Emotion-Cluster Classification of Infant Cries Using Sparse Representation

Takayuki Kurokawa, Tasuku Miura, Masaru Yamashita, Tomoya Sakai and Shoichi Matsunaga
Nagasaki University, Nagasaki, Japan
E-mail: {b314024, b313047, masaru, tsakai, mat}@cis.nagasaki-u.ac.jp. Tel: +81-95-819-2700

***Abstract*— This paper proposes the use of sparse representation of infant cries for the classification of emotion clusters. In the recording of infant cries near an infant's mouth, short-tme noises such as popping and collision noises are frequently recorded. The short noises increase the difficulty of achieving an accurate emotion classification performance. In the proposed approach, input cry signals are divided into "resonant cry" components and sound components with short time duration, which are mainly short noises, using sparse representation; then the extracted "resonant cry" signals are used for classification. In our preliminary investigation of sparse representation, it was confirmed that the stationary components of the "resonant cry" signals and short noises were divided effectively. To evaluate the proposed approach, we performed emotion classification experiments to distinguish between the two major emotion clusters that were obtained by clustering the results of subjective opinion tests. The proposed method achieved a higher classification performance of 73.1% compared to 70.0% for the conventional method, which did not use sparse representation, demonstrating the usefulness of the proposed approach for the classification of infant cries including short noises.**

I. INTRODUCTION

Infants convey their demands to their parents and/or childminders by crying, which has a particularly important role to express their situation, especially in an urgent situation [1]. However, it is difficult to correctly estimate the emotions that infants express through their cries (i.e., the causes of crying), particularly for people who have minimal experience in childcare. However, mothers and childminders with sufficient experience in childcare can frequently sense an infant's emotion correctly through their cries. Yet, it is true that no one can know an infant's true emotion. Based on this empirical knowledge, our research was designed to detect the emotions that experienced mothers and childminders commonly sensed in infant cries. It is our opinion that this detected emotion can be useful for parents who are performing childcare for the first time.

Several previous emotion detection studies have focused on an acoustic analysis of infant cries [2, 3, 4]. Of the types of emotions in cries, pain was a traditional research area in the detection of emotions [5]. The distinction of "hunger" and "sleepiness" cries has also been studied [6]. These studies used simple matching techniques based on power and spectral features. We develop an emotion-cluster classification procedure that is based on a maximum likelihood approach using hidden Markov models (HMMs) [7]. In the proposed approach, two major emotion clusters are adopted to distinguish [8], one cluster consisting of emotional states such as "sleepiness" and "psychological dependence" (craving for his/her mother's attention), and the other cluster comprising "anger," "sadness," and "hunger." These clusters were obtained by clustering of the results of subjective opinion tests requesting childminders emotion states in infant-cry signals. The classification procedure demonstrated that the stochastic approach was promising in the classification of emotion clusters. However, in the study, cry samples, which were recorded near the mouth of each infant by each infant's mother at home using a digital voice recorder, were used. An infant frequently moves his/her limbs and head while crying. Hence, these samples contained many short-time noises such as popping and collision noises. Popping noise frequently occurred when the infant breath reached the microphone on the recorder; collision noise occurred by the collision of the infant and the recorder or some other object. Such noise pollution in the cry sounds hinders the achievement of a relatively high degree of correct detection of infant emotion. This is because there is a tendency for silence periods containing short noises to be misrecognized as other kinds of sound uttered by the infant. The acoustic features expressing emotion states are mainly involved in "resonant cry" periods, where the cry signal contains several types of sound such as "resonant cry", breath, cough, and bubbling. Hence, to obtain improved performance, an extraction technique for resonant-cry components excluding short noises is required.

Addressing this issue, we propose an emotion classification method using the sound of a resonant-cry component extracted using sparse modeling. In the proposed approach, we assume that cry sounds can be represented as a linear combination of stationary cry sound and short-duration sound including popping and collision noises. After the extraction of the stationary cry (resonant cry) components, we perform HMM-based emotion classification to distinguish between the above-mentioned two major emotion clusters using the extracted sound of the resonant-cry component. Sparse representation has been previously applied in the field of analysis of lung sounds [9]; however, it has not yet become applicable to the automatic classification of types of lung sounds because of the diversity of acoustic features. In this paper, we demonstrate the usefulness of spare representation of infants' cries containing short noises through the emotion-

cluster classification experiments.

