Recognizing Child's Emotional State in Problem-Solving Child-Machine Interactions

Serdar Yildirim
Computer Engineering Department
Mustafa Kemal University, Hatay, Turkey
serdar@mku.edu.tr

Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory
University of Southern California, Los Angeles
shri@sipi.usc.edu

ABSTRACT

The need for automatic recognition of a speaker's emotion within a spoken dialog system framework has received increased attention with demand for computer interfaces that provide natural and user-adaptive spoken interaction. This paper addresses the problem of automatically recognizing a child's emotional state using information obtained from audio and video signals. The study is based on a multimodal data corpus consisting of spontaneous conversations between a child and a computer agent. Four different techniques– k-nearest neighborhood (k-NN) classifier, decision tree, linear discriminant classifier (LDC), and support vector machine classifier (SVC)– were employed for classifying utterances into 2 emotion classes, *negative* and *non-negative*, for both acoustic and visual information. Experimental results show that, overall, combining visual information with acoustic information leads to performance improvements in emotion recognition. We obtained the best results when information sources were combined at feature level. Specifically, results showed that the addition of visual information to acoustic information yields relative improvements in emotion recognition of 3.8% with both LDC and SVC classifiers for information fusion at the feature level over that of using only acoustic information.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Speech recognition and synthesis*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Natural language*

General Terms

Languages, Human Factor, Design

Keywords

Emotion recognition, child-computer interaction, spoken dialog systems

ICMI-MLMI '09 Workshop on Child, Computer and Interaction November 5, 2009, Cambridge, MA, USA

1. INTRODUCTION

There has been increasing interest in the design of spoken interfaces for applications such as interactive tutoring, games, and instructional materials for children. Notably, it is desirable to provide children with computer interfaces that allow more active engagement. This can be facilitated by allowing human-computer interaction (HCI) systems to robustly handle natural modalities of human communication to process and interpret the underlying content and intent. One of the steps toward building natural and responsive spoken interfaces is to enable automatic emotion recognition capability within a spoken dialog system framework. However, building such interfaces especially for the younger children is a challenging task. Children are still learning linguistic rules of social and conversational interaction. They interact with the computer differently from adults. Hence it is important to study how emotions are communicated by children, and can be automatically processed by machines.

In the last few years, significant research has been done on the automatic recognition of emotion from spoken language [6]. The goal is to recognize emotional state of the user using segmental and/or supra-segmental information obtained from speech. Information such as speech prosody (pitch, duration, and energy) and spectral parameters, language use patterns, and discourse cues have been widely explored in the context of emotion recognition [9]. In this paper we focus on the prosody parameters of speech signal for recognizing emotional state of young children.

In addition to the verbal channel, nonverbal cues such as hand and head movements and facial expressions also play an important role in human-human spoken interaction. Given that, several recent studies have addressed the use of information obtained from the visual channel in addition to that of the audio channel to explore the multiple aspects of human communication. Chen et al. [5] included gestural information for improving sentence boundary detection. Recently, audio-visual user uncertainty cues using spontaneous conversations between a child and a computer in a problem-solving setting were analyzed in [3]. Facial and vocal modalities were effectively combined to improve emotion recognition in [10]. In this paper, we adopted a simple approach based on motion properties estimated from motion intensity changes between video frames to model visual information [15].

Even though emotions can be grouped into several classes, recognition of a large set of emotions is often not needed in the context of application dependent spoken dialog design. Hence many research efforts have usually focused on a lim-

ited number of emotion categories such as negative vs. non-negative [9], frustration/annoyance vs. neutral [1], and polite, frustrated, and neutral [14]. In [14], the authors focused on automatically detecting frustrated, polite and neutral attitudes from the child's speech and studied their differences as a function of age and gender. Their study was based on a dialog corpus from children 7 to 14 years old playing a voice activated computer game. In this study, we addressed the problem of recognizing *negative*, and *non-negative* emotions of children in pre-literate age groups, 4 to 6 years old, in problem-solving child-machine interactions.

The rest of the paper is organized as follows. The speech data and the emotion labeling procedure are described in Section 2. Feature extraction from the different data sources is discussed in Section 3. Experimental results and discussion are provided in Section 4.

2. MULTIMODAL DATABASE

In this work, we used data from the Little Children's Interactive Multimedia Project [15]. The database consists of audio-visual data from 50 children of ages 4-6 years. A subset of the data from 10 children for which transcription and annotation were completed was used for the experiments reported in this paper. The detailed information about data collection procedure, transcription and annotation are given below.

