EMOTION DETECTION FROM INFANT FACIAL EXPRESSIONS AND CRIES

Pritam Pal, Ananth N. Iyer and Robert E. Yantorno,
Electrical & Computer Engineering Department, Temple University
12th & Norris Streets, Philadelphia, PA 19122-6077, USA Tel: 215-204-6984
E-mail: pritam@temple.edu, aniyer@temple.edu , robert.yantorno@temple.edu
http://www.temple.edu/speech_lab

ABSTRACT

A new system for translating the infant cries from its facial image and cry sounds is presented in this paper. The system is designed to analyze the facial image and sound of the crying infant to derive the reason why the infant is crying. The image and the sound represent the same cry event. The image processing module determines the state of certain facial features, certain combinations of which determine the reason for crying. The sound processing module analyzes the data for the fundamental frequency and the first two formants and uses k-means clustering to determine the reason of the cry. The decisions from the image and sound processing modules are then fused using a decision level fusion system. The overall accuracy of the image and sound processing modules are 64% and 74.2%, respectively, and that of the fused decision is 75.2%.

1. INTRODUCTION

Crying is the first and only tool of communication for an infant. To most people (with the exception of the infant's mother), the cries seem to be uniform and indistinguishable from each other. Thus most adults are unable to distinguish them accurately on the basis of the facial image and sound. Consequently, it is useful to automate a system of classification using technology to draw a reliable decision as to why an infant is crying.

The cry types that are dealt with in the research presented here are pain, hunger, anger, sadness and fear. The system described here deals with analyzing the image of the crying face and the sound, simultaneously and independently, to classify the cry event into one of the five types. The image and the sound are recorded from the same cry event and are thus correlated. These are input to their respective processing blocks, each of which determines the reason for the cry. The two decisions are then fused employing a decision-level fusion system to obtain a single reason for the cry. The block diagram for the system is illustrated in Figure 1.

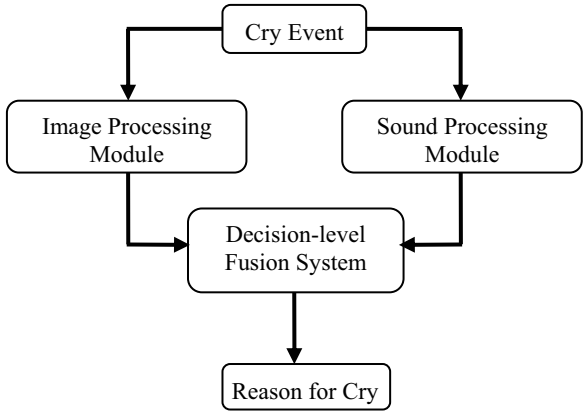

FIGURE 1: Block diagram of the infant cry interpretation system.

2. IMAGE PROCESSING MODULE

The image processing module of the system takes the image data, i.e. the image of the infant's crying face and processes it to infer the reason for its crying. The processing involves detecting the changes that occur in certain key features, like the mouth, the eyes and the eyebrows. Different combinations of the state of the mouth (open/closed), the eyes (open/closed) and the position of the eyebrows (raised-up/lowered-down) describes the basis for the infant's cry [1][2]. These different combinations along with the reason depicted by each are summarized in Table 1. Figure 2 shows the block diagram of the image module.

TABLE 1: Facial feature states for infant emotion detection.

Mouth state (open/closed)	Eyes State (open/closed)	Eyebrows Position (up/down)	Reason of Cry
Closed	Closed	Up	**Sad**
Closed	Open	Down	**Anger**
Closed	Open	Up	**Hunger**
Open	Closed	Down	**Pain**
Open	Open	Raised	**Fear**

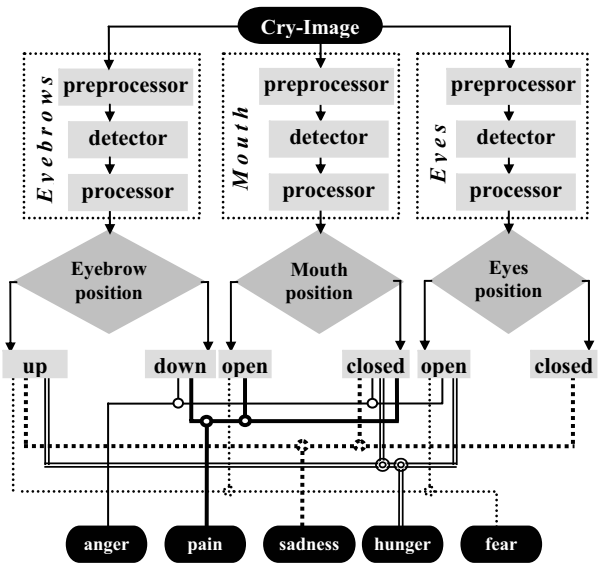

FIGURE 2: Block diagram of the image module.