II. Infant Cries for Emotion-Cluster Classification

A. Subjective Opinion Test for Determination of Emotion

We prepared two sets (Set-A and Set-B) of infant cries, where the primary difference between the two sets was distance from the infant to the microphone. For both sets, the recording sampling frequency was 16 kHz. The age of the infant subjects ranged from eight to 13 months; it was inevitable that a variety of noises would be mixed during the recording because the recording was performed at home. The aim of this study is to improve classification performance for infant cries, where popping and collision noises caused frequent contamination.

For Set-A, 11 mothers recorded infant cries near each infant, at home. The average duration of the recorded data was approximately 30 s. According to our investigation of these samples, 62% of the samples contained popping noises and/or collision noises. After recording each cry, the infant's mothers and three baby-rearing experts judged the emotions expressed in the samples (subjective opinion test). Ten types of emotion tags were defined: "psychological dependence," "anger," "sadness," "fear," "surprise," "hunger," "sleepiness," "excretion," "discomfort," and "pain." They assessed the emotions that they considered to be correlated with the cause of the crying of the infants. The intensities of all emotion types were ranked on a scale ranging from zero (no emotional content) to four (full emotional content). The mothers considered the cries, and the facial expressions and behaviors of their infants. The experts judged the scale using the acoustic data only.

B. Two Emotion Clusters for Classification

Emotion clustering was performed using approximately 1200 results from the subjective opinion tests using the crying samples of Set-A. The detailed clustering algorithm is described in [8]. From the clustering process using these results, there were two major emotion clusters among the generated clusters, i.e., cluster C1 consisted of emotional states such as "sleepiness" and "psychological dependence," whereas cluster C2 comprised "anger," "sadness," and "hunger." We assumed that the clustering results were correct and used these two emotion clusters in the subsequent classification experiments of emotion clusters.

C. Acoustic Labels for HMM-based Classification

A cry sample is a time signal. We considered that a cry sample was composed of five types of segments with specific acoustic characteristics. To identify the emotions expressed in a cry using an HMM-based method, we defined these segments according to their acoustic features and assigned a symbol to each segment for transcription. All of the collected data was hand-labeled using these symbols, whereas the noises were not labeled. If we assume that a cry sample $\mathbf{s}$ is comprised of n segments and we let the i-th segment be s_i $(1 \le i \le I)$,which contains the beginning time and end time information, then

$$\mathbf{s} = s_1 s_2 \cdots s_i \cdots s_I \quad (1)$$

The five types of acoustic segments (labels) are 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 resonant cry segment and silence were the two of segment types that were observed most frequently.

III. Emotion Classification Using Spares Representation of Infant Cries

A. Definition of Sparse Solution for Cry Sound

We assume that the cry signal vector $\mathbf{y}$ of a sample consists of signal-component vector $\mathbf{y}_k$ and residual vector $\boldsymbol{\varepsilon}$,

$$\mathbf{y} = \sum_k \mathbf{y}_k + \boldsymbol{\varepsilon} = \sum_k \mathbf{A}_k \mathbf{x}_k + \boldsymbol{\varepsilon} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon} \quad (2)$$

In this equation, we assume that signal component $\mathbf{y}_k$ $(1 \le k \le K)$ is expanded in a known basis matrix $\mathbf{A}_k$, and $\mathbf{x}_k$ is a sparse vector of coefficients. We can obtain a sparse solution for an l_1-regularization problem:

$$\min_{\mathbf{x}} \left(\frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1 \right) \quad (3)$$

where λ is a value to control the sparsity of the solution. Some efficient algorithms for this problem are provided in [10, 11], and we use FISTA algorithm [12].

B. Extraction of Resonant Cry Component

Using the abovementioned sparse representation, we attempt to extract the resonant cry component excluding the short noises. In the resonant cry period, there are specific frequencies with large amplitude; these frequency peaks exist continuously for a certain time. Conversely, short noise sounds have wide-raging frequency components because of their pulsating waveform. According to the acoustic features of these two types of sounds and essential difference of these acoustic features, we adopt a sinusoidal basis for the resonant-cry sounds and a wavelet basis for the short noise sound. These two base types were used in the previous study to analyze lung sounds [9].

