computer and Information Sciences, Nagasaki University, Nagasaki, JAPAN
[2]Department of Translational Medical Sciences, Nagasaki University, Nagasaki, JAPAN
{mat, yamauchi, masaru}@cis.nagasaki-u.ac.jp

10.21437/Interspeech.2007-606

Abstract

This paper proposes an emotion clustering procedure for emotion detection in infants' cries. Our clustering procedure is performed using the results of subjective opinion tests regarding the emotions expressed in infants' cries. Through the procedure, we obtain a tree data structure of emotion clusters that are generated by the progressive merging of emotions. Emotion merging is carried out on the condition that the objective function concerning the ambiguity of emotions that were detected in the opinion tests is minimized. Clustering experiments are performed on the results of opinion tests completed by infants' mothers and baby-rearing experts. The experimental results show that the proposed clustering, which considers the evaluation rank of each emotion, is superior to the clustering that is only concerned with the detection/nondetection of each emotion. Based on the clustering results, we performed a preliminary recognition experiment on two emotion clusters. According to the recognition results, the proposed emotion cluster achieves a detection rate of 75%, which shows the effectiveness of the proposed clustering procedure.

Index Terms: emotion recognition, infants' cry, clustering

1. Introduction

Crying is an important means by which infants convey their intentions to their parents [1]. It is supposed that around the age of two months, it is possible to gradually distinguish between the different emotions in an infant's cry based on its sound. However, in general, it is difficult to understand the emotions that infants wish to express through their cries (the causes of crying), particularly for people who are not sufficiently experienced in childcare. For this reason, they may be unable to satisfy the wants of crying infants. It is somewhat easier to deal with an infant whose emotions are automatically detected in his/her cry, as this helps to avoid unwanted treatment.

From the viewpoint of emotion detection, a number of studies have been conducted on the acoustic analysis of an infant's cry [2, 3]. An infant's pain is one of the traditional research topics for the detection of emotions [4]. In addition, classification between "hunger" and "sleepiness" has been studied in recent years [5]. There are also some emotion detection products currently available in the market [6]. These employ simple matching techniques using acoustic features.

We have been collecting infants' cries and studying an emotion detection procedure that is based on a maximum likelihood approach using hidden Markov models (HMMs) [7]. Our study required infants' mothers to judge the emotions expressed in the samples. Further, with the assistance of baby-rearing experts, we conducted subjective opinion tests on the emotions expressed in each sample. Assuming that the judgments of each infant's mother were correct, we performed some detection experiments. However, we faced some problems during our experiments. The mothers frequently selected two or more emotions for each cry sample, and the characteristics of emotion detection were very different among the mothers—for example, one mother detected the emotion of "sadness" more frequently than the other mothers. Furthermore, the agreement rate between the emotions judged by the mothers and those detected by the experts was not high.

In an attempt to address the abovementioned problems, we propose an emotion clustering procedure (clustering of emotion types) for emotion detection. The concept of "emotion cluster" is introduced in order to deal with the ambiguity of the detected emotions. According to the results of the opinion tests, the mothers and the experts tend to detect the same set of emotions through many cry samples. Next, we merge these emotions into one emotion cluster by using the clustering technique. Emotion merging is carried out on the condition that the objective function concerning the ambiguity of detected emotions is minimized. Clustering experiments are performed on the results of subjective opinion tests that were completed by infants' mothers and baby-rearing experts. Finally, emotion detection experiments are performed on two emotion clusters, which are obtained by using the proposed clustering method.

2. Corpus of infants' cries

Our corpus of infant cries comprises the waveform data, the tag of emotions detected by mothers and baby-rearing experts, and transcriptions using the labels of the acoustic segments.

2.1. Subjective opinion tests

All the mothers were required to record their infants' cries over several days at home using a digital recorder. A total of 500 cries by 25 infants were recorded. The average duration of the recorded data was approximately 30 seconds. The infants ranged from 8 to 13 months of age.