A method for processing the eyes, mouth and the eyebrows to detect their state is outlined in Figure 2. A grid representing an image of a crying infant is shown in Figure 3 below, where the lower oval represents the mouth and the upper two circles represent the eyes. The shaded portions above the eyes represent the eyebrows. The gray levels of each row are successively added to generate the vertical gray-level-vector. The plot of this vector gives a representation of the gray-level intensity along the vertical direction. As seen from Figure 3, the plot of vertical gray-level-vector shows low values of intensity corresponding to regions of the mouth and the eyes (since the mouth and the eyes, when open, show darker regions compared to other regions of the face). Thus the width of these regions, if above a certain threshold, tells us that the mouth and the eyes are open. Also, a low-intensity peak is seen along the position of the eyebrows. The distance of this peak from the top extreme of the eye ('L' in Figure 3) is used to infer whether the eyebrows are up or down.

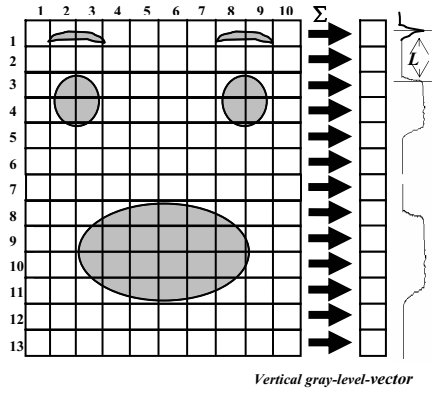

FIGURE 3: Grid of idealized face image to find facial feature states.

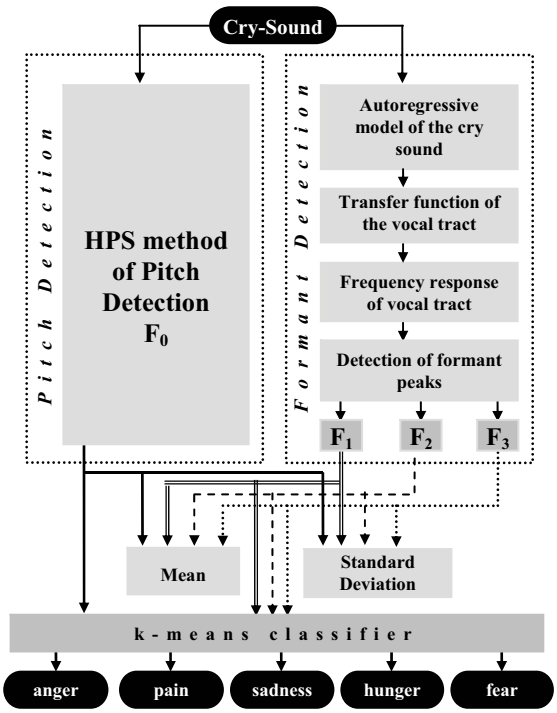

FIGURE 4: Block diagram of the sound module.

3. SOUND PROCESSING MODULE

The sound processing module of the system takes the sound of the infant's cry and processes it to infer the reason for its crying [3]. The primary focus is in extraction and analysis of the fundamental frequency (F_0) and the first three formants F_1, F_2 and F_3, of infant vocalization [4][5]. These parameters contain important information regarding the emotional state of the infant. Since F_0 for infant cries varies widely and rapidly, the Harmonic Product Spectrum (HPS) method has been used to obtain F_0. The block diagram of the sound module is shown in Figure 4.

4. DECISION LEVEL FUSION

In spite of the fact that the image and the sound data are attributable to the same cause, due to discrepancies in data acquisition and performance unreliability of the modules, the image and sound decisions may be different. Therefore a system that will fuse the two decisions of the modules based on the reliability of their performance is depicted in Figure 5. This approach was developed by [6]. The confusion matrixes, depicting the image and sound module performances, are obtained by inputting individual classifiers with a number of cry data files and noting the performance of the module on the basis of hits and misses. For the classifier, its truth has uncertainty. From the knowledge of its confusion matrix, such an uncertainty can be described by the conditional probabilities which may be defined as (3rd block of Figure 5):

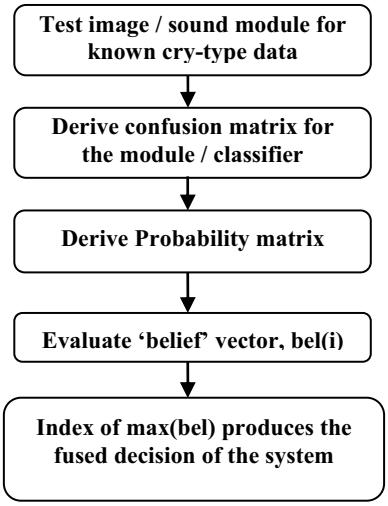

FIGURE 5: Flow diagram for the decision-level fusion process.

