Emotion Classification of Infants' Cries using Duration Ratios of Acoustic Segments

K. Kitahara[1], *S. Michiwiki*[1], *M. Sato*[1], *S. Matsunaga*[1], *M. Yamashita*[1] *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
{b606231, b603443, b607244, mat, masaru}@cis.nagasaki-u.ac.jp

10.21437/Interspeech.2011-473

Abstract

We propose an approach to the classification of emotion clusters using prosodic features. In our approach, we use the duration ratios of specific acoustic segments—resonant cry segments and silence segments—in the infants' cries as prosodic features. We use power and pitch information to detect these segment periods and use normal distribution as a prosodic model to approximate the occurrence probability of the duration ratios of these segments. Classification experiments on two major emotion clusters are carried out using samples of recorded cries of 11 infants. When the detection performance for the segment periods is about 75%, an emotion classification rate of 70.8% is achieved. The classification performance of our approach using the segment duration ratios was significantly better than that of the classification method using power and spectral features, thereby indicating the effectiveness of using prosodic features. Furthermore, we describe a classification method using both spectral and prosodic features with a slightly better performance (71.9%).

Index Terms: emotion classification, infants' cry, prosodic feature

1. Introduction

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

From the viewpoint of emotion detection, several studies have focused on the acoustic analysis of infants' cries [2, 3]. An infant's pain is one of the traditional research topics in the detection of emotions [4]. In addition, the classification between "hunger" and "sleepiness" has been studied in recent years [5]. These studies employ simple matching techniques based on spectral features.

We made recordings of infants' cries (hereinafter referred to as "cry samples") and developed an emotion detection procedure based on the maximum likelihood approach using hidden Markov models (HMMs) [6]. With the assistance of the infants' mothers and baby-rearing experts, we conducted subjective opinion tests on the emotions expressed in each sample. In our study, these mothers and experts frequently selected two or more emotions for each cry sample and the characteristics of the detected emotions were very different. Further, the rate of agreement between the mothers and the experts for the detected emotions was not high. However, they tended to detect the same set of emotions in the course of listening to many cry samples. We then proposed the use of "emotion clusters" generated by an emotion clustering procedure in order to cope with the ambiguity of the detected emotions [7]. In our previous studies, a high level of emotion classification could not be achieved using power and spectral features as acoustic features.

In an attempt to address the abovementioned problems, we propose an approach to emotion cluster classification using prosodic features. According to our investigation of cry samples with different prosodic features, we found that there is no difference in the average pitch frequency between two major emotion clusters. However, a remarkable difference is observed in the following two duration ratios: the total duration of all the resonant cry segments and the total duration of all the silence segments in a cry sample. Then, we use the ratios of the duration of these segments to the duration of the entire cry sample for emotion cluster classification. The duration ratios for each cluster are described using a normal (Gaussian) distribution. Classification experiments on emotion clusters are carried out using statistical formulation to compare the classification performances for acoustic features using (1) prosodic features, (2) spectral features and (3) a combination of prosodic and spectral features.

2. Corpus of infants' cries

Our corpus of infants' cries comprises waveform data, tag of emotions detected by the mothers and baby-rearing experts, and transcriptions that use the labels of the acoustic segments [7].

2.1. Subjective opinion tests

Eleven mothers were required to record their infants' cries over several days at home using a digital voice recorder. We collected a total of 342 cry samples from 11 infants for the experiments. The average duration of the recorded data was approximately 30 seconds. The age of the infants ranged from 8 to 13 months.

After recording each cry, the infants' mothers and three

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

Figure 1: *Example of the emotion table*

Copyright © 2011 ISCA

baby-rearing experts judged the emotions expressed in the samples (subjective opinion test). During the test, the mothers took into consideration not only the cries but also the infants' facial expressions, behaviors, etc. Ten types of emotion tags were defined: contentment (psychological dependence), anger, sadness, fear, surprise, hunger, sleepiness, excretion, discomfort, and pain. They assessed the emotions that caused the infants to 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 was ranked on a scale ranging from 0 (the emotion is not contained at all) to 4 (the emotion is contained fully) in this table. They were permitted to choose two or more emotions.

2.2. Hand labeling of acoustic segment

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

$$z = s_1 s_2 \cdots s_i \cdots s_N , \quad (1)$$

The labels were defined using the following five types of acoustic segments: a silent segment (28%), an inspiratory sound segment including hiccough (14%), a glottal sound segment (a cry that sounds like a cough, 9%), a resonant segment (a harmonic cry and a spasmodic cry, 43%), and a miscellaneous segment (babbling, cooing, etc., 6%) [7].

