2008 IEEE International Conference on
Robotics and Automation
Pasadena, CA, USA, May 19-23, 2008

Auditory Mood Detection for Social and Educational Robots

Paul Ruvolo, Ian Fasel, and Javier Movellan
University of California San Diego
{paul, ianfasel, movellan}@mplab.ucsd.edu

Abstract— **Social robots face the fundamental challenge of detecting and adapting their behavior to the current social mood. For example, robots that assist teachers in early education must choose different behaviors depending on whether the children are crying, laughing, sleeping, or singing songs. Interactive robotic applications require perceptual algorithms that both run in real time and are adaptable to the challenging conditions of daily life. This paper explores a novel approach to auditory mood detection which was born out of our experience immersing social robots in classroom environments. We propose a new set of low-level spectral contrast features that extends a class of features which have proven very successful for object recognition in the modern computer vision literature. Features are selected and combined using machine learning approaches so as to make decisions about the ongoing auditory mood. We demonstrate excellent performance on two standard emotional speech databases (the Berlin Emotional Speech [1], and the ORATOR dataset [2]). In addition we establish strong baseline performance for mood detection on a database collected from a social robot immersed in a classroom of 18-24 months old children [3]. This approach operates in real time at little computational cost. It has the potential to greatly enhance the effectiveness of social robots in daily life environments.**

I. Motivation

The development of social robots brings a wealth of scientific questions and technological challenges to the robotics community [4], [5], [6], [7], [8], [9], [10]. Social environments are complex, highly uncertain, and rapidly evolving, requiring subtle adaptations at multiple time-scales. A case in point is the use of robots in early childhood education, an area of research that we have been pursuing for the past 3 years as part of the RUBI project [3]. At any given moment the students in a classroom may be crying, laughing, dancing, sleeping, overly excited, or bored. Depending on the mood the robot must choose different behaviors so as to assist the teachers in achieving their educational goals. In addition much of the work of teachers in early education occurs at mood transition times, e.g., transition from play time to sleep time. Social robots capable of recognizing the current mood could potentially assist the teachers during these transition periods.

The goal of the RUBI project is to explore the potential use of social robots to assist teachers in early childhood education environments [3], [11]. As part of the project for the last three years we have conducted more than 1000 hours of field studies immersing social robots at the Early Childhood Education Center at UCSD. A critical aspect of these field studies is to identify the perceptual problems that social robots may face in such environments and to develop perceptual primitives for those such problems.

One of the phenomena we identified from early on is that over the course of a day the mood of the classroom goes through dramatic changes and that much of the work of the teacher occurs when they need to maintain a desired mood, or to make mood transitions, e.g. transitioning from playtime to naptime. Human teachers are indeed masters at detecting, influencing, and operating within the classroom moods. As such we identified detection of such moods as a critical perceptual primitive for social robots.

Here we investigate a novel approach to detecting social mood based on auditory information. The proposed approach emerged out of our previous experience developing visual perception primitives for social robots, and the realization of the critical role that auditory mood plays in early childhood education settings. In the following sections we describe the proposed approach, evaluate it using two standard datasets from the emotion recognition literature and finally test it on a mood detection task for a social robot immersed in an early childhood education center.

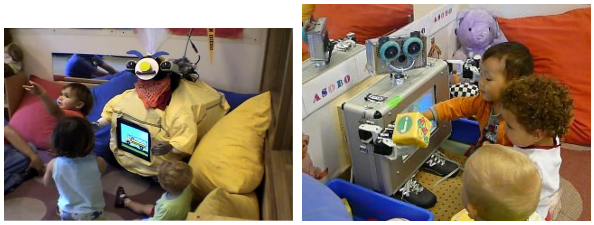

Fig. 1. *Two of the robots developed as part the RUBI project.* **Top:** *RUBI-1, the first prototype was for the most part remote controlled.* **Bottom:** *RUBI-3 (Asobo) the third prototype teaches children autonomously for weeks at a time*

II. Automatic Recognition of Audio Categories

Recognition of audio categories has recently become an active area of research in both the machine perception and robotics communities. Problems of interest include recognition of emotion in a user's voice, music genre classification, language identification, person identification, and in our case, auditory mood recognition. The robotics community has also recognized the importance of this area of research. For example, in [12] auditory information is used to determine the environment a robot is operating in (e.g. street, elevator, hallway). Formally all these problems reduce to predicting a category label for given audio samples and thus are a prime target for modern machine learning methods.

