Detecting communication errors from visual cues during the system's conversational turn

Sy Bor Wang, David Demirdjian, and Trevor Darrell

Computer Science and Artificial Intelligence Laboratory, MIT
32 Vassar Street, Cambridge, MA 02139, USA
{sybor, demirdji, trevor}@csail.mit.edu

ABSTRACT

Automatic detection of communication errors in conversational systems has been explored extensively in the speech community. However, most previous studies have used only acoustic cues. Visual information has also been used by the speech community to improve speech recognition in dialogue systems, but this visual information is only used when the speaker is communicating vocally. A recent perceptual study indicated that human observers can detect communication problems when they see the visual footage of the speaker during the system's reply. In this paper, we present work in progress towards the development of a communication error detector that exploits this visual cue. In datasets we collected or acquired, facial motion features and head poses were estimated while users were listening to the system response and passed to a classifier for detecting a communication error. Preliminary experiments have demonstrated that the speaker's visual information during the system's reply is potentially useful and accuracy of automatic detection is close to human performance.

Categories and Subject Descriptors

H.1.2 [**User-Machine System**]: Human Information processing; I.5.4 [**Computing Methodologies**]: Pattern Recognition Applications—*computer vision*

General Terms

Reliability, Human Factors, Experimentation

Keywords

Visual Feedback, System Error Detection, Conversational systems

1. INTRODUCTION

Many spoken dialogue systems have difficulty detecting the occurrence of communication errors (e.g., errors made by the speech recognizer). By our definition, a communication error occurs when the automatic speech recognition system misinterprets the user and makes an erroneous reply. Considerable research has been invested in monitoring audio cues to detect such communication errors. Various researchers have shown that human users change their speaking style when they encounter a problem with a conversational system. For example, users tend to speak slower or louder when speech recognition errors occur. These problems motivated the monitoring of prosodic aspects of a speaker's utterances. Different studies using prosodic features to detect communication errors automatically have achieved varying results [10, 12, 1].

A recent perceptual study [2] indicated that the detection can be improved significantly if visual modalities are taken into account. This study showed that human observers performed better at recognizing communication errors when they were given the visual footage of the speaker when the speaker is listening to the system's response as compared to when only audio recordings were provided. This insight motivates us to detect communication errors automatically by using the non-verbal facial expressions of the speakers when the system is making its reply. Figure 1 illustrates the reaction of a user experiencing a speech recognition error from a conversational system. Notice that when the system is giving its reply, the co-occurring facial expressions can be indicative of the communication state. In most existing work, when a system makes an erroneous reply, errors are detected only during the speaker's next turn in the conversation, but this insight shows that we can detect errors within the same turn. Systems can benefit from this early feedback to improve the quality of the conversation.

In this paper, we describe work in progress towards the development of a communication error detector. Two datasets are collected: the first dataset consists of actors articulating facial expressions on demand. The second dataset consists of subjects interacting with a conversational system in a natural setting. In both datasets, facial motion features and head poses were estimated and passed to a classifier for training and evaluation. We describe preliminary experiments which show that automatic communication error detection based on visual cues while a user is listening is useful, and that this detection performance is similar in accuracy compared to human observers.

2. RELATED WORK

Related work can be mainly divided into three categories, namely, audiovisual emotion recognition (audiovisual), sys-

ICMI'07, November 12-15, 2007, Nagoya, Aichi, Japan.

Figure 1: Diagram illustrating a speaker encountering a communication error. Notice that during the system's conversational turn (from time $t1$ to time $t2$), where it is giving an erroneous reply, the speaker twitched his eyes naturally to signal an error. This suggests that visual analysis during the system turn may be useful.

tem error detection (audio) and facial expression analysis (visual).

There has been considerable research activity in the domain of detecting communication problems using audio cues. Litman et al. [10] developed a rule based system error detector using prosody statistics. Hirschberg et al. [9] used a combination of prosody, speech recognition features, dialog history and the conversation turn to detect errors. This combination of different audio features proved very useful. Oviatt [12] showed that there is a pattern of hyper-articulation when there are system errors, which leads to worse recognition results. However, a separate study by Ang et al. [1] suggested that hyper-articulation may not be a good predictor of frustration.

Visual features have also been used to recognize facial expressions. Two comprehensive reviews can be found in [7, 13]. Most existing work encode facial expressions based on the basic movements of facial features called action units (AUs), as inspired by Ekman [6]. Recently, Cohen et al. [4] proposed a Tree Augmented Classifier that learns the dependencies between facial features.