2.3. Two emotion clusters based on results of subjective opinion test

For each cry sample, a mother or baby-rearing expert selected one or more emotions from among ten types of emotions. It is supposed that when she selected one emotion e_i from among the emotion cluster set $E = \{e_1, \cdots, e_I\}$, she also simultaneously selected emotion e_j $(i \neq j)$. Our clustering algorithm is based on the criterion of maximum reduction of the ambiguity of selection of emotions/emotion-clusters, taking into account the formulation concerning conditional entropy. The emotion clusters e_i and e_j, which maximize the reduction in ambiguity, are selected and merged into one emotion cluster. A detailed formulation of this clustering is described in [7].

Clustering was performed using approximately 1200 results of subjective opinion tests for about 300 cry samples collected from 10 infants. According to the clustering results, if we subsequently set two emotion clusters for emotion detection, cluster C1 would consist of emotional states such as "sleepiness" and "contentment", and cluster C2 comprise "anger," "sadness," and "hunger." We use these two emotion clusters in the following classification experiments.

Using the clustering criterion [7] and the subjective opinion test results, all the cry samples were divided into three sub-sets: C1, C2, and "others." The "others" set includes samples that cannot be classified clearly as C1 or C2 because of the similarity in the test results for each cluster or do not belong to those clusters. The samples in C1 and C2 are used for emotion cluster classification, and the number of samples for each cluster is shown in Table 1.

Table 1. Number of cry samples in emotion clusters C1 ("sleepiness" and "contentment") and C2 ("anger," "sadness," and "hunger") for evaluation sets

Emotion cluster	C1 [%]	C2 [%]
Number	127 [50.2]	126 [49.8]

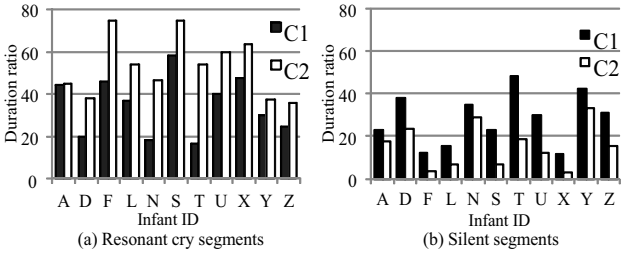


Figure 2: *Duration ratios for each segment type and infant*

Table 2. Average (standard deviation) of pitch frequency for each emotion cluster [Hz]

Emotion cluster	C1	C2
Average (s.d.)	406 (108)	410 (113)

3. Prosodic features in infants' cries

3.1. Duration ratios as prosodic features

Infants' cries include a variety of prosodic features such as pitch information. The pitch frequency is often used to detect emotions in adult speech [8]. Hence, we investigated the pitch frequency for each emotion cluster C1 and C2 using our database. The results are shown in Table 2; "average" indicates the averaged frequency for all the frames where pitch can be detected in the cry samples. The results show that there is no difference in the averaged frequency among the two emotion clusters.

Next, we investigated the duration ratios of the five types of segments for each emotion cluster C1 and C2 and we found a consistent difference in the duration ratios of the silent segments and resonant segments. Figure 2(a) shows the duration ratios of the resonant cry segments for each emotion cluster, and Figure 2(b) shows those of the silent segments for each emotion cluster. The duration ratio for each infant, from A to Z, is shown in these figures. It can be seen that for each infant, the duration ratio of the resonant cry segments for emotion cluster C1 is smaller than that for cluster C2. In contrast, the duration ratio of the silent segments for C1 is larger than that for C2. The results also show that the difference in the duration ratios among the infants is considerably large.

3.2. Modeling of duration ratios

The histograms of duration ratios of resonant segments for each cluster are shown in Figure 3. These histograms were obtained using all 253 samples. As described before, the mean value of the duration ratio of the resonant segments in emotion cluster C1 (a) is smaller than that in cluster C2 (b). The opposite phenomenon is observed for the silent segments. Table 3 shows the mean values and standard deviation for

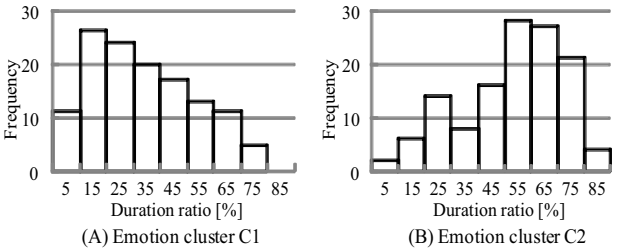


Figure 3: *Histograms of duration ratios of resonant segments*

each emotion cluster and each segment type. In this paper, normal distribution is adopted to describe the occurrence probability of duration ratios using these figures (mean and standard deviation). In the classification test, the probability density function for each cluster and for each segment is used to calculate the prosodic likelihood for the input data.