2.1 Data Collection Protocol

The experiment was conducted using a Wizard of Oz (WoZ) paradigm, where a hidden human agent manipulates a computer's behavior. A WoZ tool was designed using Visual Basic that enabled scheduling and replay of audio/video events; the content itself was a combination of recorded and synthesized speech and audio prompt and graphics. Each session took approximately 30 minutes and began with a warm-up briefing by the experimenter. This is followed by a briefing by the computer agent that parallels the human briefing. Next came the experimental battery, which included a set of five tasks, comprised of pattern recognition, sorting and category membership. Following the five tasks, subjects were debriefed by the computer agent, and then once again by the experimenter in an analogous format. Prior and subsequent to the computer interactions, a researcher conversed with each child both (1) to brief and debrief them on issues such as prior computer experience as skill levels, comfort level with our games, attitude toward the computer agent, and the other preferences, as well as (2) to allow comparisons of child-adult and child-computer agent discourse.

2.2 Transcription and Annotation

The transcription and annotation of the audio-video data were carried out in several stages. First, audio data from each session were transcribed by a native speaker of English and further double-checked by a second native speaker of English. Next, the transcriptions were imported, utterance-by-utterance, into the Praat tool [4] to allow for aligning of the transcribed material with their acoustic counterpart. The output of this process was further imported into a multi-layer annotation tool to encode additional verbal and non-verbal information and to synchronize with the visual information. A multi-layer annotation board was constructed using the Anvil multimodal annotation tool [8]. Along with the speech transcription and acoustic information (e.g., pitch and intensity contours), discourse information, such as repairs (e.g., repetitions, clarifications, corrections), speech acts (e.g., opening, providing information, acknowledging) and pacing strategies (e.g., topic termination, anticipated response, topic shift) as well as gestural information, such as hand/head movements (e.g., pointing, nodding, yes/no type head shakes), body postures and facial expressions were encoded in a synchronized multi-layer manner. These annotated data were used for the analysis; previous work on the analysis and detection of disfluency boundaries using audio-visual information can be found in our previous work [15].

In this paper we focused on recognizing an emotional state, *negative* vs. *non-negative*, of a child using audio-visual information. In addition to the annotation process explained above, user utterances were also encoded independently into two emotion categories, *negative* and *non-negative*, by two different labelers. The kappa agreement between the two labelers was 0.67. We consider only the 515 utterances that both labelers agreed on. In this study, negative emotions represent anger and frustration, whereas non-negative emotions represent neutral or positive emotions, i.e., joy or happiness. Labelers also took video information into consideration.

3. FEATURE EXTRACTION

3.1 Acoustic Feature Extraction

We used 20 different parameters comprising utterance level statistics corresponding to F0 (fundamental frequency), energy, and duration to model the emotional information carried in the speech acoustic signal. The *get_f0* function of the ESPS program was used to obtain pitch contours. The pitch contours were smoothed using a 3-point filter. Intensity contours of each utterance were obtained using the Praat speech processing tool. Details of the parameters used are summarized below.

- F0: Mean, median, standard deviation, maximum, minimum, range, linear regression coefficients.
- Energy: Mean, median, standard deviation, maximum, minimum, range, linear regression coefficients.
- Duration: utterance duration, average voiced and unvoiced duration, inter-word silence duration, longest voiced portion duration, speaking rate.

3.2 Visual Feature Extraction

We used a gradient method to estimate the 2D image velocity $\mathbf{v}(\mathbf{x}, t)$. In this method, velocity is computed by calculating the optical flow field from the spatial and temporal derivatives of image intensity [2].

$$\nabla I(\mathbf{x}, t).\mathbf{v}(\mathbf{x}, t) + I_t(\mathbf{x}, t) = 0 \quad (1)$$

where $\nabla I(\mathbf{x}, t)$ is the spatial intensity gradient, $\mathbf{v}(\mathbf{x}, \mathbf{t})$ is the image velocity and $I_t(\mathbf{x}, t)$ is the partial temporal derivative of image intensity $I(\mathbf{x}, t)$. 2D velocity can be obtained by solving Equation 1. For each frame, the average velocities for the horizontal (v_x) and vertical (v_y) directions were calculated by averaging the corresponding speeds of all pixels. Average speed of each frame using v_x and v_y were also calculated.

To model visual information, we calculated 21 utterance level statistics calculated from the vertical, horizontal and average speed trajectories obtained from the corresponding video sequence for each utterance. These statistics are mean,

Table 1: 10 best features selected by SFFS method.