After recording each cry, the infants' mothers judged the emotions expressed in the samples (subjective opinion test). In doing so, the mothers took into consideration not only the cries but also the infants' facial expressions, behaviors, etc. Ten kinds of emotion tags were prepared: pampered (psychological dependence), anger, sadness, fear, surprise, hunger, sleepiness, excretion, discomfort, and painfulness.

Emotion	Anger	Sadness	Hunger	Surprise	……
Rank	0	4	2	0	……

Figure 1: *Example of the emotion table*

The mothers assessed the emotions that caused the cry and recorded this assessment by filling in the emotion table. An example of this table is shown in Figure 1. The intensity of the emotions is to be ranked on a scale ranging from 0 (the emotion is not contained at all) to 4 (the emotion is contained fully). The mothers were permitted to choose two or more emotions. We provided the mothers with some examples, as follows, in order to explain each emotion:

- Sadness: her/his mother goes away
- Anger: her/his favorite toy is taken away
- Surprise: sudden loud sounds

We also conducted the subjective opinion tests on 200 recordings with the help of baby-rearing experts. Their judgments were made solely on the basis of the recordings. A subjective evaluation was performed in the same way as the judgment of the mothers. Before the test, we also provided the experts with some examples to explain each emotion.

These two opinion tests showed that the agreement between the emotions judged by the mothers with the highest rank and those detected by the experts was not high. For example, the agreement rate for the samples that the experts judged as having the top evaluation rank 4, was 41% [7].

2.2. Hand labeling of acoustic segment

We consider a cry to be composed of segments with acoustic characteristics. In order to recognize the emotions in the cry using a statistical method, we defined the segments according to their acoustic features and assigned a symbol to each segment. The corpus was hand-labeled using the symbols. Suppose a cry z comprises N segments, and let the i-th segment be s_i $(1 \le i \le N)$,

$$z = s_1 s_2 \cdots s_i \cdots s_N , \tag{1}$$

where the beginning time of segment s_{i+1} is the end time of segment s_i. In order to capture acoustic features precisely, following ten kinds of acoustic segments are prepared: a silent segment, a breath segment, a glottal sound segment (a cry that sounds like a cough), a typical cry segment, a babbling segment, a cooing segment, and so on [7].

3. Emotion clustering

For each cry sample, the mothers selected one or more emotions from among ten types of emotions. We supposed that when a mother or a baby-rearing expert selected one emotion e_i from among the emotion cluster set $X = \{e_1, \cdots, e_I\}$, she also simultaneously selected the emotion e_j $(i \neq j)$, where e_j is a single element among $Y = \{e_1, \cdots, e_{i-1}, e_\phi, e_{i+1}, \cdots, e_I\}$. The emotion e_ϕ indicates that she selected only one emotion e_j. Our clustering algorithm is based on the criterion of maximum reduction of the ambiguity of selection of emotions/emotion-clusters. Next, taking into account the formulation concerning conditional entropy, we set the objective function H as follows. We assume that emotion clusters m and n are merged into one cluster r.

$$\begin{aligned}\hat{m},\hat{n} &= \underset{m,n\,(m\neq n)}{\arg\min}\, H(Y \mid X) \\ &= \underset{m,n\,(m\neq n)}{\arg\max} \Big\{ \sum_{i(i\neq m,n)} P(r)P(i \mid r)\log P(i \mid r) + P(r)P(\phi \mid r)\log P(\phi \mid r) \\ &+ \sum_{i(i\neq m,n)} P(i)P(r \mid i)\log P(r \mid i) + \sum_{i(i\neq m,n)} \sum_{j(j\neq m,n)} P(i)P(j \mid i)\log P(j \mid i) \Big\} \end{aligned} \tag{2}$$

where $P(r) = P(m) + P(n)$. Terms $P(\cdot)$ and $P(\cdot \mid \cdot)$ are occurrence probability and conditional probability of each emotion cluster, respectively $(0 \le P(\cdot), P(\cdot \mid \cdot) \le 1)$. The $\hat{m}$ -th and $\hat{n}$ -th emotion clusters, which minimize the function H, are selected and merged into one emotion cluster. In our formulation, we devised the following two types of calculations for $P(\cdot)$ and $P(\cdot \mid \cdot)$: Method I is a merging method that considers the detection/nondetection of each emotion expressed in a cry and Method II is the method that takes into account the value of the evaluation rank.

