Automatic Cry Detection in Early Childhood Education Settings

Paul Ruvolo and Javier Movellan
Institute for Neural Computation
University of California San Diego
San Diego, CA 92093-0445
Email: { paul, movellan } @mplab.ucsd.edu

***Abstract*—We present results on applying a novel machine learning approach for learning auditory moods in natural environments [1] to the problem of detecting crying episodes in preschool classrooms. The resulting system achieved levels of performance approaching that of human coders and also significantly outperformed previous approaches to this problem [2].**

I. Introduction

As part of the RUBI project for the last three years we have conducted more than 1000 hours of field studies immersing social robots at an Early Childhood Education Center (ECEC) at UCSD [3]–[5]. Our experience suggests that an important factor for the success of social robots in early education environments is the need to develop perceptual primitives to allow them to recognize basic classroom moods and to respond appropriately to these moods. It would be inappropriate, for example for a robot to sing and dance when the children are taking a nap, or to rest and do nothing when the children are crying. Due to the well known role that crying plays in early social interaction [6], [7] it is not surprising that crying episodes are a critical component of the classroom life and one for which robot assistants could be particularly useful for human teachers.

II. Prior Work

Much work has been published on behavioral studies that analyze the cry production process as well as the nature of crying as a means of guiding infant and caretaker interactions [7].

Fig. 1. *Two of the robots developed as part the RUBI project.* **Top:** *RUBI-1, the first prototype was for the most part remote controlled.* **Bottom:** *RUBI-3 (Asobo) the third prototype teaches children autonomously for weeks at a time.*

The majority of the prior work on automatic analysis of cry has focused on diagnosis of clinical conditions in infants such as severe hearing loss [8] [9]. There are important differences between this prior work and the work presented here. In clinical settings one can typically assume pristine, noise-free conditions, and the focus is on learning to detect subtle difference between cries that may be used for diagnostic purposes (e.g., to identify babies with severe hearing loss) [8]. In contrast, here we focus on developing perceptual primitives for social robots that need to operate in the noisy and unpredictable conditions of daily life. As such the focus is on robustness, i.e., spotting crying episodes in very noisy and unpredictable environments.

The problem of extracting knowledge from an auditory signal is sometimes known as auditory-scene analysis. Robust real-time auditory scene analysis has been studied in a variety of domains such as searching large audio databases [2], automatically analyzing emotional content from speech [10] [11], person identification, language identification, and music genre identification [12]. Formally all of these problems reduce to predicting a category label for given audio samples and thus are a prime target for modern machine learning methods.

One system that uses a machine learning approach for detecting auditory phenomena in noisy environments is SOLAR (Sound Object Localization and Recognition) [2]. SOLAR is designed to detect "sound objects", e.g., gunshots, doors opening and closing, laughter, in environments with high background noise. This system uses a cascaded architecture to create a detector with a very low false positive rate while keeping the true positive rate as high as possible. The motivation being that when using SOLAR as a front-end for searching a large segment of audio for a relatively rare auditory object, a low false positive rate is necessary to avoid the set of returned clips from being dominated by false positives.

Recently we proposed a novel approach to auditory scene analysis. The approach effectively converts the auditory signal into video, and applies machine learning methods that have been shown to work well in the visual domain. In the past we showed that the approach attained state of the art performance on auditory emotion recognition datasets [13]. Unfortunately the available datasets for emotion recognition research are typically collected in noise-free laboratory environments, and thus it was unclear whether the approach would generalize

Train-Time Algorithm

1) Compute 2-d Sonogram image from the raw audio signals. (see Figure 3)
2) Use Gentle-Boost to choose a set of Spatio-Temporal Box Filters to solve the binary classification problem.

Run-Time Algorithm

1) Compute 2-d Sonogram image from the raw audio signals. signal (see Figure 3)
2) Apply bank of Spatio-Temporal Box Filters selected during the training process.
3) Combine output of the filters to make a binary classification decision.

Fig. 2: General description of our approach to cry detection at train-time and run-time.

well in the difficult conditions of daily life.