Acoustic Features	Visual Features
$F0\_mean$	s_y_range
$F0\_min$	s_y_min
E_min	s_max
$F0\_median$	s_mean
$unvoiced\_dur$	s_median
$F0\_reg\_coeff$	s_min
E_std	s_range
E_mean	s_std
E_median	s_y_mean
E_range	s_x_mean

median, standard deviation, maximum, minimum, range, and linear regression coefficients of each trajectory.

3.3 Feature Selection

Some of the acoustic and visual features that are summarized above may be irrelevant for emotion recognition and therefore they can degrade the classifier performance. To eliminate irrelevant features from the base feature set and thereby to reduce the large amount of acoustic and visual features, we employed two different feature selection and reduction techniques, sequential Forward Feature Selection (SFFS) and principal component analysis (PCA).

Sequential Forward Feature Selection (SFFS) starts with the single best feature according to some criterion from the whole feature set. The subsequent features are added from the remaining features which maximize the feature selection criterion. In this paper we choose to use Nearest neighbor classifier classification accuracy as the feature selection criterion. The selection stops when the number of features added reaches the pre-set number. Principle component analysis (PCA) finds a linear transformation that projects high-dimensional data onto a lower dimensional space.

In this study, we reduced the dimension of feature vectors to 15 by applying SFFS. The list of the first ten acoustic and visual features in decreasing order of importance is given in Table 1. Note that Pitch related features are the most discriminative followed by Energy. Visual features related to vertical speed seem to be more discriminative than other visual features.

4. EXPERIMENTAL RESULTS

In this section, we present experimental results on recognizing the emotional state in the children's speech using acoustic and visual information. The task was to identify the emotional state (i.e., valence), *negative* vs. *non-negative*, of spoken utterances.

4.1 Methodology

We tested four different classifiers: Linear discriminant classifier (LDC), k-nearest neighbor (k-NN), support vector machines (SVC), and decision tree (J48). LDC is a classifier which assumes that each class has a Gaussian probability density with common covariance, and k-NN classifies a test object according to the most frequent class labels amongst the k-nearest neighbors. The number of neighborhoods was set experimentally to 3 for both acoustic and visual information for the k-NN classifier. The third method that we used was the support vector machine classifier (SVC) [12, 7]. The idea of SVM classifier is to map the training data to a higher dimensional feature space via nonlinear mapping $\mathbf{\Phi}(.)$ and construct a separating hyperplane to achieve maximum margin between classes. This can be achieved by using *kernel function* which is the dot product of support vectors and feature vectors.

$$K(\mathbf{x}, \mathbf{x}_i) = \mathbf{\Phi}(\mathbf{x}).\mathbf{\Phi}(\mathbf{x}_i) \tag{2}$$

Then, the decision is made using the nonlinear decision function given below.

$$f(\mathbf{x}) = sign(\sum_i y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b) \tag{3}$$

where α_i's are Lagrange multipliers and b is the bias term. In this study, we used polynomial kernel with degree 2. The advantage of the SVC approach is that SVC tends to be less prone to overfitting problem than other classifiers.

In this study, we also tested performance of the decision tree classifer. We used a J48 decision tree in WEKA [13] as a decision tree based classifier for this binary classification problem. J48 is an implementation of C4.5 decision tree [11] in WEKA.

In this work, we evaluated both feature level and decision level information fusion techniques to combine acoustic and visual information streams. In the feature level integration, the acoustic and visual features were combined and a single classifier was built using the new combined feature set. In the decision level integration, a separate classifier was built for each information source, and the output posterior probabilities obtained from the classifiers were combined using a chosen fusion function. Some common fusion functions are average, product, maximum, and minimum. In this work, we used the product fusion rule in which posterior probabilities are multiplied and the maximum is selected.

The performance of the classifiers was evaluated by leave-one-speaker-out cross-validation, i.e., data from one child was used for testing while the rest were used for training. The procedure was repeated until data from all children were used for testing. In this study, 515 utterances (265 non-negative, 250 negative) were considered. The baseline was the case when all data were classified as non-negative.

Table 2: Classification accuracy, in percent, for acoustic and visual information with different feature selection approaches. Acou: Acoustic, Vis: Visual

	Base		SFFS		PCA	
	Acou	Vis.	Acou	Vis.	Acou	Vis.
k-NN	64.7	59.4	65.2	60.2	64.3	59.1
LDC	62.5	58.7	65.7	61.5	62.4	59.2
SVC	64.0	58.0	66.3	61.6	63.2	58.9
D.Tree	64.3	58.3	-	-	-	-
Baseline	51.4	51.4	51.4	51.4	51.4	51.4

4.2 Results and Discussion

Table 2 shows the classification performances of different classifiers using all features and also the effects of sequential forward feature selection (SFFS) and the feature reduction by PCA on the classifiers for both acoustic and visual information. Performance improvement over baseline is significant for both information sources for all classifiers (as tested by one-sample t-test, $p < 0.001$) in terms of classification