A comprehensive survey of using audiovisual cues for emotion recognition can be found in [5]. The most recent and relevant work was presented by Zeng et al. [18], who used a voting method to combine the facial expression and prosody-based emotion classifiers to improve affect recognition. Chen [3] fused the scores of facial feature and prosody classifiers occurring sequentially in time. Yoshitomi et al. [17], used a weighted sum of neural network classifiers to combine audio and visual features. Song et al. [15] proposed a tripled hidden Markov Model (HMM) to perform audio visual emotion recognition. The key advantage of this model was allowing asynchrony of the audio and visual observation sequences while preserving their natural correlation over time.

We believe our work is the first one to detect communication errors visually during the conversational system's reply. Our work is motivated by a perceptual study conducted by Barkhuysen et al. [2], which showed that human observers performed better at detecting errors when they only view the visual footage of the speaker listening to the response of the system than when they only use audio cues. Previous works only use the speaker's audio and/or visual features in the next conversational turn to detect errors, but this psychological insight shows that we could detect the errors earlier within the current turn.

3. APPROACH

We propose detecting the communication errors visually when the speaker is listening to the system's reply. During the interaction, video footage of the speaker is segmented per conversation turn. Each segment's boundaries are illustrated in Figure 1, where the segment of interest starts from time $t1$ and ends at time $t2$. At time $t1$, the system starts its reply and at time $t2$, the speaker clicks the "click-to-talk" button to communicate in the next conversation turn. For each frame in the video segment while the user is listening, facial motion features and head pose features are estimated. These visual features are subsequently evaluated by Hidden Markov Models (HMMs) [14] trained for communication error detection.

In our framework, head motion is decomposed into two distinct components. The first component consists of the 3D rigid motion, δ, of the head. The second component consists of the local motion, V_f, generated by the non-rigid parts of the face (e.g. mouth, lips, eyes). Our head motion estimation algorithm consists of three components: estimating the rigid motion component of the head, using this rigid motion to compensate for the global motion of the head and estimating the local motion of the non-rigid parts.

3.1 Head Motion and Pose Estimation

Head motion is estimated using a 3D model-based tracking algorithm. In our approach, the head of the user is modeled using a 3D mesh, which consists of a set of 3D points $M_i = (X_i, Y_i, Z_i)$.

Let $\mathbf{R}$ and $\boldsymbol{t}$ be the rotation and translation of the head between time t and $t+1$. In case of small motions, the rotation $\mathbf{R}$ can be parameterized by:

$$\mathbf{R} = \mathbf{I} + \begin{pmatrix} 0 & -w_z & w_y \\ w_z & 0 & -w_x \\ -w_y & w_x & 0 \end{pmatrix} \quad (1)$$

The rigid transformation can then be parameterized by a 6-vector $\delta = (w_x, w_y, w_z, \boldsymbol{t})^\top$. Given the location and orientation of the model (*i.e.* location of the points M_i) at time t, the rigid motion δ of the head is estimated as follows:

Let M_i' be the 3D location of point M_i at time $t+1$. Let m_i and m_i' be the respective projections of M_i and M_i' in the image. Let f and (u_0, v_0) be the focal length and principal point location of the camera respectively. Using the camera as the reference, we can write:

$$m_i = \frac{1}{Z_i}\mathbf{P}M_i \qquad m_i' = \frac{1}{Z_i'}\mathbf{P}M_i' \quad (2)$$

where $\mathbf{P}$ is the camera-projection matrix defined as

$$\mathbf{P} = \begin{pmatrix} f & 0 & u_0 \\ 0 & f & v_0 \end{pmatrix} \quad (3)$$

If the face moves rigidly then all points M_i move according to the same rigid transformation, *i.e.* $M_i' = \mathbf{R}M_i + \boldsymbol{t}$. Let $\vec{u}_i = m_i' - m_i$ be the optical flow estimated at point m_i. By combining eq. (2) and (1) for all i's, δ can be found as the solution of a linear system

$$\mathbf{A}\delta = \boldsymbol{b} \quad (4)$$

where $\mathbf{A}$ is a matrix and $\boldsymbol{b}$ a vector, whose entries depend on f, (u_0, v_0), M_i and $\vec{u}_i$ (for all i's).

If the head performs a perfectly rigid motion (*i.e.* all points M_i perform the same rigid transformation), eq. (4) can be solved exactly using a standard linear least-squares technique. However, in order to account for outliers, i.e., points M_i corresponding to non-rigid parts of the face (e.g., eyes, mouth), a robust estimation algorithm needs to be used. Because of its simplicity and performance, we chose to employ the RANSAC [8] algorithm. The RANSAC algorithm is able to find the dominant rigid motion of the face even if half the points M_i perform some 'outlier' motions.

Our face tracking algorithm is initialized with the Viola-Jones face detector [16]. In order to compensate for drift, we implemented our face motion and pose estimation algorithm in a keyframe-based estimation framework similar to [11].