In this paper we explore an approach to recognition of auditory categories inspired by machine learning methods that have recently revolutionized the computer vision literature [13], [14]. It is interesting to note that historically machine perception in the auditory and visual domain have evolved in similar ways. Early approaches to object detection in computer vision were typically based on compositions of a small set of high-level, hand-coded feature detectors. For example human faces were found by combining the output of hand-coded detectors of eyes and other facial features [15]. Instead, modern approaches rely on a large collection of simple low-level features that are selected and composed using machine learning methods.

Similarly, much of the pioneering work on recognition of auditory categories was initially based on the composition of a small collection of hand-coded high-level features (e.g., pitch detection, glottal velocity detection, formant detection, syllable segmentation) [16], [17], [18]. An alternative approach, which is the one we explore in this document, is based on the use of machine learning methods on a large collection of simple, light-weight features. While such methods have been recently explored with some success [19], here we introduce significant changes. For example, while [19] uses learning methods to select from a pool of 276 low-level audio features, here we utilize a new collection of $2,000,000$ light-weight spatio-temporal filters, orders of magnitude larger than what has appeared in the previous literature. The potential advantage of the approach proposed here is three-fold: (1) It removes the need to engineer domain specific features such as glottal velocity that apply only to human speech. This characteristic is vital for auditory mood detection in which salient auditory phenomena are not constrained to human speech. (2) The approach relies on general purpose machine learning methods and thus could be applied to a wide variety of tasks and category distinctions. (3) The pool of auditory features was designed to be computationally lightweight and to afford real-time detection in current hardware, a critical issue for social robot applications.

Figure 2 describes the steps involved in the proposed approach. First the auditory signal is preprocessed and converted into a Sonogram, which is an image-like representation of the acoustic signal. A bank of spatio-temporal filters is then applied to the Sonogram image and combined to make a set of binary classifiers. The output of these classifiers are then combined into an n-category classifier.

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 multiple binary classification problems.
3) Combine the output of the binary classifiers using multinomial logistic regression to produce an n-category classifier.

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 into binary classifiers.
4) Combine output of binary classifiers into n-category classifier.

Fig. 2: General Description of the Approach at Train-time and Run-time

III. 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. To obtain the Sonogram, Short Term Fast Fourier Transforms (STFT) are first computed over 50 millisecond windows overlapping by 25 ms and modulated by Hanning function. The energy of the different frequencies are then integrated into 24 frequency bands according to the Bark model [20], which uses narrow bands in low frequency regions and broader bands in high frequency regions. The energy values from the 24 Bark bands are then transformed into psychoacoustical Sone units of perceived loudness. This is done by transforming the energy of each band into decibels, transforming decibel values into Phon units using the Fletcher-Munson equal-loudness curves [20], and finally applying the standard phon-to-sone non-linearity to convert into Sone units [20]. The main advantage of working with Sone units is that they are directly proportional to the subjective impression of loudness in humans [20].

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.