III. A Database of Audio from a Preschool Setting

We recorded a full day of audio from the preschool environment. The audio was recorded using an iPod with a microphone attachment. A typical day at ECEC is divided into several periods (examples are free play, nap-time, and group singing). Each activity has a distinct auditory signature. This required the creation of a database that included audio from the complete breadth of activities at ECEC. From the six hours of audio collected from the preschool, forty minutes of audio were labeled by human coders. The labeled audio contains examples of each of the major periods of activity in the ECEC schedule. The labeling task was presented as a two alternative forced choice task between "cry present" and "cry not present". The clips were labeled by two different labelers using non-overlapping 2 second sliding windows. The agreement between labelers was 94% on the 2 second length chunks. The label (cry versus not cry) of long audio clips was assigned based on the most frequent label observed across all the non-overlapping 2 second chunks inside the clip. Of the forty minutes of the database that was coded by humans approximately 25% of the data contains toddler crying episodes.

The forty minute database of audio from the preschool was segmented into 27 episodic chunks. These audio chunks represent a continuous interval of audio that has been labeled by human coders.

IV. Our Approach to Detecting Cries

Our system [1] (see Figure 2) for detecting infant cries is inspired by recent advances in real-time visual object detection. Rather than using a small-set of hand crafted and domain specific features we use machine learning methods to select and combine features from an ensemble of several million simple light-weight features (see [1] for a more complete explanation of our approach).

A. Front End: Auditory Signal Processing

We use a popular auditory processing front end, motivated by human psychoacoustic phenomena. It converts the raw audio-signal into a 2-dimensional Sonogram, where one dimension is time and the other is frequency band, and the value for each time $\times$ frequency combination is the perceived loudness of the sound. The first step in creating the Sonogram is to take the Short-Time Fourier Transform (STFT) which converts the original 1-d temporal signal into a 2-d spectral-temporal representation. The energy of the different frequencies are then integrated into 24 frequency bands according to the Bark model [14], which uses narrow bands in low frequency regions and broader bands in high frequency regions. The energy values from the 24 Bark bands are transformed into decibels, then into Phon units using the Fletcher-Munson equal-loudness curves [14], and finally applying the standard phon-to-sone non-linearity to convert into Sone units [14]. The main advantage of working with Sone units is that they are directly proportional to the subjective impression of loudness in humans [14].

The result of these transformations is a 2-d, image-like representation of the original signal. An example of a transformed audio signal is shown in Figure 3.

B. Spatio-Temporal Box Filters

Box filters [15]–[17] are characterized by rectangular, box-like kernels, a property that makes their implementation in digital computers very efficient. Their main advantage over other filtering approaches, such as those involving Fourier Transforms, is apparent when shift-variant filtering operations are required [17]. Box filters [15]–[17] have recently become one of the most popular features used in machine learning approaches to computer vision [18] because of their efficient computational properties along with their ability to be combined using boosting methods to create very accurate classifiers. In this paper we employ a class of features, called Spatio-temporal Box Filters (STBFs) [1] that generalize the basic box filter for use in real-time machine perception problems in the auditory domain. STBFs are designed to capture critical properties of signals in the auditory domain. The first is periodic sampling in time to capture properties such as beat, rhythm, and cadence (see Figure 3). This is especially important in the context of detecting cries due the highly rhythmic structure of infant crying episodes [19]. The second is the temporal integration of filter outputs via five summary statistics: mean, min, max, standard deviation, and quadrature pair. All but the last are self-explanatory. Quadrature pairs are a popular approach in the signal processing literature to detect modulation patterns in a phase independent manner. In our case each STBF has a quadrature pair which is identical to the original STBF but half a period out of phase. Each of these summary statistics can be seen as a way of converting local evidence of the auditory category to a global estimate.

We use six types of box filter configurations (see Figure 4). The specific configuration of the box filters explored in this document is taken directly from the computer vision literature