3.1. Merging based on the detection/nondetection of each emotion (Method I)

In this method, the evaluation rank is used only to eliminate the small value. If "three or more" values are effective, then "two or less" values are ignored. Terms $P(\cdot)$ and $P(\cdot \mid \cdot)$ are calculated as follows:

$$P(i) = C(i) \Big/ K = \frac{1}{K} \sum_k \delta(i,k) , \tag{3}$$

where $C(i)$ is the number of samples in which emotion e_i was detected,

$$\delta(i,k) = \begin{cases} 1, & \text{if emotion } e_i \text{ was detected on the } k\text{ - th data} \\ 0, & \text{otherwise} \end{cases}$$

and K is the number of samples. Conditional probabilities are $P(j \mid i) = C(i,j)/C(i) = \sum_k \delta(i,j,k) \Big/ \sum_k \delta(i,k)$ and

$$P(i \mid r) = \sum_k \delta(i,r,k) \Big/ \left(\sum_k \delta(i,k) + \sum_k \delta(r,k) - \sum_k \delta(i,r,k) \right),$$

where $C(i,j)$ is the number of samples in which both emotions e_i and e_j were detected. (If these emotions are detected in the k-th sample, then $\delta(i,j,k) = 1$; otherwise, $\delta(i,j,k) = 0$)

3.2. Merging based on evaluation rank (Method II)

Method II considers the evaluation rank ($S_{i,k}$: rank for emotion e_i on the k-th sample, $0 \le S_{i,k} \le 4$). Since this rank is a value that is related to human perception, we change it to $S'_{i,k} = u^{S_{i,k}}$, if $S_{i,k} \neq 0$. If $S_{i,k} = 0$, we set $S'_{i,k} = 0$, where u is a positive constant more than one. (In our experiments, the constant u is 4.) Next, the value $S'_{i,k}$ is normalized as $w(i,k) = S'_{i,k} \Big/ \sum_i S'_{i,k}$. Taking into account the emotion rank, w is used in place of occurrence probabilities, so that the following approximation is introduced.

$$\begin{aligned} P(i) &\approx \frac{1}{K} \sum_k w(i,k)\delta(i,k), \\ P(j \mid i) &\approx \sum_k w(i,j,k)\delta(i,j,k) \Big/ \sum_k w(i,k)\delta(i,k) \\ &\approx \sum_k w(i,k)w(j,k)\delta(i,j,k) \Big/ \sum_k w(i,k)\delta(i,k). \end{aligned} \tag{4}$$

Since the estimation of $w(i,j,k)$ is difficult, we use $w(i,k)w(j,k)$ instead.

3.3. Clustering procedure of Methods I and II

The following is our emotion-clustering procedure:

Step-1. Select major emotions from among ten emotions using the threshold concerning $P(i)$ in Eq. (3) or (4).

Step-2. Select two emotions (clusters) by calculating Eq. (2) and merge them into one cluster.

Step-3. Repeat the merging process in Step-2 progressively until the desired number of emotion clusters is attained.

Such emotion clustering sometimes generates a very large emotion cluster, which is not appropriate for automatic detection. We impose the merging constraints based on $P(r)$ (Eq. (3) or (4)) in order to avoid this situation.