IV. Spatio-Temporal Box Filters

Box filters [21], [22], [23] 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 non shift-variant filtering operations are required [23]. Box Filters became popular in the computer graphics community [21], [22], [23] and have recently become one of the most popular features used in machine learning approaches to computer vision [13]. In this paper we propose a spatio-temporal generalization of Box Filters (STBF) designed for 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. 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

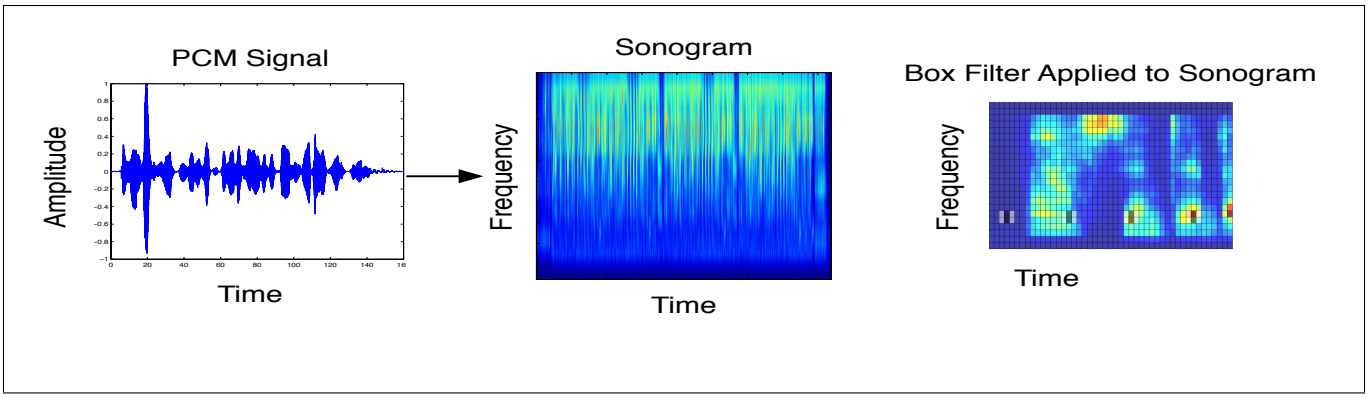

Fig. 3: Depicted above is the original 1-d temporal audio signal (left), the Sonogram (middle) and a STBF superimposed on a Sonogram (right). The STBF output serves as the input to the learning framework described in section IV-A.

modulation patterns in a phase independent manner. In our case each STBF has a quadrature pair which is identical to the original STBF but phase shifted by half a period. 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 [13], 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 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 in 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.

Figure 3 shows one of the extensions studied in this document, in this case a simple filter is periodically applied to a Sonogram. The total number of features used in this work is approximately $2,000,000$. All combinations of the 5 summary statistics, 20 sampling intervals, and $20,000$ basic box filters are considered.

A. Training

We use Gentle-Boost [24] to construct a strong classifier that combines a subset of all possible STBFs. Gentle-Boost is a popular method for sequential maximum likelihood estimation and feature selection. At each round of boosting, a transfer function, or "tuning curve", is constructed for each

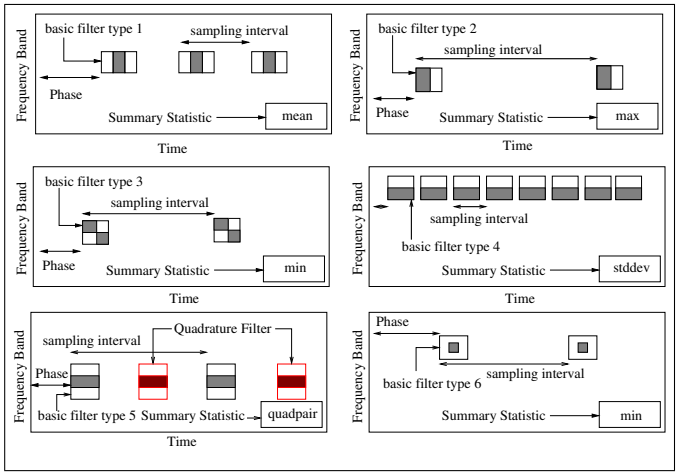

Fig. 4: Shown above are several examples of spatio temporal box filters. Each of the six basic features are shown. For each simple filter, the sum of the pixels in the black rectangle are subtracted from the sum of the pixels in the white rectangle. The output of each repetition of the simple filter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

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 [25] 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 holdout set. In this way, Gentle-Boost simultaneously builds a classifier and selects a subset of good STBFs.