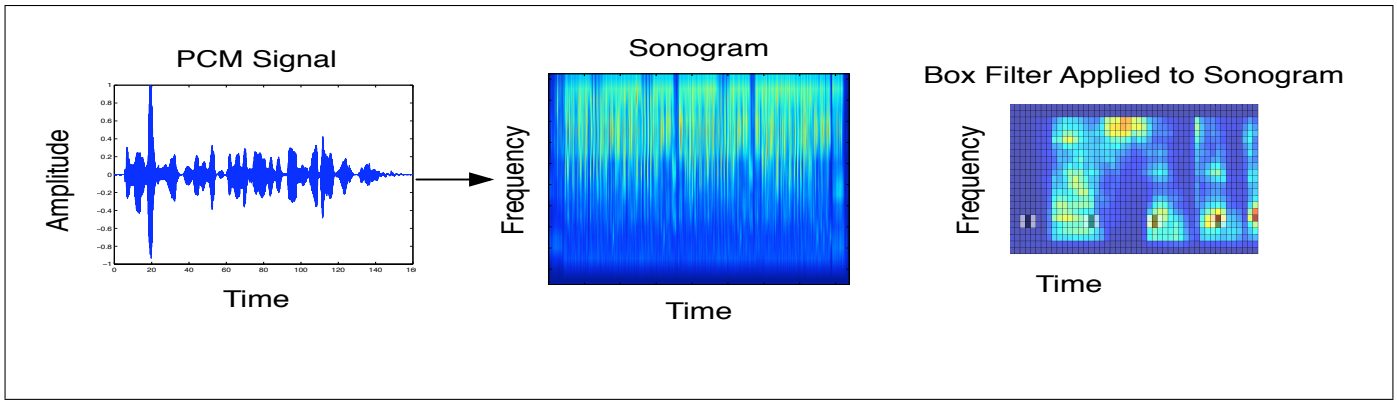

Fig. 3: Depicted above is the original 1-d temporal audio signal (left), the Sonogram (middle) and a STBF superimposed on a Sonogram (right). The periodic sampling of the box filter on the Sonogram (right) allows our system to detect rhythmic and cadence properties of the audio signal. The STBF output serves as the input to the learning framework described in section IV-C (This figure is reprinted from [1]).

[18], because they appear to compute quantities important for describing a Sonogram. In the vision literature, the response of the box filter to an image patch is given by the sum of the pixel brightnesses in the white area minus the sum of the pixel brightnesses in the black area (pixels not encompassed by the box filter are ignored). Similarly, the response of a Box filter to a portion of a Sonogram is the sum of the spectral energies of the frequency / time cells that fall in the white region minus the sum of the spectral energies of the cells fall in the black region. In the auditory domain these filters compute partial derivatives with respect to time or frequency band of the spectral energy. For instance filters of type 2 compute the partial derivative of loudness with respect to time for a particular frequency band. Filters of type 3 compute the second partial derivative with respect to frequency and time. Filters of type 4 compute the the partial derivative of loudness with respect to frequency at a specific time location. These low-level time and frequency derivatives have been shown to be useful features in sound classification [12].

The total number of features used in this work is $2,000,000$. All combinations of the 5 summary statistics, 20 sampling intervals, and $20,000$ basic box filters are considered.

C. Training

We use Gentle-Boost [20] to select and combine a subset of all possible STBFs. At each round of boosting, an optimal transfer function, or "tuning curve", is constructed for each STBF which maps feature response to a real number in $[-1, 1]$. Each tuning curve is computed using non-parametric regression methods to be the optimal tuning curve for the corresponding STBF at this round of boosting (see [21] for details). The feature + tuning curve that yields the best improvement in the Gentle-Boost loss function is then added into the ensemble, and the process repeats until performance no longer improves on a hold-out set. In this way, Gentle-Boost simultaneously builds a classifier and selects a subset of good STBFs.

To speed up search for the best feature to add (since brute-force search through all 2×10^6 possible features would be very expensive) we employ a search procedure known as Tabu Search which tries a random sample of the full set of features and then focuses on trying "nearby" features to the best features from the initial set [22]. In this respect our approach is similar to the work of Pachet and Roy [23] in which a genetic algorithm technique is used to select from a large pool of audio features.

The amount of time needed to train a classifier scales linearly with the number of examples. On a standard desktop computer it takes approximately 1 hour to train a classifier on the human coded subset of the database of audio from ECEC (described in Section III).

V. Results

Both our method and SOLAR were evaluated using leave one episode out cross validation (see Section III). Each approach was trained to classify 4-second audio clips as containing or not containing toddler crying. The binary labelers were determined from the human labels using the procedure outlined in Section III. Our system outperformed SOLAR with area under the ROC of .9467 and .9093 respectively. This difference was statistically significant ($p < 0.0001$). Figure 5 shows the results of this comparison in more detail. Of particular importance is the portion of the ROC that corresponds to a low false positive rate. The fact that our method exhibits better performance in this portion of the curve is encouraging given that SOLAR's cascaded architecture is motivated by the desire to minimize the false positive rate.

Figure 5 shows our system's time-accuracy trade-off function. The area under the ROC curve for 8 second clips was 0.97. This is comparable to the human-to-human area under the ROC 0.981, when using the labels of one coder to predict those of another coder on 8 second clips.

