Emotion Detection in Infants' Cries Based on a Maximum Likelihood Approach

S. Matsunaga[1], S. Sakaguchi[1], M. Yamashita[1], S. Miyahara[1], S. Nishitani[2] and K. Shinohara[2]

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

ABSTRACT

This paper proposes a new procedure based on a maximum likelihood approach using hidden Markov models to detect infants' emotions through their cries. Our procedure uses stochastic acoustic models for each kind of emotion. The acoustic models are generated using infant's cries that are labeled segmentally according to their acoustic features. The procedure detects segment sequences with the highest likelihood among all kinds of emotions. The results of our preliminary recognition experiments on two emotions using three types of segment labeling show that the proposed procedure is applicable to emotion detection in an infant's cry and that the detailed transcription of acoustic segments is useful. In this paper, using detailed transcriptions, we broaden the experiments to include five emotions. Assuming the judgment of each infant's mother to be correct, we compared the result of the experiment and that of a subjective opinion test. We conducted the opinion test with the help of three child-rearing experts. Emotion recognition using the proposed procedure displays a favorable comparison with the judgment of our experts, showing the validity of the proposed procedure.
Index Terms: emotion detection, infants' cry, acoustic model

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 emotion that the infants wish to express through the cry (the causes of crying), particularly for those people who do not have enough experience in childcare. Therefore, they may fail to satisfy the infant's wants. In medical science, it is often observed that there are some differences with regard to the expression of emotions through a cry between an infant with cerebral disorders or autism spectrum disorders and a non-disordered infant [2]. In these cases, if the infant's emotions are automatically detectable through a cry, this will ease the situation, as unwanted treatment can be avoided and appropriate medical treatment can be administered to the disordered infants at an early stage.

A number of studies have been conducted on the acoustic analysis of an infant's cry from the viewpoint of emotion detection [3, 4]. An infant's pain is one of the traditional research topics for the detection of emotions [5]. In recent years, "hunger" and "sleepiness" have also been studied [6]. These researches apply the heuristic rules that are based on the threshold of spectral features. Furthermore, some emotion-detection products are also available in the market [7]. These employ simple matching techniques using acoustic features. In these conventional studies, the features observed in the cry over a short period of time (few seconds) were investigated. However, it is very difficult to identify the characteristic portion that helps clarify the emotions expressed in a cry. Thus far good performance has not been achieved.

In order to achieve improved performance, we propose a new procedure to detect infants' emotions that is based on a maximum likelihood approach. We consider the entire cry as a time sequence of acoustic segments and assume that the characteristics of emotions are scattered throughout the cry. The procedure comprises a training process and a test process. In the training process, the acoustic models of each segment unit are generated for each kind of emotion. Hidden Markov models (HMMs) are used in this modeling. In the test process, the acoustic likelihood of an input cry is calculated using the acoustic models, and the segment sequence with the highest likelihood from among all kinds of emotions is detected. In this paper, we first examine how detailed the segment labels should be in order to detect emotions correctly. Next, we conduct a subjective test to investigate how well child-rearing experts can recognize the emotions in an infant's cry. Finally, assuming the judgment of each infant's mother to be correct, we compare the results of the experiment and the subjective test.

2. EMOTIONS IN A CORPUS OF INFANTS' CRIES

We prepared the corpus of infant cries to evaluate the performance of emotion detection. The age of the infants in this study is around ten months. All of them are non-disordered infants. Our corpus comprises the waveform data, the tag of emotions, transcriptions using the labels of the acoustic segments, and the information regarding the time period in which noise is included.

2.1 Cry corpus and tag of emotions

All the mothers were required to record their infants' cries

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

Figure 1: Example of the emotion table

over several days at home using a digital recorder. A total of 402 cries by 23 infants were recorded (11 male infants and 12 female infants). The average duration of the recorded data was approximately 30 seconds. The infants ranged from 8 to 13 months of age, and the average age was 10.6 months.

After recording each cry, the infants' mothers judged the emotions expressed in the samples. 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. 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

The mothers chose the following five emotions most frequently: pampered, anger, sadness, hunger, and sleepiness. These accounted for approximately 95% of all the samples in the corpus.