At each round of boosting, an optimal tuning curve is constructed and training loss is computed for each feature under consideration for being added to the ensemble. 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 [26]. First, a random set of n filters are selected and

evaluated on the training set, and are used to initialize the "tabu list" of filters already evaluated in this round. The top $k \leq n$ of these filters are then used as the starting points for a series of local searches. From each starting filter, a set of new candidate filters are generated by replicating the filter and slightly modifying its parameters (sampling interval, phase, etc.). If the best feature from this set improves the loss, that feature is retained and the local search is repeated until a local optimum is reached.

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 a dataset of audio that is roughly 40 minutes in length.

V. Evaluation

In order to benchmark the proposed approach we performed experiments on two standard datasets of emotional speech and one on data we collected ourselves from a preschool. Once we confirmed that the approach produced competitive performance we evaluated it on a mood detection task in a social robot immersed at an early childhood education center.

A. Recognition of Emotion from Speech: Berlin Dataset

First the system was tested on the Berlin Emotional Database [1]. The dataset consists of acted emotion from five female and five male German actors. Each utterance in the database was classified by human labelers into seven emotional categories: anger, boredom, disgust, fear, joy, neutral, and sadness. Five long utterances and five short utterances are given by each speaker for each of seven emotional categories. Speech samples that are correctly classified by at least 80% of the human labelers were selected for training and testing.

To ensure speaker independence, we performed 10-fold leave one out cross validation. That is we trained our system 10 times each time leaving one speaker out of the training set and testing performance on the speaker left out. Each classifier consisted of 15 STBFs selected by the GentleBoost algorithm. In order to make a multi-class decision, we trained all possible non-empty subsets of emotions versus the rest. For a seven-way classification experiment this makes a total of 63 binary classifiers. To make the final classification decision, multinomial ridge logistic regression [27] was applied to the continuous outputs of each of the 63 binary detectors. The confusion matrix of the final system on the hold out set is presented in table I. The overall recognition rate on this seven-way classification task was 75.3%. These results are superior to several other published approaches [28]. Although it falls short of the best in the literature performance of 82.7% [29], we believe this is because the work in [29] used many optimizations to tailor their system to classification of human speech, a route that we wish to avoid for the sake generality. Also of note is that our approach is quite novel, and performed well despite this being our first attempt to employ these features. Thus we believe this approach shows great potential for improvement as we begin exploring the parameters of the technique in greater detail.

The lightweight nature of STBFs allows us to easily compute the responses of each of the 63 classifiers in real-time. Even using an inefficient run-time implementation this system can provide an estimate of the current emotion every 50 ms on a standard desktop.

B. Determining Emotional Intensity : Orator dataset

The ORATOR dataset [2] contains audio from 13 actors and 14 non-actors reciting a monologue in German. The actors were instructed to deliver the monologue as if they were in a variety of settings, such as talking to a close friend or delivering a speech. The non actors spoke spontaneously. Contrary to the Berlin dataset, in ORATOR the specific emotion categories were not explicitly prompted, but rather were situationally based. Single sentence segments of the original monologue were labeled by non-German speaking native English speakers. Each labeler was asked to rate the speech sample on seven different emotional dimensions: agitation, anger, confidence, happiness, leadership, pleasantness, and strength. This highlights another key difference between Berlin and ORATOR. In the Berlin dataset each audio clip belongs to one of a mutually exclusive set of emotions, however, the ORATOR emotions are not mutually with each monologue being rated on a continuum for each of emotional dimension. The resulting dataset consists of 150 audio samples of approximately 6 seconds each, labeled by a total of 20 labelers on 7 different emotional dimensions.

We trained a series of binary detectors to distinguish the top n versus the bottom n samples in each emotional dimension. By increasing the value of n the task becomes harder since the system is forced to correctly discriminate more subtle differences. We used two different values of n: 25 and 50 which correspond to using one third and two thirds of the original samples respectively. The consensus label for each sample was computed by taking the mean judgment across all labelers.