$$P(x \in C_i / e_k(x) = j) = \frac{n_{ij}^{(k)}}{\sum_{i=1}^{M} n_{ij}^{(k)}}, i = 1,..., M \quad .....(1)$$

Here, 'i' corresponds to known cry-types and 'j' corresponds to classified cry-types by individual classifier where M=5 (for the 5 different types of cries). Once individual conditional probability matrixes for the two classifiers have been obtained, a 'belief' vector is derived as (4th block of Figure 5):

$$bel(i) = \eta \prod_{k=1}^{K} P(x \in C_i / e_k(x) = j_k) \quad .....(2)$$

The index of the maximum value in the belief (column) vector, corresponds to the fused decision of the system. For example, if the maximum value index is 1, then the fused decision is 'Pain', etc. (2 -- 'Hunger', 3 -- 'Fear', 4 -- 'Sadness, 5 -- 'Anger').

5. RESULTS

5.1. Image Processing Results

The image data was obtained from open sources over the Internet and was hand-labeled. As discussed in Section 2, the three important features of the infant face that tell about the reason for the cry are the mouth, the eyes and the eyebrows. An example of detection of the state of each of these features is presented here. The method for the detection was discussed in Section 2. In Figure 6, the detection of the state of the mouth, lines 1, 2 & 3 represent the upper, middle, and lower limit of the mouth, respectively. These are obtained from the plot of the vertical-gray-level vector to the right of the figure.

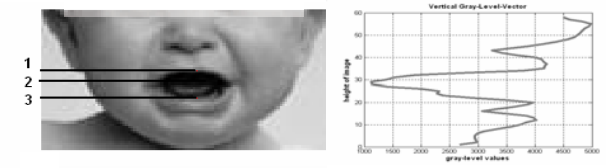

FIGURE 6: Detecting the mouth and its state.

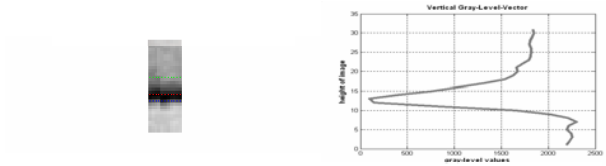

FIGURE 7: Detecting the eye and its state.

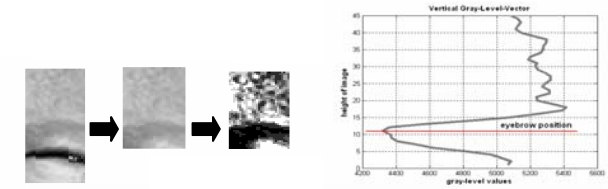

FIGURE 8: Detecting the eyebrow and its state.

Since the distance between lines 1 & 3 is above a threshold of 5 pixels (determined empirically), the mouth is detected as 'open'.

In Figure 7, similar logic is applied to detecting if the eyes are 'closed'. Here a threshold of 3 pixels has been set as the threshold. In Figure 8, the eyebrows are detected after enhancing the contrast of the image. This is necessary since for many infants the eyebrows are not as prominent as in adults. The low-intensity peak in the gray-level vector plot at the right of the figure gives the position of the eyebrows. Since the distance of this peak from the top extreme of the eye is below a threshold of 7 pixels (determined empirically), the eyebrows have been detected as 'down'.

5.2. Sound Processing Results

The sound data was obtained from [7]. As discussed in Section 3, the primary focus of the sound processing module is in the analysis of the fundamental frequency (F_0) and the first three formants F_1, F_2 and F_3, of infant vocalization. These data are clustered using a k-means classifier to obtain the optimum combination of the parameters that give maximum separation between the cry types. Thus, different combinations of the parameters F_0, F_1, F_2 and F_3 were used in the clustering process. From the results it is concluded that the combination of F_0, F_1 and F_2 produce the best clustering, according to the different cry types. Therefore this combination has been used in the sound processing module to distinguish the cry type of the infant cry sounds.

5.3. Overall Results

As discussed in Section 4, decision-level fusion is used to fuse the independent decisions of the image and sound processing modules. The individual accuracy of the two modules, as well as the fused decision (output of the overall system), is presented in Figures 9-11.

In Figure 9 the image processing module results show a higher accuracy for the pain type of cries as compared to the others. This can be explained from the fact that pain is an extreme emotional state which invokes intense facial expressions. Consequently, the algorithms perform better at detecting the key features which attain exact states as listed in Table 1.