3.2 Facial Motion Features

Let δ be the dominant rigid motion of the face. The facial motion features are defined as the head motion-compensated optical flow, i.e., the optical flow between the images I_t and I_{t+1} from which the motion δ has been 'subtracted'. The facial motion features correspond to the local non-rigid motion generated by the muscles of the face only (e.g., lips, jaw, eyes) independently from the global head motion.

In order to estimate the motion-compensated optical flow, we proceed as follows:

Let $N_i = (X_i^N, Y_i^N, Z_i^N)$ be the 3D coordinates of point M_i undergoing the exact rigid transformation δ, *i.e.* $N_i = \mathbf{R}M_i + \boldsymbol{t}$. Let $n_i = \frac{1}{Z_i^N}\mathbf{P}N_i$ be the projection of N_i in the image. We defined the motion-compensated optical flow $\vec{u}_i^0$ as the discrepancy between the rigid motion-induced flow $\vec{v}_i = n_i - m_i$ and the measured optical flow $\vec{u}_i$.

$$\vec{u}_i^0 = n_i - m_i - \vec{u}_i = \frac{1}{Z_i^N}\mathbf{P}(\mathbf{R}N_i + \boldsymbol{t}) - m_i - \vec{u}_i$$

In our framework, the facial motion features are defined as:

$$V_f = (\vec{u}_1^0, \ldots, \vec{u}_N^0)$$

3.3 Classification

Once the visual features are estimated, we project the features into a ten dimensional subspace using principle component analysis for dimensionality reduction. These projected features are subsequently used to train two HMMs [14], one trained based on data consisting of communication errors and the other trained on data with no communication errors. During evaluation, test sequences are evaluated under each model and the model with the highest likelihood is selected as the recognized communication state. The performance of these HMMs are evaluated on two datasets described below.

4. DATASETS

Two datasets, one consisting of staged expressions while the other consisting of natural ones, are used to evaluate the effectiveness of our approach in communication error detection.

4.1 Staged dataset

In this dataset, each subject is asked to pose in front of a camera and demonstrate non-verbal facial expressions that would occur when he/she is listening to the response of a conversational system. For communication errors, the subject presumes the system is giving an incorrect reply. Subjects posed facial expressions depicting varying levels or frustration or confusion. For non-communication errors or neutral situations, each subject presumes the system is giving a correct reply. To vary the intensity of the expressions, scenarios containing different numbers of correct system replies are presented before the subject is asked to articulate his/her facial expression. Eight posed facial expressions, each lasting on average three seconds, are collected per subject. Data from a total of ten subjects is collected and the facial motion and head pose features are extracted for our experiments.

4.2 Natural dataset

We want to investigate how effective our approach is in detecting communication errors in a natural setting. For preliminary evaluation, we acquired the dataset collected by Barkhuysen et al. [2] for system questions. In this dataset, subjects have made a train ticket reservation and are listening to system verification questions such as "So you want to travel to Amsterdam?". Nine subjects' video footage during the system's verification response is used for our experiments. More details of this dataset could be found in [2].

5. EXPERIMENTS AND RESULTS

On the staged dataset, we used leave-one-out cross validation (nine subjects' dataset were used for training and the remaining one was used for testing), with randomly generated partitions. For each test set, we plot the ROC curve for communication error detection. We compute the mean and standard deviation of these individual ROC curves and plot an averaged ROC curve with error bars shown in Figure 2. From this figure, the detection rate was significantly above chance; however, this performance was evaluated on a staged dataset and the performance on a natural dataset may differ.

On the natural dataset, we conducted several experiments. In the first experiment, we compared the performance of a user-independent model versus human performance. We conducted a nine-fold cross validation and made sure our Hidden Markov models are not trained on any data coming from a subject in the test set. In the same manner as described in the experiments on the staged dataset, we compute and plot the averaged ROC curve with error bars for communication error detection. This is shown in Figure 3, where the averaged ROC curve was plotted in a solid blue line. To compare against human performance, we conducted a human perception experiment, where video segments evaluated by the HMMs were also evaluated by four human observers. These four human observers did not view any sample video footage of the subjects before the evaluation. The results of the four human observers were denoted as four green circles in Figure 3. The difference in detection

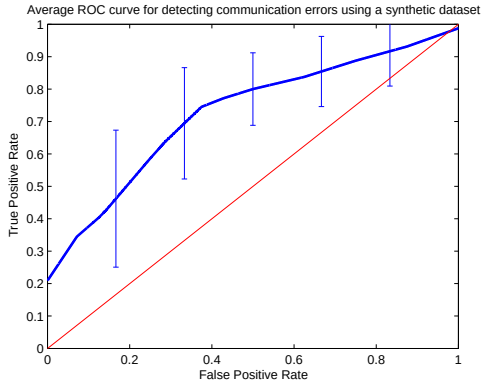

Figure 2: Average ROC curve and error bars for detecting communication errors on a staged dataset. The red line, which is used as a reference here, represents detection accuracy at 50%.