Table II shows the results of 14 binary classification experiments. Our approach shows performance comparable to that of the average human labeler on each task, which is considerably better that the previously reported performance on this dataset [2]. In addition to being able to successfully place a binary label on each emotional axis, the approach also achieved human-like performance at estimating continuous emotional intensity, i.e., the correlation coefficients between the detector outputs and the continuous emotion labels on an hold-out set were comparable to those of individual coders (See Table III). This ability is crucial for social robotics applications where the degree of a specific social mood is desired.

We computed several descriptive statistics of the learned features for solving this task. The most popular temporal integration function is mean, followed by the quadrature pair. This suggests that some form of phase invariance may be critical for learning the emotional characteristics of speech. The most popular frequency bands were in the range of $100 - 200$ Hertz, which contain the pitch of the average conversational male and female voice.

	Anger	Boredom	Disgust	Fear	Happy	Neutral	Sadness
Anger	**.9051**	0	0	.0238	.071	0	0
Boredom	.0281	**.7817**	.0412	.0094	0	.0677	.0720
Disgust	.1492	.0589	**.6061**	.0232	.0419	.0278	.0929
Fear	.0909	0	.0166	**.6278**	.0302	.1405	.0939
Happy	.3782	0	.0185	.0370	**.4804**	.0859	0
Neutral	.0219	.1334	0	.0414	.0452	**.7581**	0
Sadness	0	.0953	.0152	0	0	.0231	**.8665**

TABLE I

A CONFUSION MATRIX FOR THE BERLIN EMO DATABASE. THE CELL IN THE ITH ROW AND JTH COLUMN REPRESENTS THE FRACTION OF SAMPLES WITH OF EMOTION I CLASSIFIED AS EMOTION J. THE RECOGNITION RATE USING 10-FOLD LEAVE ONE SPEAKER OUT CROSS VALIDATION IS 75.3%.

	agitated	angry	confident	happy	leadership	pleasant	strong
STBFs (25 vs. 25)	.133	.033	.133	.2	.1	.2	.133
average labeler (25 vs. 25)	.082	.124	.123	.162	.105	.211	.142
STBFs (50 vs. 50)	.2166	.233	.2166	.266	.233	.233	.1833
average labeler (25 vs. 25)	.1885	.244	.183	.2225	.181	.2905	.2115

TABLE II

COMPARISON ON THE ORATOR DATASET OF THE PERFORMANCE OF VARIOUS APPROACHES ON THE BINARY CLASSIFICATION TASK OF RECOGNIZING THE TOP N EXAMPLES OF A SPECIFIC EMOTIONAL CATEGORY FROM THE BOTTOM N EXAMPLES OF THAT CATEGORY. THE QUANTITY REPORTED IS THE BALANCED ERROR RATE (THE PERCENT CORRECT WHEN THE TRUE POSITIVE RATE EQUALS THE TRUE NEGATIVE RATE). NOTE THAT LOWER NUMBERS ARE BETTER SINCE THAT IMPLIES A PARTICULAR APPROACH WAS BETTER ABLE TO MODEL THE CONSENSUS OF THE 20 HUMAN LABELERS.

	agitated	angry	confident	happy	leadership	pleasant	strong
STBFs	.48	.6	.53	.42	.5	.43	.52
average labeler	.43	.43	.49	.32	.42	.23	.41

TABLE III

THE FIRST ROW SHOWS THE CORRELATION COEFFICIENT BETWEEN THE OUTPUT OF THE TRAINED CLASSIFIER AND THE AVERAGE INTENSITY ASSIGNED BY THE LABELERS (RECALL THAT EACH LABELER PROVIDED AN ESTIMATE OF INTENSITY FOR EACH EMOTIONAL DIMENSION). THE SECOND ROW SHOWS THE AVERAGE CORRELATION COEFFICIENT BETWEEN THE INTENSITY RATING OF A PARTICULAR LABELER AND THE AVERAGE INTENSITY ASSIGNED BY ALL OF THE LABELERS.

