Emotion Classification of Infant Cries with Consideration for Local and Global Features

Kazuki Honda, Kazuki Kitahara, Shoichi Matsunaga, Masaru Yamashita and Kazuyuki Shinohara
Nagasaki University, Nagasaki, Japan
E-mail: {b608416, b606231, mat, masaru} @cis.nagasaki-u.ac.jp Tel: +81-95-8192700

Abstract— **In this paper, we propose an approach to the classification of emotion clusters in infant cries with consideration for frame-wise/local acoustic features and global prosodic features. Our proposed approach has two main characteristics as follows. The emotion cluster detection procedure is based on the most likely segment sequence, which delivers the emotion cluster as a classification result. This is obtained based on a maximum likelihood approach using the frame-wise likelihood and the global prosodic likelihood. We exploit the duration ratios of resonant cry segments and silent segments as prosodic features, while the duration ratios are calculated using the derived segment sequence. The second characteristic is the use of pitch information, in addition to conventional power and spectral information, during the modeling of frame-wise acoustic features with hidden Markov models. The classification performance (74.7%) of our proposed approach with added pitch information was better than (71.5%) the classification method using only power and spectral features. The proposed method based on a maximum likelihood approach using both frame-wise and global features also achieved better performance (75.5%).**

I. INTRODUCTION

Infants convey their needs to their parents and/or carers by crying, which has a particularly important role in urgent communication [1]. However, it is difficult to determine the emotions that infants express through their cries (i.e., the causes of crying), particularly for people who are inexperienced in childcare. In fact, no one can know an infant's true emotion. However, mothers and carers with sufficient experience in childcare can sense an infant's emotion more correctly based on their cries.

Based on this empirical knowledge, we determined that the aim of our research would not be to detect the true emotion of infants directly, but instead we aimed to detect the emotions that experienced mothers and carers commonly sensed in infant cries. We made recordings of infant cries and conducted subjective opinion tests to determine the emotions expressed in each sample with the assistance of the mothers of infants and baby-rearing experts. In the tests, mothers and baby-rearing experts frequently selected two or more emotions for each cry sample and the rate of agreement among them for the detected emotions was not high. However, we found that they tended to detect the same set of emotions when listening to many cry samples. We then developed a method for detecting "emotion clusters" generated by an emotion clustering method using the results of the subjective opinion tests [2].

Several previous emotion detection studies have focused on an acoustic analysis of infant cries [3, 4]. Infant pain is a traditional research area in the detection of emotions [5]. The distinction of "hunger" and "sleepiness" cries has also been studied in recent years [6]. These studies used simple matching techniques based on power and spectral features.

We developed an emotion detection procedure that was based on a maximum likelihood approach using hidden Markov models (HMMs) [7], which also used power and spectral features. The pitch frequency is useful for detecting emotions in adult speech [8]. However, there is a problem with using pitch information to detect emotions in infant cries. Our procedure only used frame-wise/local acoustic features, so we proposed an approach for emotion cluster classification where the duration ratios of resonant cry segments and silent segments were used as global acoustic features to improve the classification performance [9]. This was because we observed a significant difference in the two duration ratios among emotion clusters. However, this approach was practical but it was not based on the maximum likelihood approach. There was a theoretical problem that the duration ratios derived from HMMs and the duration ratios obtained from the global acoustic features were not identical.

To address these problems, we propose an emotion cluster classification approach based on the maximum likelihood approach with consideration for both frame-wise acoustic information and the segment information of duration ratios. We used the pitch frequency as an acoustic feature parameter in the modeling of HMMs to obtain the most likely segment sequence with greater certainty. This is because pitch information is useful for detecting resonant cry segments or silent segments. Classification experiments are conducted with two emotion clusters based on a statistical formulation to evaluate the utility of pitch information as the acoustic parameter and to evaluate the proposed maximum likelihood approach using both the frame-wise likelihood and the prosodic likelihood obtained from the segment duration ratios.

II. CORPUS OF INFANT CRIES

Our corpus of infant cries comprised waveform data, tags for the emotions detected by mothers and baby-rearing experts, and transcriptions using the labels on acoustic segments [7].