4. Emotion classification methods

4.1. Formulation of emotion classification

We formulate the classification of emotion clusters as follows. Given the acoustic evidence observation q, our process of emotion detection seeks to find the most likely segment sequence, $\hat{z}$, and the emotion cluster $\hat{e}$ that gives $\hat{z}$, satisfying

$$P(\hat{e},\hat{z}\mid q)=\max_{e,z}P(e,z\mid q)\,. \tag{2}$$

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

$$P(e,z\mid q)=P(e)P(z\mid e)P(q\mid e,z)/P(q) \tag{3}$$

where $P(e)$ is the a priori occurrence probability of emotion cluster e. According to our investigation results of the cry samples, if the number of data samples for two emotion clusters are almost equal, then we can assume that the probability $P(e)$ is irrelevant in our experiments. Moreover, the term $P(q)$ is not related to z and e, and therefore, it is also irrelevant.

$P(z|e)$ represents the occurrence probability that segment sequence z will occur in emotion cluster e. One of the main characteristics of our approach is that this term is calculated using prosodic features: the segment duration ratio for each cluster is derived using power and pitch.

The term $P(q\mid e,z)$ is the probability that when an infant utters a sequence z caused by emotion cluster e, acoustic evidence q will be observed. In this paper, this term is calculated using HMMs wherein spectral features (mel-cepstrum) and power are used as feature parameters.

Thus, we can apply the emotion detection procedure to Eq. (2) as follows:

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

Method 1 using prosodic feature

According to the abovementioned meaning of each term, we adopt the following equation in the classification using only prosodic features:

$$\hat{e}=\arg\max_{e}\left(\log P(e,\tilde{z}\mid q)\right)\approx\arg\max_{e}\left(\log P(\tilde{z}\mid e)\right) \tag{5}$$

$$\log P(\tilde{z}\mid e)\approx\alpha\log P\left(r_c\mid e\right)+\log P\left(r_s\mid e\right) \tag{6}$$

where r_c and r_s are the duration ratios for the resonant cry and silence segments, respectively, and they are detected using power and pitch in the segment sequence $\tilde{z}$. The probability for each segment type is calculated using the normal distribution function for each emotion cluster. α is a weighting factor of the contribution of each segment.

Method 2 using spectral feature

When we use the power and spectral information as acoustic features in our experiments, we use the following formulation to detect the emotion cluster [6]:

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

We perform the maximization in Eq. (7) by using the likelihood of HMMs. These models of each segment are generated for each type of emotion cluster.

Table 3. Average of duration ratios for each emotion cluster (standard deviation) [%]

Cluster\Segment	Resonant cry	Silence
C1	33.6 (19.5)	29.6 (17.8)
C2	53.0 (19.3)	16.3 (15.0)

Method 3 using prosodic and spectral features

Finally, using Eqs. (4), (5) and (7), we try to detect the emotion cluster using both prosodic features and spectral features as follows:

$$\begin{aligned}\hat{e}&=\arg\max_{e}\left(\log P(e,\tilde{z}\mid q)\right)\\&\approx\arg\max_{e}\left(\log P(\tilde{z}\mid e)+\beta\log P\left(q\mid e,\hat{z}\right)\right)\end{aligned} \tag{8}$$

β is a weighting factor of contribution of the spectral information and prosodic information.

4.2. Duration detection using prosodic features

In our emotion detection approach, it is desired that the segment periods should be detected accurately to calculate the prosodic likelihood for each cluster. We use the following decision rules for each analysis frame in a cry sample using pitch and power information:

1. If the power is larger than a pre-defined threshold *th*1 and pitch can be detected on this frame, this frame is regarded as a resonant segment frame.
2. If the power is less than a pre-defined threshold *th*2 and pitch is not detected on this frame, this frame is regarded as a silent segment frame.

In our experiments, we use the pitch extraction tool coded by Mr. Kazunori Imoto, Kyoto University [9].

5. Classification experiments and results

The cry data were sampled at 16 kHz. Every 10 ms, the power and pitch frequency were computed using a Hamming window.

5.1. Segment detection experiment

A detection experiment of resonant and silent segments using the method described in Section 4.2 is performed. All 253 cry samples of emotion clusters C1 and C2 were used, and the thresholds *th*1 and *th*2 were determined experimentally. The agreement rate per frame between the detected segments and hand-labeled segments is shown in Table 4. This table shows that our method achieved about 75% detection rate. The main types of misclassification are the confusion between inspiratory sounds and silence and that between resonant sounds and miscellaneous segments.