C. Detecting Mood in a Preschool Environment

The original motivation for our work was to develop perceptual primitives for social robots. Here we present a pilot study to evaluate the performance of our approach in an actual robot setting. The study was conducted at Room 1 of the Early Childhood Education Center (ECEC) at UCSD and it was part of the RUBI project, whose goal is to explore the use of social robots in early childhood education. The experiment was conducted on a robot, named Asobo, that has been autonomously operating in Room 1 of ECEC for weeks at a time, teaching the children materials targeted by the California Department of Education.

Through discussions with the teachers at Room 1 we identified three basic moods: crying, singing / dancing, and background (everything else). Detection of these moods could result in new robot abilities with tangible benefits: (1) The robot could help reduce crying. (2) The robot could help improve the atmosphere in the classroom by dancing and singing when other children are dancing and singing. (3) The robot could avoid inappropriate behaviors, like dancing and singing when the teachers are reading to the children.

We collected a database of audio from one full day at ECEC and coded into the three moods described above. We extracted non-overlapping audio segments of eight seconds each. There were 79 examples of crying, 72 examples of playing and singing, and 151 examples from the background category. We used 80% of each of these categories and 20% for testing. In order to test the time accuracy tradeoff, we ran our detector with various length intervals sampled from the test set. For instance, to test the performance using 4 seconds of audio, a sliding window of duration 4 seconds was slid over all audio samples in the test set.

Figure 5 shows the time/accuracy tradeoff function of the system. When given 8 second audio segments, the system achieves an accuracy of 90%. As expected the performance declines if shorter audio segments are used, and it is basically at chance with less than 600 millisecond segments.

The obtained levels of performance are very encouraging considering this was a non-trivial task in a very challenging environment. We are currently in the process of developing new behaviors for Asobo to respond to the perceived mood. Preliminary anecdotal evidence is encouraging. For example, we observed a child immediately stopped crying when Asobo asked “Are you OK?”. This behavior could be made more effective if a system to localize the source of audio signals was integrated in to the system [30]. In this case ASOBO could direct his gaze to the crying child before offering his concern.

In addition, the mood detector could also be potentially used for robots to learn on their own how to behave so

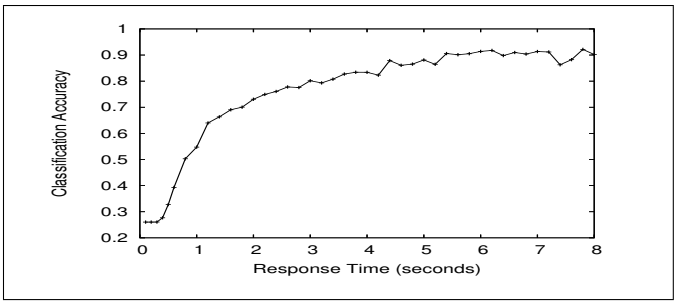

Fig. 5: Time Accuracy Tradeoff Function: the performance on the three-way classification task as a function of the time interval of sound used for classification. Using approximately half a second for classification results in a decision slightly above chance. The maximum performance is attained when using a 7.8-second sample.

as to accomplish classroom goals. For example, reduction of crying and increase of playing could be used as a reinforcement signal for the robot to learn to improve the atmosphere of the classroom.

VI. Conclusion

We identified automatic recognition of mood as a critical perceptual primitive for social robots, and proposed a novel approach for auditory mood detection. The approach was inspired by the machine learning techniques for object recognition that have recently proven so successful in the visual domain. We proposed a family of spatio-temporal box filters that differ in terms of kernel, temporal integration method, and tuning curve. The advantage of the proposed approach is that it removes the need to engineer high-level domain specific feature detectors, such as glottal velocity detectors, that apply only to human speech. Instead we let machine learning methods select and combine light-weight, low-level features from a large collection. In addition the filters are designed to be computationally efficient thus allowing real time mood detection at little computational cost, an aspect critical for robot applications.