Table 3: Performance of classifiers in terms of f-score for each class. Results given for all classifiers are based on SFFS feature sets. Neg.:Negative

	Acoustic		Visual	
	Non-Neg.	Neg.	Non-Neg.	Neg.
k-NN	0.66	0.65	0.60	0.63
LDC	0.64	0.68	0.59	0.62
SVC	0.64	0.70	0.60	0.61
Decision Tree	0.63	0.68	0.59	0.57

Table 4: Classification accuracy results for acoustic, visual features and the combination of acoustic and visual features.

	k-NN	LDC	SVC	DecTree
Acoustic Only	65.2	65.7	66.3	64.3
Visual Only	60.2	61.5	61.6	58.3
Feature Level	63.4	68.2	68.3	65.2
Decision Level	65.6	66.6	66.8	65.6
Baseline	51.4	51.4	51.4	51.4

accuracy. Even though k-NN performed better than other classifiers when base features were used, the performance differences among classifiers are not statistically significant.

For the feature selection experiments, we obtained statistically significant classification performance improvement by using SFFS feature sets over using all features for both information sources (one-sample t-test, $p < 0.001$). As can be observed from the table, SVC performed better than k-NN, LDC and decision tree classifiers. It is worth noting that the feature sets obtained by PCA, which reduced the feature dimensions, showed comparable performance with the base feature sets. Table 3 shows the performance of classifiers based on SFFS feature sets in terms of f-score for each emotion class for both information streams.

Next, we investigated the combination of the visual information with acoustic information in order to improve the classification performance in terms of classification accuracy. Two information fusion techniques were used: feature level and decision level. The results are shown in Table 4. Overall, combining visual information with acoustic information leads to performance improvements in emotion recognition. We obtained the best results when information sources were combined at the feature level. Specifically, results showed that the addition of visual information to acoustic information yields relative improvements in emotion recognition of 3.8% for LDC and SVC classifiers for information fusion at the feature level over that of using only acoustic information. This improvement is also statistically significant (one-sample t-test, $p < 0.01$).

5. FUTURE WORK

In this study, we focused on an approach towards automatic recognition of a child's emotional states using information gathered from audio and visual channels. It is well known that the problem of modeling visual information is a challenging task, thus, in this preliminary research we adopted a simple approach based on motion intensity changes between video frames to model visual information. In future work, we plan to carry out an integrated analysis of the verbal and gestural (i.e. hand and head movements, and facial expressions) characteristics of children's speech as a function of emotional state. We believe such an analysis may yield further improved results.

Other sources of information that have been shown to be useful for emotion recognition are language, dialog and discourse cues. As we annotate more data, we plan to investigate and utilize these information sources in conjunction with the audio/visual information considered in this paper.

6. REFERENCES

[1] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of ICSLP*, Denver, CO, 2001.

[2] J. L. Barron, D. J. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journal on Computer Vision*, 12:43–77, 1994.

[3] M. Black, J. Chang, and S. Narayanan. An empirical analysis of user uncertainty in problem-solving child-machine interactions. In *Proceedings of the Workshop on Child, Computer and Interaction*, Chania, Greece, 2008.

[4] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 4.3.19) [computer program]. 2005.

[5] L. Chen, Y. Liu, M. Harper, and E. Shriberg. Multimodal model integration for sentence unit detection. In *Proceedings of ICMI*, State College, PA, 2004.

[6] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001.

[7] M. A. Hearst. Trends and controversies: Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, 1998.

[8] M. Kipp. Anvil - a generic annotation tool for multimodal dialogue. In *Proceedings of the Eurospeech*, pages 1367 – 1370, 2001.

[9] C. M. Lee and S. Narayanan. Towards detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–302, 2005.

[10] A. Metallinou, S. Lee, and S. Narayanan. Audio-visual emotion recognition using gaussian mixture models for face and voice. In *Proceedings of IEEE International Symposium on Multimedia*, Berkeley, CA, 2008.

[11] R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[12] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, NY, 1995.

[13] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

[14] S. Yildirim, C. M. Lee, S. Lee, A. Potamianos, and S. Narayanan. Detecting politeness and frustration state of a child in a conversational computer game. In *Proceedings of the Eurospeech*, Lisbon, Portugal, 2005.

[15] S. Yildirim and S. Narayanan. Automatic detection of disfluency boundaries in spontaneous speech of children using audio-visual information. *IEEE Transaction On Audio, Speech, and Language Processing*, 17(1):2–12, 2009.