Let $\mathbf{A}_C$ and $\mathbf{A}_W$ be a discrete cosine transform matrix and a wavelet transform matrix, respectively. We assume that the main component of infant-cry sounds can be expressed by only $\mathbf{A}_C$ and $\mathbf{A}_W$, which indicates $K = 2$ in (2). Then, we assume that the sparse representation of a cry signal $\mathbf{y}$ is described as follows:

$$\mathbf{y} = [\mathbf{A}_C \ \mathbf{A}_W] \begin{bmatrix} \mathbf{x}_C \\ \mathbf{x}_W \end{bmatrix} + \boldsymbol{\varepsilon}. \quad (4)$$

Using the sparse solution to (4), two main cry-signal components are generated as $\mathbf{y}_C = \mathbf{A}_C \mathbf{x}_C$ and $\mathbf{y}_W = \mathbf{A}_W \mathbf{x}_W$. That is, the original cry signal is divided into two characteristic signals, signal $\mathbf{y}_C$ which contains mainly the resonant cry component, and signal $\mathbf{y}_W$, which contains mainly the short noise component. Then, we expected that the performance of the emotion classification using signal $\mathbf{y}_C$ would be superior to that using the original signal $\mathbf{y}$, that was used in the previous studies [7, 8].

C. Classification of Emotion Cluster

We formulate the classification of the emotion clusters based on a maximum likelihood approach, which delivers the emotion cluster as a classification result. Given the acoustic evidence observation $\mathbf{q}$ for an unknown cry sample $\mathbf{y}$ or $\mathbf{y}_C$, the process of emotion detection aims to identify the most likely segment sequence, $\hat{\mathbf{s}}$, and the emotion cluster $\hat{e}$ that yields $\hat{\mathbf{s}}$, which satisfies

$$p(\hat{e},\hat{\mathbf{s}}\,|\mathbf{q}) = \max_{e,\mathbf{s}} p(e,\mathbf{s}\,|\mathbf{q}), \quad (5)$$

$$\hat{\mathbf{s}} = \hat{s}_1\hat{s}_2\cdots\hat{s}_i\cdots\hat{s}_{I'}. \quad (6)$$

where p indicates the probability value. The right-hand side of (5) can be rewritten according to Bayes' rule as

$$p(e,\mathbf{s}\,|\mathbf{q}) = p(e)p(\mathbf{s}\,|\,e)p(\mathbf{q}\,|\,e,\mathbf{s})/p(\mathbf{q}) \quad (7)$$

where $p(e)$ is the *a priori* occurrence probability of emotion cluster e. We assumed that the probability $p(e)$ was equal (0.5 for two emotion clusters) in our experiments. Moreover, the term $p(\mathbf{q})$ was not related to $\mathbf{s}$ and e, hence it was considered irrelevant. The term $p(\mathbf{s}\,|\,e)$ represents the occurrence probability that segment sequence $\mathbf{s}$ will occur in emotion cluster e. In our method, this term is calculated using an emotion-dependent segment bigram. The segment bigram for each emotion cluster was trained using the acoustic labels for the recorded data described in Section II. The term $p(\mathbf{q}\,|\,e,\mathbf{s})$ is the probability that acoustic evidence $\mathbf{q}$ is observed when an emotion cluster of infant is e and a segment sequence of the cry sample from the infant is $\mathbf{s}$. This term is calculated using acoustic HMMs. Thus, we can apply the emotion detection procedure to (5) as follows:

$$\begin{aligned}\hat{e},\hat{\mathbf{s}} &= \operatorname*{argmax}_{e,\mathbf{s}} \log p(e,\mathbf{s}\,|\mathbf{q}) \\ &\approx \operatorname*{argmax}_{e,\mathbf{s}} (\alpha \log p(\mathbf{q}\,|e,\mathbf{s}\,) + \log p(\mathbf{s}\,|e)) \end{aligned} \quad (8)$$

where α is a weighting factor for the contribution of the likelihood derived from the acoustic HMMs.