A. Subjective Opinion Tests

Eleven mothers were required to record infant cries at home for several days using a digital voice recorder. It was inevitable that a variety of noises would be mixed during the

TABLE I
RATIOS OF THE NUMBERS OF EACH TYPE OF SEGMENT [%]

Silence	Inspiration	Glottal	Cry	Miscellaneous
28	14	9	43	6

TABLE II
NUMBER OF CRY SAMPLES IN EACH EMOTION CLUSTER C1 ("SLEEPINESS" AND "CONTENTMENT") OR C2 ("ANGER," "SADNESS," AND "HUNGER")

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

TABLE III
DETECTION RATE OF PITCH FREQUENCY [%]

Silence	Inspiration	Glottal	Cry	Miscellaneous
5	38	48	75	36

TABLE IV
AVERAGE DURATION RATIOS μ (STANDARD DEVIATION σ) [%]

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

recording. We collected a total of 342 cry samples from 11 infants to use in the experiments. The average duration of the recorded data was approximately 30 s and the age of the infants ranged from 8 to 13 months.

After recording each cry, the infant's mothers and three baby-rearing experts judged the emotions expressed in the samples (subjective opinion test). During the test, the mothers considered the cries and the facial expressions, behaviors, etc. of their infants at the time of the recording. 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 they considered to be correlated with the crying of infants and these assessments were recorded in a table. The intensities of the emotions were ranked on a scale ranging from 0 (no emotional content) to 4 (full emotional content) in the table. Subjects were allowed to select two or more emotions.

B. Hand Labeling of Acoustic Segments

We considered that a cry was composed of segments with specific acoustic characteristics. 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, whereas the noises were not labeled. If we suppose that cry $\mathbf{z}$ is comprised of n segments and we let the i-th segment be s_i $(1 \le i \le n)$, then

$$\mathbf{z} = s_1 s_2 \cdots s_i \cdots s_n \ . \quad (1)$$

The labels were defined using the following five types of acoustic segments: a silent segment, an inspiratory sound segment including hiccough, a glottal sound segment (a cry that sounds like a cough), a resonant cry segment (a harmonic cry and a spasmodic cry), and a miscellaneous segment (babbling, cooing, etc.) [7]. The relative ratios of the numbers of each segment in the corpus are shown in Table I. The resonant cry and silence were the two major segment types.

C. Two Major Emotion Clusters for Classification

Automatic emotion clustering was performed using approximately 1200 results from the subjective opinion tests with about 300 cry samples collected from 10 infants. Based on the clustering results [2], we specified two emotion clusters for emotion detection, i.e., cluster C1 consisted of emotional states such as "sleepiness" and "contentment", while cluster C2 comprised "anger," "sadness," and "hunger." We used these two emotion clusters in the subsequent experiments.

Using the clustering criterion and the subjective opinion test results, we divided all the cry samples into three sub-sets: C1, C2, and "others." The "others" set includes samples that could not be classified clearly as C1 or C2 because of similar test results for each cluster or because they did not belong to those clusters. The samples in C1 and C2 were used for emotion cluster classification, while the number of samples in each cluster is shown in Table II.

III. ACOUSTIC FEATURES FOR CLASSIFICATION

A. Pitch Information as Acoustic Parameter

Infant cries include a variety of prosodic features such as pitch information. However, it is difficult to capture the pitch frequency perfectly in human utterances. Furthermore, there were many periods where pitch was not detected in the cry samples. Thus, we investigated the pitch frequency detection rate for each segment type using our corpus. In our experiments, we used a pitch extraction tool coded by Mr. Kazunori Imoto, Kyoto University [10]. The results are shown in Table III. Each figure represents the ratio of the number of frames where pitch was detected relative to the total number of frames for each segment type. The pitch frequency was not detected in about 25% of the frames in the cry segments whereas the pitch frequency was detected as a mixture of noises in about 5% of the frames in the silent segments. These results showed that deterministic rules were not applicable to acoustic segment detection. Thus, we introduced the pitch frequency as an acoustic parameter for the stochastic modeling of segment HMMs. If the pitch was not detected in a sample frame, the pitch value was set to 0 for the frame.

B. Modeling of Duration Ratios

Based on our analysis of the cry samples, we identified the following characteristic of the duration ratios of segments [9]: the mean value of the duration ratio of the resonant segments in emotion cluster C1 was lower than that in cluster C2. The opposite effect was observed in the silent segments. Table IV shows the mean value μ and standard deviation σ for each emotion cluster and each segment type. A normal distribution $N(\mu,\sigma)$ was used to describe the occurrence probability of duration ratios with these values. In the classification test, the duration ratios were calculated using the most likely segment sequence, which was derived by with consideration for both the frame-wise acoustic features using the pitch information and the prosodic features of the segment ratios. The probability density function for each cluster and each segment was used to calculate the global prosodic likelihood.