We performed a cursory cursory analysis of the features learned by our system. To this effect we computed a histogram of the number of features selected by our system for classification that fall into each of the 24 frequency bands of the

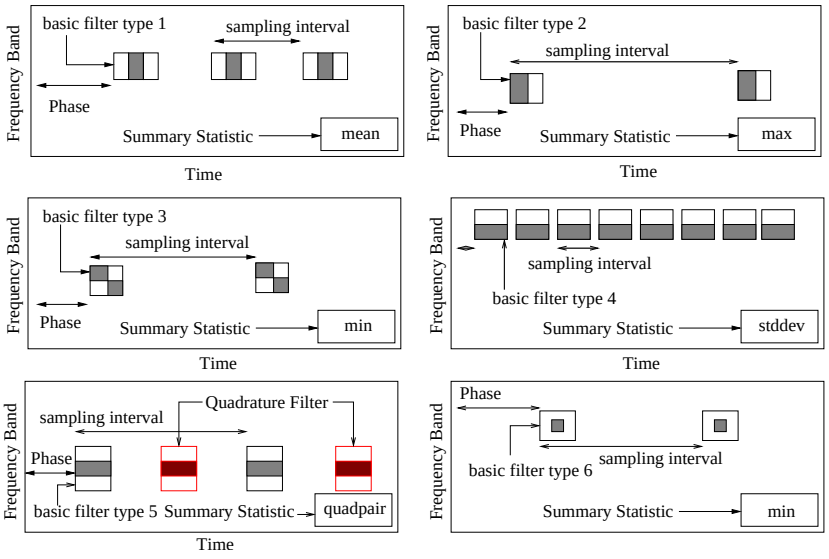

Fig. 4: Shown above are several examples of spatio temporal box filters. Each of the six basic features are shown. For each simple filter, the sum of the pixels in the black rectangle are subtracted from the sum of the pixels in the white rectangle. The output of each repetition of the simple filter yields a time series that is fed into the summary statistic specific to the particular spatio-temporal feature. This figure also appears in [1].

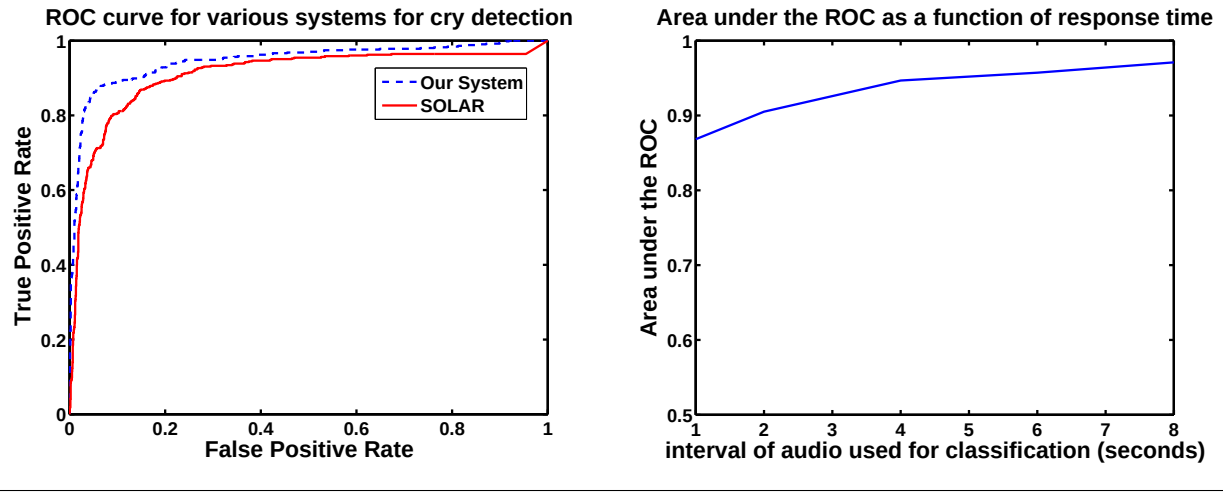

Fig. 5: *Left:* comparison of various approaches on the task of predicting whether a given 4-second window contained a child crying. Our system significantly ($p < 0.0001$) outperforms (d-prime of .9467) previously published work (d-prime of .9093) [2]. *Right:* the effect on performance of using various length windows of audio for classification. This graph demonstrates that the performance of our system substantially increases when a longer window of audio is used to make the classification decision.

bark-scale. The maximum of this histogram yields a frequency band that contains the fundamental frequency of infant cries.