2.2 Hand labeling

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 w conprises N segments, and let the i-th segment be s_i $(1 \le i \le N)$,

$$w = s_1 s_2 \cdots s_i \cdots s_N , \qquad (1)$$

where the beginning time of segment s_{i+1} is the end time of segment s_i.

In order to examine how detailed segmentation should be carried out in order to recognize the emotions, we used three types of segmentations: Label-1, 2, and 3. Label-1 comprises only a silent segment and a sound segment. The sound segment in Label-2 is divided into two kinds of segments: an utterance segment and a breath segment. In Label-3, the utterance segment is divided into several segments: a glottal sound segment (a cry that sounds like a cough), a typical cry segment, a babbling segment, a cooing segment, and so on. These segments are easily recognizable by humans. The hierarchical construction of these labels is shown in Figure 2.

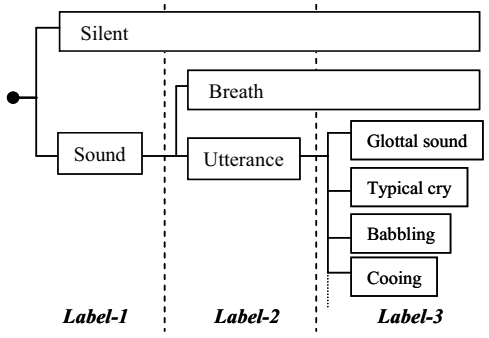

Figure 2: Hierarchical structure of the segment labels

Table 1: Emotion rank by experts and agreement rate

Emotion rank	4	3 or more	2 or more
Agreement rate [%]	41	32	29

3. SUBJECTIVE OPINION TEST

We conducted a subjective opinion test with the help of three baby-rearing experts. Their judgments were made solely on the basis of the recordings. The purpose of this test was to investigate how well the experts could recognize the emotions in a cry. A subjective evaluation was performed in the same way as the judgment of the mothers, as described in section 2.1. We also provided the experts with some examples to explain each emotion. We used 100 recordings uttered by five infants. Each cry includes one of the five most frequently chosen emotions described in section 2.1. The agreement between the emotions judged by mothers with the highest rank and the emotions detected by experts is shown in Table 1. The three agreement rates are shown for each emotion rank judged by the experts. For example, the agreement rate for the samples that the experts judged as having the evaluation rank 4, is 41%. This table shows that with regard to the samples that the experts judged, if the emotions are included completely, the agreement rates are higher. This means that there exists the possibility of recognizing the emotions in a cry automatically. Table 2 shows the content by percentage c and the average number of the kinds of emotions n that were selected by the experts for each cry. The emotions in the leftmost column are those that were detected by the infants' mothers with the highest rank, and the emotion ranks in the topmost row are those that were judged by the experts. The content by percentage c indicates that $c = 100 \times A / M$ [%].

Here, A is the number of samples for which there was an agreement between the mothers and the experts, and M is the number of the samples for each emotion judged by the mother. For example, the childcare experts selected 0.37 kinds of emotions as having the rank 4; these samples were the ones that the mothers had judged as "anger" with the highest rank. The experts detected the same emotion of "anger" with the rank 4 in 22% of the samples that the mothers had judged as "anger." The table shows that the detection of the emotion of "anger" is comparatively easy, while "sadness" is a difficult

Table 2: Content by percentage c and average number of selected emotions n by experts

Emotion rank [experts] / Emotion [mothers]	4		3 or more		2 or more	
	c	n	c	n	c	n
Pampered (17)	4	0.02	25	1.1	59	1.9
Anger (18)	22	0.37	46	1.0	63	1.7
Sadness (11)	9	0.27	15	0.76	18	1.3
Hunger (18)	2	0.17	28	0.96	50	2.0
Sleepiness (36)	9	0.21	31	0.98	54	2.0
Average (100)	9.3	0.23	30.7	0.97	51.7	1.8