IV. Experiments

A. Examples of Sparse Representation of Infant Cries

We examined the sparse representation-based signal extraction of infant cry signals recorded near each infant (Set-A). To extract short noises, we set $\mathbf{A}_W$ to be the Daubechies tap-10 wavelet algorithm. Fig. 1 displays a typical example of an infant cry (waveform and spectrogram) including short noises in our collected samples for each emotion cluster. In each example, the original signal $\mathbf{y}$ and two extracted signals are presented: $\mathbf{y}_C$ intended to consist mainly of resonant cry signals and $\mathbf{y}_W$ designed to consist of short noises. This indicates that a popping noise in the left example (cluster C1) was virtually separated into signal $\mathbf{y}_W$ and a collision noise in the right example (C2) was also moved into signal $\mathbf{y}_W$; cry signal $\mathbf{y}_C$ for each sample contains practically no short noises. These acoustic phenomena indicate the effectiveness of the sparse representation. However, a beginning period of resonant cry sound in the right column was separated into $\mathbf{y}_W$ (indicated using dotted ellipse); this indicates the loss of information concerning the resonant-cry component and could result in a reduction of the performance of emotion detection.

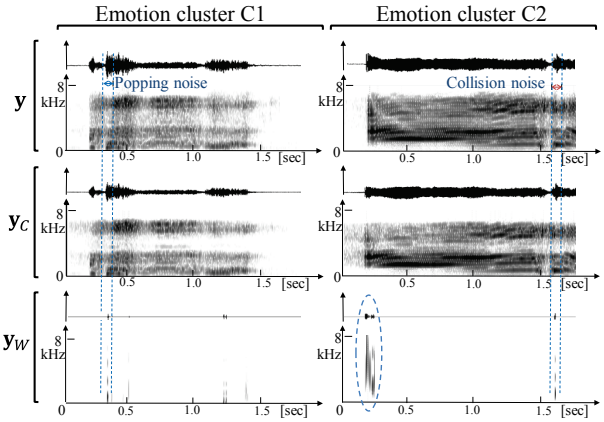

Fig. 1 Waveforms and spectrograms of original signal $\mathbf{y}$ and extracted signals $\mathbf{y}_C$ and $\mathbf{y}_W$ for each emotion cluster.

It is our opinion that the reason for this phenomenon was that same transform values of the sparse representation were applied for all cry samples.

B. Experimental Conditions for Emotion Classification

We performed HMM-based emotion classification experiments using the two sets of acoustic data: one set consisting of the original data $\mathbf{y}$ to demonstrate the performance of conventional approach, and the other set consisting of the extracted data $\mathbf{y}_C$ to indicate the performance of the proposed approach. Our emotion recognition system comprises a training process and a test process. For each process, acoustic feature parameters were extracted using the same procedure. Acoustic HMMs of each acoustic segment described were generated for each kind of emotion cluster in the training process. The amount of the training data concerning $\mathbf{y}$ and $\mathbf{y}_C$ was equal. In the test process, the acoustic likelihood of an input cry was calculated using these acoustic models, and the emotion $\hat{e}$ which provided the segment sequence $\hat{\mathbf{s}}$ with the highest total likelihood, which was the summed likelihood of the acoustic likelihood and the segment sequence likelihood, was detected as indicated in (8).

For each cry data, every 10 m, a vector of 12 mel-frequency cepstral coefficients (MFCC) and power was computed using a 25-m Hamming window. The segment HMMs were 3-state 2-mixture, context-independent models. A silent model was shared among emotion clusters.

C. Classification using Sparse Representation

We performed classification experiments to distinguish between the two emotion clusters C1 and C2 using the cry samples in Set-A. The number of samples for each emotion cluster was virtually equal as indicated in Table I. In the

TABLE I
NUMBERS OF SAMPLES IN EACH EMOTION CLUSTER C1 ("SLEEPINESS" AND "DEPENDENCE") OR C2 ("ANGER," "SADNESS," AND "HUNGER")

Data set	No. of infants	C1	C2
Set-A	11	127	126
Set-B	11	100	100

classification experiments, we performed leave-one-out cross-validation for each sample. Classification rates using the original signal $\mathbf{y}$ (conventional method) and the extracted cry signal $\mathbf{y}_C$ (proposed method) are displayed in Table II. This table confirms that the classification performance (73.1%) of the proposed method was superior to that of the conventional method (70.0%). Next, we investigated the performance for each infant. The performances of eight infants of the 11 increased; those of three infants decreased. Then, for each infant we investigated the ratio of the number of samples including short noises in all samples uttered by the infant and the degree of increase in the classification performance using the proposed method; we calculated correlation coefficient ρ between them. The value of the correlation coefficient was not significant, however, it was positive (0.17). The above-mentioned results indicate that the proposed sparse representation of infant cries was effective for the emotion classification of infant cries including short noises.