The approach provided excellent performance on the problem of recognizing emotional categories in human speech, comparing favorably to previous approaches in terms of accuracy while being much more general. A pilot study in a classroom environment also confirmed the very promising performance of the approach.

In the near future, and as part of the RUBI project, we are planning to incorporate the mood detector for the robot Asobo to operate as a sort of social "Moodstat", i.e. a device that helps achieve a desired mood, in an analogous way as thermostats help maintain a desired temperature level.

References

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

[2] H. Quast, "Automatic recognition of nonverbal speech: An approach to model the perception of para- and extralinguistic vocal communication with neural networks," Master's thesis, University of Gottingen, 2001.

[3] J. Movellan, I. Fasel, F. Tanaka, C. Taylor, P. Ruvolo, and M. Eckhardt, "The RUBI project: a progress report," in *Human Robot Interaction (HRI)*, Washington, D.C., 2007.

[4] R. W. Picard, *Affective Computing*. The MIT Press, 1997.

[5] R. A. Brooks, C. Breazeal, M. Marjanovic, B. Scassellati, and M. M. Williamson, "The Cog Project: Building a Humanoid Robot," *Lecture Notes in Artificial Intelligence*, vol. 1562, pp. 52–87, 1999.

[6] C. Breazeal, *Designing Sociable Robots*. Cambridge, MA: MIT Press, 2002.

[7] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A Survey of Socially Interactive Robots," *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 143–166, 2003.

[8] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial," *Human-Computer Interaction*, vol. 19, no. 1-2, pp. 61–84, 2004.

[9] H. Kozima, C. Nakagawa, and Y. Yasuda, "Interactive Robots for Communication-care: A Case-study in Autism Therapy," in *Proceedings of the 2005 IEEE International Workshop on Robot and Human Interactive Communication*, 2005, pp. 341–346.

[10] J. Peter H. Kahn, B. Friedman, D. R. Perez-Granados, and N. G. Freier, "Robotic Pets in the Lives of Preschool Children," *Interaction Studies*, vol. 7, no. 3, pp. 405–436, 2006.

[11] F. Tanaka, A. Cicourel, and J. R. Movellan, "Socialization between todlers and robots at an early childhood education center," *Proceedings of the National Academy of Sciences*, In Press.

[12] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Matarić, "Where am I? Scene recognition for mobile robots using audio features," in *IEEE International Conference on Multimedia & Expo (ICME)*, 2006.

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

[14] I. Fasel and J. Movellan, "Segmental boltzmann fields," *(unpublished manuscript)*, 2007.

[15] K. C. Yow and R. Cipolla, "Probabilistic framework for perceptual grouping of features for human face detection," *Proc. Second Intl Conf. Automatic Face and Gesture Recognition,*, vol. 4, pp. 16–21, 1996.

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

[17] P. Mertens, "The prosogram : Semi-automatic transcription of prosody based on a tonal perception model," *Proceedings of Speech Prosody*, 2004.

[18] M. Lang, B. Schuller, and G. Rigoll, "Hidden markov model-based speech emotion recognition," *Acoustics, Speech, and Signal Processing Proceedings*, 2003.

[19] B. Schuller, S. Reiter, R. Müller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker independent speech emotion recognition by ensemble classification," *Multimedia and Expo ICME*, 2005.

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

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

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

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

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

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

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

[27] J. R. Movellan, "Tutorial on multinomial logistic regression," *MPLab Tutorials. http://mplab.ucsd.edu*, 2006.

[28] W. D. Zhongzhe Xiao, E. Dellandrea and L. Chen, "Two-stage classification of emotional speech," *International Conference on Digital Telecommunications*, 2006.

[29] E. A. Thurid Vogt, "Improving automatic emotion recognition from speech via gender differentiation," *Language Resources and Evaluation Conference*, 2006.

[30] J. Hershey and J. Movellan, "Audio vision: Using audiovisual synchrony to locate sounds," 2000.