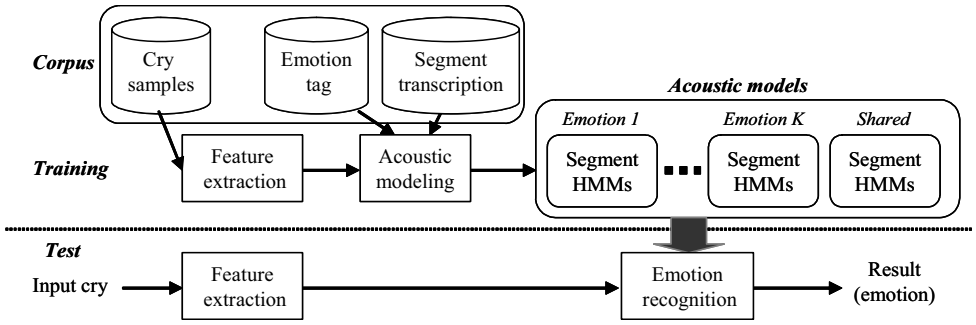

Figure 3: Architecture of emotion recognition system

emotion to detect. It is also shown that even if the experts select approximately 1.8 emotions from among the five emotions per sample, the agreement rate (content by percentage) remains 51.7%.

4. EMOTION RECOGNITION ALGORITHM

We assume that the features of emotions can be detected more reliably in an infant's complete cry, rather than in a specific portion of the cry. In our approach, the acoustic likelihood of the input, which represents the infant's complete cry, is calculated by using the acoustic models for each kind of emotion. Accordingly, the segment sequence with the highest likelihood is detected among all kinds of emotions.

The complete cry w was expressed by the series of the acoustic segments s in section 2.2. Given acoustic evidence observation q, the proposed process of emotion recognition is to find the most likely segment sequence, $\hat{w}$, and the emotion $\hat{e}$ which gives $\hat{w}$, satisfying

$$P(\hat{e}, \hat{w} \mid q) = \max_{e,w} P(e, w \mid q) \quad (2)$$

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

$$P(e, w \mid q) = \frac{P(e, w) P(q \mid e, w)}{P(q)}, \quad (3)$$

where $P(e, w)$ is a priori probability that the segment sequence w will be occurred on the emotion 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 $P(e, w)$ in detecting the emotion $\hat{e}$. $P(q \mid e, w)$ is the probability that when the infant utters the sequence w caused by the emotion e the acoustic evidence q will be observed. Since $P(q)$ is not related to w and e, it is irrelevant to recognition. Then, we can apply the emotion recognition procedure to Eq. (2) as follows:

Table 3: Acoustic modeling for segment sets

Segment / Labels	Sound segment	
	Breath	Utterance
Label-1	One model for each emotion	
Label-2	One for each	One for each
Label-3	One for each	3 models for each
Label-2-1	One shared model	One for each
Label-2-2	One for each	One shared model

$$\hat{e}, \hat{w} = \arg\max_{e,w} P(e, w \mid q) \approx \arg\max_{e,w} P(q \mid e, w) \quad (4)$$

In this paper, we perform this maximization by using the likelihood of HMMs.

5. RECOGNITION EXPERIMENTS

5.1 Emotion recognition system

The architecture of our emotion recognition system is shown in Figure 3. The system comprises a training process and a test process. Acoustic feature parameters were extracted in the feature extraction module. Acoustic models of each segment are generated for each kind of emotion in the training process. In the test process, acoustic likelihood of an input cry is calculated by using these acoustic models, and the emotion $\hat{e}$ which gives the segment sequence $\hat{w}$ with the highest likelihood is detected.

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 8-mixture, context-independent models. These models were generated for each segment and each emotion (emotion-dependent). A silent model was shared among emotions (emotion-independent).

5.2 Recognition with regard to two emotions

The preliminary recognition experiment was conducted on two kinds of emotions, "anger" and "pampered," using 35 cry samples uttered by one infant. In these data, the mother identified the 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, for none of the samples had the mother judged the emotions as both anger and pampered. We performed a leave-one-out cross validation on all these data.

We used the three sets of acoustic models based on the labels. These sets (corresponding to Label-1, 2, and 3 in section 2.2) are shown in Table 3. The Label-1 set comprises a shared silent model and an emotion-dependent sound model. In the Label-2 set, a shared silent model, an emotion-dependent breath model, and an emotion-dependent utterance model are included. The Label-3 set includes a shared silent model, an emotion-dependent breath model and 3 emotion-dependent utterance models (a typical cry model, a glottal sound model, and a garbage model).