4. Emotion recognition

In our recognition approach, the acoustic likelihood of the input is calculated by using the acoustic models for each kind of emotion/emotion-cluster. Accordingly, the segment sequence with the highest likelihood is detected among all types of emotions or emotion clusters. Given acoustic evidence observation q, the our process of emotion recognition is to find the most likely segment sequence, $\hat{z}$, and the emotion cluster $\hat{e}$ which gives $\hat{z}$, satisfying

$$P(\hat{e},\hat{z}\mid q)=\max_{e,z}P(e,z\mid q) \tag{5}$$

The right-hand side of the above equation can be rewritten according to Bayes' rule as

$$P(e,z\mid q)=P(e,z)P(q\mid e,z)/P(q) \tag{6}$$

where $P(e,z)$ is a priori probability that the segment sequence z will be occurred on the emotion cluster e. Although we calculated the occurrence probabilities of the segments for each emotion using our cry corpus, there was no significant difference among them. Then, we neglect the term $P(e,z)$ in detecting the emotion cluster $\hat{e}$. $P(q\mid e,z)$ is the probability that when the infant utters the sequence z caused by the emotion cluster e the acoustic evidence q will be observed. Since $P(q)$ is not related to z and e, it is irrelevant to recognition. Then, we can apply the emotion recognition procedure to Eq. (5) as follows:

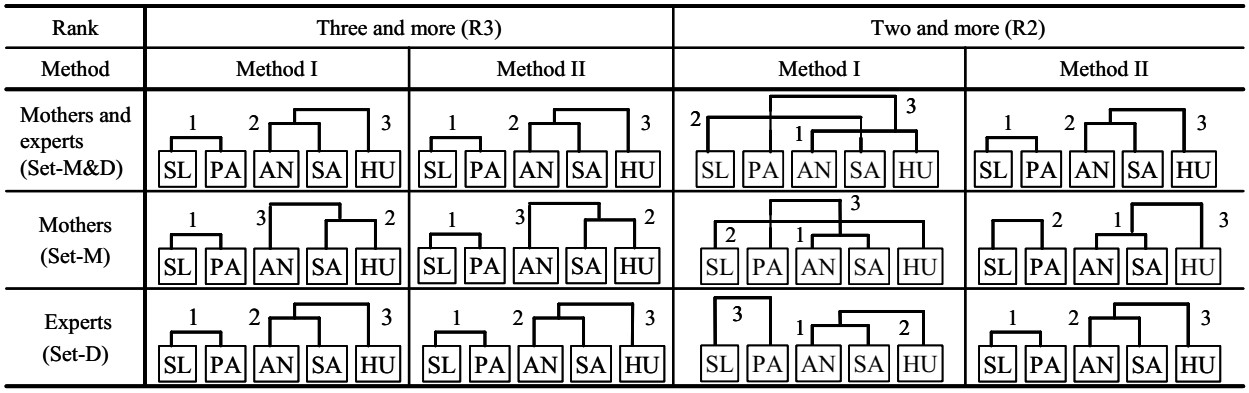
$$\hat{e},\hat{z}=\arg\max_{e,z}P(e,z\mid q)\approx\arg\max_{e,z}P(q\mid e,z) \tag{7}$$

In our experiments, we perform this maximization by using the likelihood of HMMs. Acoustic models of each segment are generated for each kind of emotion cluster in the training of HMMs. The cry data were sampled at 16 kHz. Every 10 milliseconds a vector of 12 FFT mel-warped cepstral coefficients and power was computed using a 25-millisecond Hamming window. Segment HMMs were 3-state 6-mixture, context-independent models. These models were generated for each segment and each emotion/emotion-cluster (emotion-dependent). A silent model was shared among emotions (emotion-independent).