5.2. Classification using segment duration ratios

A classification experiment was carried out using the obtained segment duration ratios. Before the classification, hand-labeled data of 253 samples, from which test data was excluded, were used to train a normal distribution function for each emotion cluster and for each segment type. We performed leave-one-out cross-validation based on Eqs. (5) and (6) (Method 1), and the weight α is set to 0.23. The

Table 4. Segment detection performance per frame using power and pitch information[%]

Segment	Precision	Recall	F-measure
Resonant cry	72.4	81.4	77.0
Silence	64.8	82.9	73.9

Table 5. Classification performance using acoustic models with prosodic features (Method 1) [%]

Emotion cluster	C1	C2	Average
Performance	70.9	70.6	70.8

Table 6. Classification performance (upper bound) using hand-labeled data (Method 1) [%]

Emotion cluster	C1	C2	Average
Performance	68.5	76.2	72.3

Table 7. Classification performance using acoustic models with power and spectral features (Method 2) [%]

Emotion cluster	C1	C2	Average
Performance	68.5	57.1	62.8

Table 8. Classification performance using both prosodic and spectral features (Method 3) [%]

Emotion cluster	C1	C2	Average
Performance	75.6	68.3	71.9

classification performance was 70.8%, as shown in Table 5. This shows the effectiveness of using prosodic features of segment duration ratios to classify the emotions in an infant's cry.

Next, the upper bound performance of our proposed approach was investigated using the hand-labeled data (Table 6), where α is set to 1.05 to achieve the highest performance. Comparing Tables 5 and 6, our approach achieved a performance near the upper bound (72.3%) even though the segment detection rates are still around 75%. We suppose that infant-dependent modeling of duration ratios is needed to enhance the upper-bound performance.

An emotion cluster and most likely segment sequence are obtained simultaneously using Eq. (7). On the bases of this formulation (Method 2), we have previously proposed an emotion detection technique based on the maximum likelihood approach using power and spectral features [6]. To compare this technique and the proposed approach that uses power and prosodic features, we performed a classification test on same evaluation data set using power and spectral information. In this experiment, the power and a vector of 12 FFT mel-warped cepstral coefficients (spectral features) were computed and used to generate the segment HMMS that were 3-state 2-mixture models. These models were generated using the cry samples uttered by 11 infants for each segment type and each emotion-cluster. The classification performance is shown in Table 7. Comparing Tables 5 and 7, it can be seen that the performance using prosodic information is far better than that of our previous method using spectral features.

5.3. Classification using prosodic and spectral features

In the two classification experiments described in Section 5.2, only the segment duration ratios or spectral information were used. Next, we performed a classification experiment to confirm the effectiveness of using both the prosodic and spectral features. In our approach, the formulation is based on Eq. (8) (Method 3), where $P(q \mid e, \hat{z})$ is calculated using spectral and power information and $P(z \mid e)$ is obtained using segment ratios. A classification experiment was carried out under the same condition as the other experiments. The value of the weighting factor β was determined experimentally. The classification performance is shown in Table 8. Comparing Tables 5 and 8, it can be seen that the average classification performance increased slightly $(70.8\% \rightarrow 71.9\%)$; however, this increase is not significant. This result shows that the segment duration ratios plays major role in the classification of two emotion clusters C1 and C2, and a more effective method using a combination of prosodic and spectral features is needed for higher performance.

6. Conclusions

This paper proposed an approach to emotion cluster classification using segmental prosodic information. We used the ratios of the duration of resonant segments and silent segments to the duration of the entire cry sample as the prosodic feature. Normal distribution was used to model the distribution of segment ratios for each emotion cluster. The durations of these segments were estimated using the pitch and power information in the infants' cries. Even though the detection rates of the segment durations are around 75%, our proposed approach achieved 70.8% of classification rate. We showed that this new method outperformed the classification approach that used power and spectral information, indicating the usefulness of using segment duration ratios. Finally, we devised a classification method using prosodic and spectral features. Although the classification performance increased slightly, it is a future task to enhance the performance significantly by developing a model using prosodic and spectral features.

In future, we need to evaluate the effect of segment duration ratios for emotion cluster classification when the number of emotion clusters is high.

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

[7] Satoh, N., et al, "Emotion clustering using the results of subjective opinion tests for emotion recognition in infants' cries," *Proc. Interspeech* 2007, pp.2229-2232, 2007

[8] Schuller, B., et al, "The INTERSPEECH 2009 emotion challenge," *Proc. Interspeech* 2009, pp.312-315, 2009

[9] http://vision.kuee.kyoto-u.ac.jp/lecture/dsp /?menu= c_pitch