Finally, we performed an additional experiment of emotion classification using other cry samples in Set-B that contained virtually no popping or collision noises, only surrounding noises. The aim of this experiment was to confirm the influence of the proposed method for infant cries not containing such short noises. The samples were recorded distant from each infant (from approximately 2 m to 4 m) within a room at home using a fixed microphone. The infants in Set-B were different from the infants in Set-A; the number of infants in Set-B was also 11. A subjective opinion test was also performed for the samples in Set-B, and the samples belonging to each cluster were determined using the results of the test. Then, we randomly selected 100 samples among the samples in each emotion cluster (Table I). The classification performance for each type of signal ($\mathbf{y}$ or $\mathbf{y}_C$) is displayed in Table III. It was expected that the performance using the proposed method and that of the conventional method would be similar if the sparse representation was performed ideally. However, the performance using the proposed method decreased by 3.0%. This was because acoustic information of the beginning periods of resonant cry sounds was lost in signal $\mathbf{y}_C$ as described in Subsection IV.A.

V. Conclusions

This paper proposed a method for emotion cluster classification using the sparse representation of an infant's cry sound both to exclude short noises (such as popping and collision noises) and to extract "resonant cry" signals from the original cry sound. The infant cries, which were recorded near the infant's mouth, frequently contained these short noises, and these noises decreased the emotion classification performance. To address this problem, we performed the sparse representation of cry signals using a discrete cosine transform matrix and a wavelet transform matrix. Then, we performed HMM-based classification using the extracted resonant-cry signals. We compared the classification performance using the original cry signals with the performance using the extracted resonant-cry signals. The experimental results indicated that the performance (73.1%) using the resonant-cry signals, extracted by sparse modeling, outperformed the performance (70.0%) using the original cry signals, demonstrating the usefulness of the proposed method for the emotion classification of infant cries containing frequent short noises.

However, the proposed method decreased the performance for the infant cries containing virtually no short noises. Our future study is to design an applicable sparse modeling to infant cries including or not including short noises.

Table II
Classification Performance using Cry Samples Containing Frequent Short Noises [%].

Method (Signal)	C1	C2	Average
Proposed ($\mathbf{y}_C$)	73.2	73.0	73.1
Conventional ($\mathbf{y}$) [8]	77.6	63.5	70.0

Table III
Classification Performance using Cry Samples Containing Virtually No Short Noises [%]

Method (Signal)	C1	C2	Average
Proposed ($\mathbf{y}_C$)	79	80	79.5
Conventional ($\mathbf{y}$) [8]	85	80	82.5

References

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

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

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

[4] S. Nagarajan, et al., "Infant cry analysis for emotion detection by using feature extraction methods," *Proc. WRFER Int. Conf.*, pp.66-69, 2017

[5] C. Bellieni, et al., "Cry features reflect pain intensity in term newborns: an alarm threshold," Pediatric Research, Vol. 55, pp.142-146, 2004.

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

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

[9] T. Sakai, et al., "Sparse representation-based extraction of pulmonary sound components from low-quality auscultation signals," *Proc. IEEE ICASSP*, pp. 509-512, 2012

[10] S. Kim, et al., "An interior-point method for large -scale l_1-regularized least squares," *IEEE J. Selected Topics in Signal Processing*, vol.1, 4, pp. 606-617, 2007

[11] S. Wright, et al., "Sparse reconstruction by separable approximation," IEEE Trans. Signal Processing, vol.57, 7, pp. 2479-2493, 2009

[12] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *MSIAM J. Imaging Sciences,* vol.2, pp. 183-202, 2009.