IV. Emotion Classification Methods

We formulated the classification of the emotion clusters based on a maximum likelihood approach as follows. Given the acoustic evidence observation q, our process of emotion detection aimed to find the most likely segment sequence, $\hat{\mathbf{z}}$, and the emotion cluster $\hat{e}$ that yielded $\hat{\mathbf{z}}$, which satisfied

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

$$\hat{\mathbf{z}}=\hat{s}_1\hat{s}_2\cdots\hat{s}_i\cdots\hat{s}_n\,. \tag{3}$$

The right-hand side of equation (2) can be rewritten according to Bayes' rule as

$$P(e,\mathbf{z}\,|\,q)=P(e)P(\mathbf{z}\,|\,e)P(q\,|\,e,\mathbf{z})/P(q) \tag{4}$$

where $P(e)$ is the *a priori* occurrence probability of emotion cluster e. Based on the clustering results for the cry samples, the number of data samples for the two emotion clusters were almost equal, as shown in Table II. Thus, we assumed that the probability $P(e)$ was irrelevant in our experiments. Moreover, the term $P(q)$ was not related to $\mathbf{z}$ and e, so it was also considered irrelevant. The term $P(\mathbf{z}\,|\,e)$ represents the occurrence probability that segment sequence $\mathbf{z}$ will occur in emotion cluster e. In our method, this term was calculated based on the duration ratios of the silent and resonant cry segments in the segment sequence $\mathbf{z}$, $r_c(\mathbf{z})$ and $r_s(\mathbf{z})$, respectively. This was because these segment ratios were different significantly in the two emotion clusters.

$$\log P(\mathbf{z}\,|\,e)\approx\beta\log P\big(r_c(\mathbf{z})\,|\,e\big)+\log P\big(r_s(\mathbf{z})\,|\,e\big) \tag{5}$$

The probability of each segment type is calculated using the normal distribution function for each emotion cluster. β is a weighting factor for the contribution of each segment type.

The term $P(q\,|\,e,\mathbf{z})$ is the probability that acoustic evidence q will be observed when an infant utters a sequence $\mathbf{z}$ caused by emotion cluster e. This term is calculated using HMMs where the power, mel-cepstrum coefficients, and pitch frequency of each frame are used as feature parameters. Thus, we can apply the emotion detection procedure to Eq. (2) as follows:

$$\begin{aligned}\hat{e},\hat{\mathbf{z}}&=\arg\max_{e,\mathbf{z}}\big(\log P(e,\mathbf{z}\,|\,q)\big)\\&\approx\arg\max_{e,\mathbf{z}}\big(\alpha\log P(q\,|\,e,\mathbf{z})+\log P(\mathbf{z}\,|\,e)\big)\end{aligned} \tag{6}$$

where α is a weighting factor for the contribution of the frame-wise likelihood and the segmental duration likelihood.

TABLE V
CLASSIFICATION PERFORMANCE USING FRAME-WISE ACOUSTIC FEATURES

Acoustic features			C1	C2	Average [%]
Power	MFCCs	Pitch			
∨	∨ 1,⋯,12	−	68.5	74.6	71.5
∨	∨ 1,⋯,12	∨	75.6	73.8	74.7

TABLE VI
UPPER BOUND PERFORMANCE USING SEGMENT DURATION RATIOS[%]

	C1	C2	Average
Upper bound	68.5	76.2	72.3

V. Classification Experiments and Results

The cry data were sampled at 16 kHz. Every 10 ms, we computed the power, FFT mel-warped cepstral coefficients (MFCCs), and pitch frequency using a Hamming window. We performed leave-one-out cross-validation using the segment HMMs with diagonal covariance matrices of three-state two-mixture models. The weights α and β were determined to provide the highest performance.

A. Classification Using Frame-Wise Features

1) Classification using power and MFCCs (baseline): A classification experiment for emotion clusters C1 and C2 using the power and a vector of 12 MFCCs were carried out to confirm the baseline performance. This classification procedure was conducted as follows:

$$\hat{e},\hat{\mathbf{z}}=\arg\max_{e,\mathbf{z}}\big(\log P(q\,|\,e,\mathbf{z})\big) \tag{7}$$

The average classification rate weighted with the data amount ("Average") was 71.5%, as shown in Table V.

2) Classification using pitch, power and MFCCs: In addition to the power and MFCCs, we used the pitch frequency as an acoustic parameter for the HMMs. First, a classification experiment was conducted using the feature parameters of the pitch frequency, a vector of 12 MFCCs, and the power. This classification procedure is also described by Eq. (7). The classification performance is shown in Table V. The classification performance was improved from the baseline of 71.5% to 74.7%. This demonstrated the utility of pitch information to classifying emotions.