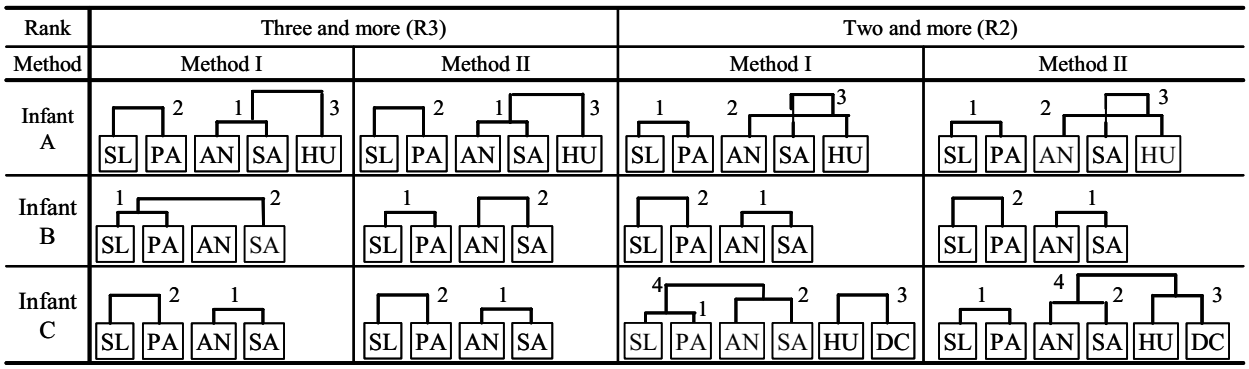
5. Recognition experiments

5.1. Emotion clustering experiments

The clustering experiments using Methods I and II were performed using the result samples of subjective opinion tests. In our first experiment, we evaluated the results samples concerning all the 25 infants; we tested three evaluation sets during this time: all test results (approximately 1100 samples, Set-M&D), a set of test results judged by infants' mothers (approximately 500 samples, Set-M), and a set of test results judged by three experts per sample (approximately 600 samples concerning 200 cry data, Set-D). The thresholds of evaluation ranks, which were used in clustering, were "three or more (R3)" and "two or more (R2)." The clustering results are illustrated in Figure 2-a. In these experiments, five major emotions were selected from among ten emotions by using Step-1. When the evaluation data was R3, Methods I and II generated the same clustering tree for each of the three sets. In the case of two emotion clusters, one cluster always contained "sleepiness" and



(2-a) Clustering trees using the results performed by infants' mothers, baby-rearing experts, and both of them



(2-b) Clustering trees for three infant sets using the results performed by infants' mothers and baby-rearing experts

Figure 2. *Obtained clustering trees using the results of subjective opinion tests* (SL: sleepiness, PA: pampered, AN: anger, SA: sadness, HU: hunger, DC: discomfort)

"pampered" and the other contained "anger," "sadness," and "hunger." These clusters were also obtained by using Method II when the evaluation set was R2.

Next, we selected three infants with top three test results (approximately 200 samples for each) for our corpus and the clustering experiments were performed on these samples. The clustering results are shown in Figure 2-b. Even though the number of major emotions was different among these infant sets, similar clustering trees were obtained by using Method II when the number of emotion clusters was two.

According to these clustering results, we concluded that Method II, which considers the evaluation rank, is superior to Method I. This is because Method II obtained the more consistent clustering tree structure for each evaluation set than Method I. If we subsequently set two emotion clusters for emotion detection, cluster C1 should consists of emotion elements such as "sleepiness" and "pampered" and cluster C2 should comprise "anger," "sadness," and "hunger." If we set the three clusters, the latter cluster should be divided. One possibility is one of "anger and sadness" and the other of "hunger." In conventional studies [5, 7], the classification experiment between "hunger and sleepiness" or "pampered and anger" was tested. Our clustering results reveal that these emotions belong to different clusters and that these classification trials are rational.

5.2. Recognition with regard to two emotion clusters

According to the clustering results, the preliminary recognition experiment was conducted on two types of emotion clusters—C1 and C2. To define the correct emotion cluster for each sample, we used $P(C1)$ or $P(C2)$ of Eq. (4). If $P(C1)>P(C2)$ for one sample, the correct emotion cluster is C1. The samples in which $P(C1)=P(C2)$ were not used for this experiment. According to this criterion, we prepared 20 samples each for cluster C1 and C2. These cry samples were uttered by one infant (infant A in Figure 2-b). We performed a leave-one-out cross validation on all these data. The results of the experiment are presented in Table 1, where averaged detection performance is 75%.