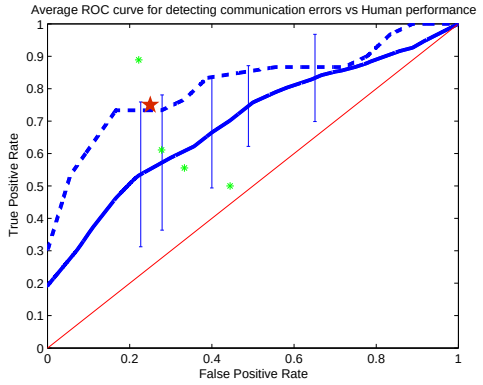

Figure 3: ROC curves, error bars and human perception performance for detecting communication errors on a natural dataset. For the user-independent experiment, the averaged ROC curve is plotted in a solid blue line with error bars. The green circles represent the performance of four human observers. For the partially user-dependent experiment, the ROC curve is plotted as a dashed blue line and the red star represents the average performance of human observers reported in [2].

accuracy between humans and HMMs was not statistically significant after performing a paired t-test ($p = 0.78$). This showed that the detection accuracy by HMMs was indistinguishable from human performance.

In the second experiment, we compared the partially user-dependent model against the human performance reported in [2]. We trained and tested HMMs, such that video clips from the same subject were used in training and testing. Note that we made sure the test set consisted of the same clips used for human evaluation in [2]. On average, each subject had sixteen video clips used in the training set and four video clips for the test set. The ROC curve is plotted in a dashed blue line in Figure 3. The average human performance reported in [2] is denoted as a red star in the same figure. From this figure, the HMMs attained similar performance as human observers.

6. CONCLUSIONS AND DISCUSSION

In this paper we describe work in progress towards the automatic detection of communication errors using visual cues during the system's conversational turn. Significant detection accuracy on the staged and the natural dataset lends evidence that such a detection formulation is useful. The visual cues captured in the natural dataset all occurred during a system's verification questions. As future work, we plan to explore the strength of this effect during other types of system response and develop better algorithms to improve detection accuracy.

7. ACKNOWLEDGMENTS

We would like to thank Stephanie Shattuck-Hufnagel, for her advice with the collection of the staged dataset, Marc Swerts and Pashiera Barkhuysen for their generous provision of their dataset for our analysis. The experiments and results would not have been possible without their prompt help. Pashiera put in a lot of work to send us the dataset.

8. 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 *ICSLP*, 2002.

[2] P. Barkhuysen, E. Krahmer and M. Swerts. Audiovisual Perception of Communication Problems. In *Speech Prosody*, 2004.

[3] L. Chen and T. S. Huang. Emotional expressions in audiovisual human computer interaction. In *ICME*, 2000.

[4] I. Cohen, N. Sebe, A. Garg, L. Chen, and T.S. Huang. Facial expression recognition from video sequences: Temporal and static modeling. In *CVIU*, volume 91(1-2), pages 160–187, 2003.

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

[6] P. Ekman. Emotion in the Human Face. Cambridge University Press,1982.

[7] B. Fasel and J. Luettin. Automatic facial expression analysis: A survey. In *Patt. Recogn.*, volume 36,pages 259–275, 2003.

[8] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In *Communications of the ACM*, volume 24(6),pages 381–395, 1981.

[9] J. Hirschberg, D. Litman and M. Swerts. Identifying User Corrections Automatically in Spoken Dialogue Systems. In *2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, 2001.

[10] D. Litman, J. Hirschberg and M. Swerts. Predicting user reactions to system error. In *ACL*, 2001.

[11] L. Morency, A. Rahimi and T. Darrell. Adaptive view-based appearance models. In *CVPR*, pages 803–810, 2003.

[12] S. L. Oviatt and R. VanGent. Error Resolution During Multimodal Human-Computer Interaction. In *SpeechCommunication*, 1998.

[13] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. In *IEEE Trans. on PAMI*, volume 22(12),pages 1424–1445, 2000.

[14] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77(2),pages 257–286, 1989.

[15] M. Song, J. Bu, C. Chen and N. Li. Audio-visual based emotion recognition - a new approach. In *CVPR*, 2004.

[16] P. Viola and M.J. Jones. Robust Real-Time Face Detection In *IJCV*, volume 57(2),pages 137–154, 2004.

[17] Y. Yoshitomi, S. Kim, T. Kawano and T. Kitazoe. Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In *IEEE International Workshop on Robot and Human Interactive Communication*, 2000.

[18] Z. Zeng, J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T. S. Huang, D. Roth and S. Levinson. Bimodal HCI-related Affect Recognition In *ICMI*, 2004.