The agreement between emotion recognition on the basis of the proposed procedure and the judgment by the infant's

mother is shown in Table 4. 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. This means that automatic emotion detection is very difficult if the infant's cry is a result of multiple emotions. With regard to the samples for which the mother selected a single emotion, the use of a set of segments with detailed labels (Label-3) achieved better agreement rates.

Table 4: Agreement rates for three types of segment labels

Label\Emotion	Single [%]	Multiple[%]	Total [%]
Label-1	16/22 [73]	6/13 [46]	22/35 [63]
Label-2	17/22 [77]	7/13 [54]	24/35 [69]
Label-3	18/22 [82]	6/13 [46]	24/35 [69]

5.3 Effectiveness of emotion-dependent models

In addition to using the Label-2 set, we conducted the recognition test using two other sets, i.e., Label-2-1 (including one shared breath model and one emotion-dependent utterance model) and Label-2-2 (including one emotion-dependent breath model and one shared utterance model). Preparing these three sets in this manner, we compared the difference in the effectiveness of using the emotion-dependent utterance and breath models separately and together.

The results of the experiments are shown in Table 5. Both the breath and utterance models should be generated for each emotion to achieve improved discrimination between "anger" and "pampered." This confirms the validity of our assumption that the acoustic features of emotions can be detected more reliably in an infant's complete cry, rather than in a specific portion of the cry.

Table 5: Agreement rates using emotion-dependent models

Label\Emotion	Single [%]	Multiple [%]	Total [%]
Label-2	17/22 [77]	7/13 [54]	24/35 [69]
Label-2-1	11/22 [50]	8/13 [62]	19/35 [54]
Label-2-2	12/22 [55]	8/13 [62]	20/35 [57]

5.4 Recognition experiments among five emotions

Finally, we conducted the recognition experiment was using 55 samples uttered by one infant. We performed a leave-one-out cross validation on all these data. In these samples, the major emotion judged by the mother was one of the following five emotions: "pampered," "anger," "sadness," "hunger," and "sleepiness." In the case of 12 samples, the mother selected two of theses five emotions as having the highest rank, and if one of them was identical to the recognition result, we considered her selection and the result to be in agreement. Table 6 shows the agreement rates of the emotions judged by the mother and the highest likelihood emotion candidate or one of the top 2 emotions candidates detected in the experiment.

Comparing this result with that of the subjective opinion test from the viewpoint of the number of candidates, the figure 51% in the Top 1 column in this table nearly corresponds to the figure 30.7% (Emotion: Average, Emotion rank: three or more) in Table 2 where the number of selected emotions is 0.97. Moreover, the figure 75% in the Top 2 column nearly corresponds to the figure 51.7% (Emotion: Average, Emotion rank: two or more) in Table 2 where the number of selected emotions is 1.8. Although the scale of this experiment is small and the test samples are not the same, these results show that emotion recognition using the proposed procedure compares favorably with the judgment by our baby-rearing experts, thus showing the validity of our procedure.

Table 6: Recognition experiments among five emotions

Candidate	Top 1	Top 2
Agreement rate [%]	28/55 [51%]	41/55 [75%]

6. CONCLUSIONS

This paper proposed a procedure based on a maximum likelihood approach using HMMs to recognize infants' emotions through their cries. This procedure was based on the assumption that the features of emotions can be detected more reliably in an infant's complete cry, rather than in a specific portion of the cry. The acoustic likelihood of an input cry was calculated, and the segment sequence with the highest likelihood from among all kinds of emotions was detected. Emotion recognition using the proposed procedure compared favorably with the judgment by three baby-rearing experts.

Future work will include pitch and prosodic features in addition to spectral features to improve emotion recognition performance.

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] Orozco, J. and Carcia, A., "Detecting pathologies from infant cry applying scaled conjugate gradient neural networks," *Proc. European Symposium on Artificial Neural Networks*, pp.349-354, 2003.

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

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

[5] 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.

[6] 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.

[7] "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.