To evaluate this detection performance, we conducted another recognition experiment on two types of emotions, "anger" and "pampered," using 35 cry samples uttered by the same infant [7]. These two emotions are elements in different clusters in the previous experiment. We assumed the judgment of the infant's mother with the highest rank to be correct. In these samples, the mother identified "anger" as the major emotion in 16 samples and "pampered" in 22 samples. The mother selected multiple emotions for 13 samples and a single emotion for 22 samples. However, none of the samples indicated that the mother had judged the emotions as both "anger" and "pampered." We also performed a leave-one-out cross validation. The agreement between emotion recognition and the judgment by the infant's mother is presented in Table 2. This table shows the agreement rates of the samples for which the mother selected a single emotion and those of the samples for which the mother selected multiple emotions. The agreement rates of the samples for which the mother selected a single emotion are shown to be high, while those of the samples for which multiple emotions were selected are low.

Table 1: *Recognition rates using emotion clusters C1 ("sleepiness" and "pampered") and C2 ("anger," "sadness," and "hunger")*

Cluster	C1 [%]	C2 [%]	Total [%]
Rate	16/20[80]	14/20 [70]	30/40 [75]

Although the scale of these experiments was small and the number of test samples was not the same, these two results suggests that automatic emotion detection using emotion clusters has detection possibilities for the samples for which multiple emotions were selected by infants' mothers. With regard to recognition performance, emotion recognition that uses the emotion clusters compares favorably with the use of judgments made by infants' mothers, thus showing the validity of the emotion cluster.

6. Conclusions

This paper proposed an emotion clustering procedure for emotion detection in infants' cries. Emotion clustering, which was formulated to minimize the objective function concerning the ambiguity of detected emotions, was performed on the results of subjective opinion tests with respect to the emotions expressed in these cries. According to clustering experiments, the proposed merging method, which considered the evaluation rank values on subjective opinion tests, was consistent with the clustering results when the number of emotion clusters was two. Using these clustering results, a recognition experiment was conducted on two types of emotion clusters. With regard to recognition performance, emotion recognition that uses emotion clusters compares favorably with the use of judgments made by infants' mothers.

In the future, to improve emotion recognition performance we will increase our training data and include pitch and prosodic features in addition to spectral features.

7. Acknowledgements

We would like to thank all subjects and their parents for their corporation of this work.

8. References

[1] Green, J. A., et al, "Infant crying: acoustics, perception and communication," *Early Development and Parenting*, vol. 4, pp.1-15, 1995.

[2] Robb, M. P. and Cacace, A. T., "Estimation of formant frequencies in infant cry," *Int. J. Pediatric Otorhinolaryngology*, 32, pp.57-67, 1995

[3] Wermke, K., et al, "Developmental aspects of infant's cry melody and formants," Medical Engineering Physics, 24, pp.501-514, 2002.

[4] Bellieni, C., Sisto, R., Cordelli, D., and Buonocore, A., "Cry features reflect pain intensity in term newborns: an alarm threshold," Pediatric Research, Vol. 55, pp.142-146, 2004.

[5] Arakawa, K., "Recognition of the cause of babies' cries from frequency analyses of their voice classification between hunger and sleepiness," *Proc. International Congress on Acoustics,* pp.1713-1716, 2004.

[6] "Alert and detection device for monitoring the physical status of babies and handicapped persons as well as their usual environment," *Int. Patent,* G08B 19/00, 2000.

[7] Matsunaga, S., et al, "Emotion detection in infants' cries based on a maximum likelihood approach," *Proc. Interspeech 2006*, pp.1834-1837, 2006

Table 2: *Recognition rates between anger and pampered based on the judgment performed by infant's mother*

Emotion	Single [%]	Multiple[%]	Total [%]
Rates	18/22 [82]	6/13 [46]	24/35 [69]