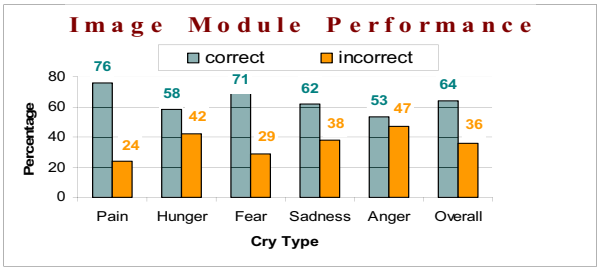

FIGURE 9: Performance of Image Module.

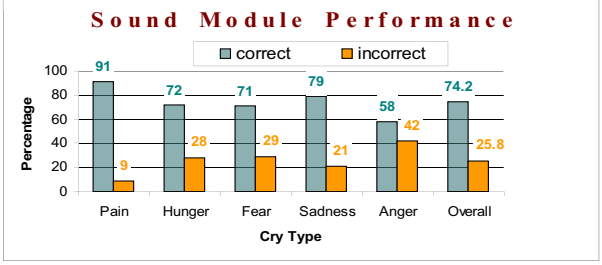

FIGURE 10: Performance of Sound Module.

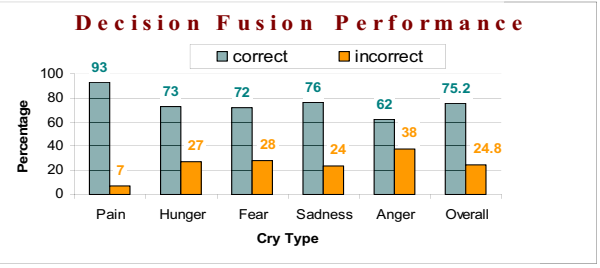

FIGURE 11: Performance of Decision Fusion.

For example, the eyes are closed in both 'pain' and 'sadness' type of cries. But the latter, not being an extreme emotion in most cases, may yield eyes which are not entirely closed. Hence there may be errors in detecting the eyes as closed in these cases.

In Figure 10, the sound processing module results show a very high accuracy for the 'pain' type of cries. This can be explained by the fact that this type of cry has a distinctive, very high fundamental frequency which is accurately detected by the pitch detection algorithm. The other types of cries have lesser distinct pitch and formants which lead to the formation of wider and more overlapping clusters, resulting in more errors in the overall detection process.

In Figure 11, the decision fusion results show that cries due to pain have the highest accuracy of detection, when compared to the other types. This is because both the sound and image processing modules perform better for the pain type of cries. For the other types of cries, as well, the decision fusion results are intuitive when the individual results of the sound and image modules are considered.

6. CONCLUSION

The cry interpreter introduced in this paper is based on the processing of the cry image and sound from an infant. The results show that each of the two modules as well as the fused decision for each type of cry performs satisfactorily in the detection process. As part of future work, an improvement in the algorithms that detect the position of the eyebrows that are not very prominent in the case of infants, can improve the overall results considerably. Also,additional metrics could be fused to increase the robustness of the sound processing module.

7. REFERENCES

[1] M. S. El-Nasr, T. R. Ioerger, J. Yen, D. H. House, and F. I. Parke, "Emotionally Expressive Agents", Proceedings of Computer Animation Conference 1999, Geneva, Switzerland; pp-48-57, 1999.

[2] S. D. Pollak, and P. Sinha, "Effects of Early Experience on Children's Recognition of Facial Displays of Emotion", Developmental Psychology 2002, Vol. 38, No. 5, 784–791

[3] H.E. Baeck, and M.N. Souza; "Study of Acoustic Features of Newborn Cries that Correlate with the Context", Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE Volume 3, 25-28 Oct. 2001 Page(s):2174 - 2177 vol.3.

[4] M. Petroni, A.S. Malowany, C.C. Johnston, and B.J. Stevens, "A New, Robust Vocal Fundamental Frequency (F_0) Determination Method for the Analysis of Infant Cries", Computer-Based Medical Systems, 1994., Proceedings 1994 IEEE Seventh Symposium on 10-12 June 1994 Page(s):223 – 228.

[5] M. Petroni, A.S. Malowany, C.C. Johnston, and B.J. Stevens, "A Cross-Correlation-Based Method for Improved Visualization of Infant Cry Vocalizations", Electrical and Computer Engineering, 1994. Conference Proceedings. 1994 Canadian Conference on 25-28 Sept. 1994 Page(s):453 - 456 vol.2.

[6] L. Xu, and A. Krzyzak, "Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition", Systems, Man and Cybernetics, IEEE Transactions on Volume 22, Issue 3, May-June 1992 Page(s):418 – 435.

[7] http://ccc.inaoep.mx/~llanto-de-bebe/pages/pages_ingles/main.html