VI. Conclusion

We argued that automatic cry detection in natural environments is a critical, and under-studied, problem. We collected a database of toddler crying episodes in a very noisy, early childhood education environment and showed that a machine learning approach worked exceptionally well at detecting crying episodes (97% correct detection in 2 alternative forced choice identification of 8 second audio clips).

We are currently exploring intermodal machine learning algorithms, that can take advantage of one sensory modality (e.g., audio) to train another modality (e.g. video). This could in principle allow for robots to learn how children look like when they cry and when they do not cry, so as to further improve their performance.

We are also focusing on the problem of integrating this perceptual primitive into a robot behavioral system. A promising avenue for achieving this is using data from human experts to help RUBI develop a control algorithm for shaping positive moods in the preschool environment.

ACKNOWLEDGMENT

This work was supported by the UC Discovery Grant 10202 and by the NSF Science of Learning Center grant SBE-0542013.

REFERENCES

[1] P. Ruvolo, I. Fasel, and J. Movellan, "Auditory mood detection for social and educational robotics," in *International Conference on Robotics and Automation*, 2008.

[2] D. H. Yan, "Solar: Sound object localization and retrieval in complex audio environments." [Online]. Available: citeseer.ist.psu.edu/738679.html

[3] M. J. R. T. F., T. C., R. P., and E. M., "The rubi project: A progress report," in *Proceedings of the 2nd ACM/IEEE international conference on human-robot interaction*, 2007.

[4] T. F., C. A., and M. J. R., "Socialization between toddlers and robots at an early childhood education center," *Proceedings of the National Academy of Sciences*, 2007.

[5] A. Howseman, O. Josephs, G. Rees, and T. R., "Special issues in functional magnetic resonance imaging," in *Human Brain Function*, R. Frackowiak, C. Frith, K. Friston, R. Dolan, J. Mazziotta, and R. Turner, Eds. London: Academic Press, 1997.

[6] J. Soltis, "The signal functions of early infant crying," *Behavior and Brain Sciences*, vol. 1, no. 27, pp. 443–458, 2004.

[7] A. S. Honig and D. S. Wittmer, "Toddler bids and teacher respones," *Child and Youth Care Forum*, vol. 14, no. 1, pp. 14–29, 1985.

[8] S. Möller and R. Schönweiler, "Analysis of infant cries for the early detection of hearing impairment," *Speech Commun.*, vol. 28, no. 3, pp. 175–193, 1999.

[9] G. Varallyay, Z. Benyo, A. Illenyi, Z. Farkas, and L. Kovacs, "Acoustic analysis of the infant cry: classical and new methods," *Engineering in Medicine and Biology Society*, vol. 1, no. 26, pp. 313–316, 2004.

[10] V. Petrushin, "Emotion in speech: Recognition and application to call centers," 1999.

[11] R. Fernandez and R. W. Picard, "Classical and novel discriminant features for affect recognition from speech," *Interspeech Proceedings*, 2005.

[12] G. Tzanetakis, G. Essl, and P. Cook, "Automatic musical genre classification of audio signals," 2001.

[13] W. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeir and B. Weiss, "A database of german emotional speech," *Interspeech Proceedings*, 2005.

[14] H. Fastl and E. Zwicker, *Psychoacoustics, Facts and Models*. Springer-Verlag, Berinlin Heidelberg, Germany, 1990.

[15] M. J. McDonnell, "Box-filtering techniques," *Comput. Graph. Image Process.*, vol. 17, no. 1, 1981.

[16] J. Shen and S. Castan, "Fast approximate realization of linear filters by translating cascading sum-box technique," *Proceedings of CVPR*, pp. 678–680, 1985.

[17] P. S. Heckbert, "Filtering by repeated integration," *Intternational Conference on Computer Graphics and Interactive Techniques*, pp. 315–321, 1986.

[18] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, 2002.

[19] P. S. Zeskind, S. Parker-price, and R. G. Barr, "Rhythmic organization of the sound of infant crying," *Developmental Psychobiology*, vol. 26, no. 6, pp. 321–333, 1993.

[20] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *Department of Statistics, Stanford University Technical Report*, 1998.

[21] J. R. Movellan and I. R. Fasel, "A generative framework for real time object detection and classification," *Computer Vision and Image Understanding*, 2005.

[22] F. W. Glover and M. Laguna, *Tabu Search*. Kluwer Academic Publishers, 1997.

[23] F. Pachet and P. Roy, "Exploring billions of audio features," *Content-Based Multimedia Indexing*, 2007.