B. Detection of Resonant Cry and Silent Segments

In our proposed approach, the segment duration ratio was obtained using the most likely segment sequence derived by the maximum likelihood approach. Based on the hand-labeled data (where the segments of the resonant cry and noise could be detected perfectly), the upper bound performance of classification using the segment duration ratios is listed in Table VI. A comparison of the upper bound in Table VI and the performances shown in Table V indicated that classification based on the frame-wise acoustic features including pitch information was superior to classification based on the segmental duration ratios.

We then investigated the detection performance of the resonant cry segments and silent segments. We performed a detection experiment using the feature parameters of the pitch frequency, a vector of 12 MFCCs, and the power. We used hand-labeled data for 253 samples, where the test data was excluded, to train a normal distribution function for the duration ratio of each emotion cluster and each segment type. The agreement rate per frame between the detected segments and the hand-labeled segments is shown in Table VII. Depending on the environmental noise during recording, there was a tendency for silent segments to be misrecognized as miscellaneous segments. We then calculated the duration ratios by transposing from the detected miscellaneous segments to silent segments. The F-measure values for the detection of the resonant cry segments or silent segments were 79.2% or 76.8%, respectively.

To evaluate the performance, we conducted another detection experiment with the conventional method (Method-1) [9] of the derivation of the duration ratios of the resonant cry and silent segments, which was based on frame-by-frame detection. This method tested whether each frame in the sample is a cry, silent, or other frame based on the pitch and power information. The experimental results are also shown in Table VII. A comparison of these results showed that the proposed detection method for the duration ratio, which was obtained from the most likely segment sequence, was more effective than the conventional method.

C. *Classification Using Frame-Wise Features and Segment Duration Ratios*

We performed a classification experiment to confirm the effectiveness of using the frame-wise acoustic features and the segmental duration features. The formulation of the proposed approach was based on Eq. (6), where the first term $P(q \mid e, \hat{z})$ was calculated using the frame-wise acoustic features (power, MFCCs, and pitch frequency) while the second term $P(\hat{z} \mid e)$ was obtained using the segment ratios. The classification performance is shown in Table VIII. The classification performance improved to 75.5%. This result demonstrates the utility of the combination of frame-wise acoustic features and segmental duration information based on a maximum likelihood approach.

Finally, we also conducted the conventional method [9] which combined the frame-wise acoustic likelihood derived from the baseline and the segmental duration likelihood obtained using the detection method (Method-1) in Section V *B*. This method and the proposed method differed with respect to whether the pitch information was used as the acoustic parameter in HMMs and whether the most likely segment sequence considered local and global acoustic features. The classification performance is also shown in Table VIII ("Conventional"). This shows that the proposed classification approach based on the derivation of the most likely segment sequence with the maximum likelihood achieved a better classification performance of 75.5%. We investigated the classification rates of four infants whose numbers of cry samples were more than 20. Though the rates for two infants were increased by using the proposed method, those for other two still remained the same.

VI. CONCLUSIONS

This paper proposed a method for emotion cluster classification based on a maximum likelihood approach with consideration for frame-wise/local acoustic features and global prosodic features. We introduced the pitch frequency as an acoustic parameter for the stochastic modeling of segment HMMs in addition to the power and MFCCs used by the conventional method. Our proposed approach allowed us to find the most likely sequence with consideration for both features, which facilitated the more accurate detection of resonant cry and silent segments. Based on the classification performance of the two major emotion clusters, the classification performance of our proposed approach was increased to 74.7% by the addition of pitch information compared with a performance of 71.5% with the classification method using only power and MFCCs. Our proposed classification method based on a maximum likelihood approach with consideration for both the local acoustic features and global prosodic features achieved a highest classification rate of 75.5% in our experiments, demonstrating the effectiveness of the derivation of the most likely segment sequence with consideration for both features.

In future studies, we need to evaluate the effect of a maximum likelihood approach using a higher number of emotion clusters.

TABLE VII
SEGMENT DETECTION PERFORMANCE PER FRAME [%]

Method	Segment	Precision	Recall	F-measure
Proposed	Cry	77.4	81.0	79.2
	Silence	57.0	96.7	76.8
Conventional (Method-1)	Cry	72.4	81.4	77.0
	Silence	64.8	82.9	73.9

TABLE VIII
CLASSIFICATION PERFORMANCE USING FRAME-WISE ACOUSTIC LIKELIHOOD AND SEGMENTAL DURATION LIKELIHOOD [%]

Method	C1	C2	Average
Proposed	77.2	73.8	75.5
Conventional [9]	70.9	75.4	73.1

REFERENCES

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

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

[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. ICA, pp.1713-1716, 2004.

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

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

[9] Kitahara, K., et al, "Emotion classification of infants' cries using duration ratios of acoustic segments," Proc. Interspeech 2011, pp.1573-1576, 2011

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