MULTIMODAL COMMUNICATION ERROR DETECTION FOR DRIVER-CAR INTERACTION

Sy Bor Wang, David Demirdjian, Trevor Darrell
Massachusetts Institute of Technology, 32 Vassar Street, Cambridge MA 02139, USA
sybor@csail.mit.edu, demirdjian@csail.mit.edu

Hedvig Kjellström
KTH (Royal Institute of Technology), CSC, SE-100 44 Stockholm, Sweden
hedvig@csail.mit.edu

Keywords: Audio-Visual Recognition, System Error Identification, Conversational systems.

Abstract: Speech recognition systems are now used in a wide variety of domains. They have recently been introduced in cars for hand-free control of radio, cell-phone and navigation applications. However, due to the ambient noise in the car recognition errors are relatively frequent. This paper tackles the problem of detecting when such recognition errors occur from the driver's reaction. Automatic detection of communication errors in dialogue-based systems has been explored extensively in the speech community. The detection is most often based on prosody cues such as intensity and pitch. However, recent perceptual studies indicate that the detection can be improved significantly if both acoustic and visual modalities are taken into account. To this end, we present a framework for automatic audio-visual detection of communication errors.

1 INTRODUCTION

In recent years, there has been an increased interest in more intelligent and emotional car interfaces. This has been motivated by the need to make driver-car interactions easier and more natural and to reduce the cognitive load of the driver, who is now confronted with multiple device, e.g. phone, radio, navigation system. To this extent, speech recognition technology has recently been introduced in the car. However, due to the difficulty of the environment (e.g. noise produced by the engine, vibrations), speech recognition is still brittle and satisfying recognition rates can be obtained only for relatively small vocabularies, limiting the extent of the driver-car interaction. A speech recognition system that can automatically detect recognition errors would allow for smoother interaction.

Many spoken dialogue systems have difficulty determining whether the communication is going well or has problems (e.g. due to poor speech recognition). Various researchers have shown that human users change their speaking style when the system misrecognizes their speech as compared to when the system correctly recognized their speech (Hirschberg et al., 2001; Litman et al., 2001; Oviatt and VanGent, 1998). For example, users tend to speak slower or louder when speech misrecognition occurs. Such a change in speaking style usually leads to worse recognition results since standard speech recognizers are trained on normal, non-hyperarticulated speech (Oviatt and VanGent, 1998). These problems motivated the monitoring of prosodic aspects of a speaker's utterances, and several studies have shown that using automatically extracted prosodic features helps in error detection (Litman et al., 2001). However, the level of effectiveness of these prosodic features differs across studies and the analysis of prosodic features are done only on user utterances and not on audio cues of users while they are listening to the system response. Such limitations hint at the possible use of additional modalities or other types of features (e.g. visual features) to improve error detection.

The co-occurrence of audio and visual modalities has been widely explored for emotion recognition. Recent work in multimodal recognition of emotions has shown that a combination of prosodic features and facial expressions improves affect recognition (Zeng et al., 2004). The primary aim of our work in this paper is to find an automatic system to detect communication errors in a conversational system. Using both visual and audio features, we com-

pare the performance of different classifiers in the unimodal stream and different audio-visual fusion strategies for identification, taking into account asynchrony between acoustic and visual user reactions, using audio and video data of user interactions with a dialogue system in a natural setting.

2 RELATED WORK

There has been limited literature on the use of low-level audio cues and visual features in automatically detecting dialogue-based system errors in an authentic environment. A perceptual study conducted by Barkhuysen et al. (Barkhuysen et al., 2004) showed that audio and visual cues are useful in detecting communication errors. The study also showed that using visual cues were very effective for detecting system errors when the user is listening in silence to the reply from the dialog manager. In this study though, subjects were specifically instructed to face a camera embedded in a cellphone while speaking to it. Knowledge of this camera could bias the subject's behavior. As shown by Sebe et al. (Sebe et al., 2004), this knowledge bias was significant for learning facial expressions. In this work, subjects were viewing movie clips in a kiosk, without any knowledge of a camera capturing their facial expressions. However, no prosody or audio cues of the subjects were collected.

Recent work done in emotion or affect recognition has explored the combined use of prosody and facial features (Zeng et al., 2004; Chen et al., 1998). Zeng et al. (Zeng et al., 2004) used a voting method to combine the facial feature and prosody classifiers to improve affect recognition. Although this work addressed the difficult task of classifying eleven emotional states, it suffers from the use of a database where subjects generated emotions upon request, which may not be the genuine expressions in an authentic environment.

In the domain of detecting communication errors, also known as system errors, audio cues have been explored widely. Oviatt (Oviatt and VanGent, 1998) showed that there is a pattern of hyper-articulation when there are system errors, which leads to worse recognition results. Litman et al. (Litman et al., 2001) and Hirschberg et al. (Hirschberg et al., 2001) automatically extracted prosodic features of a speaker's utterances and showed that these features have been useful in error detection, although the extent to which prosody is beneficial differs across studies. This implies that the accuracy of error detection can be improved by the addition of other features, e.g. visual cues, either as a combination with audio cues or simply as a separate stream.

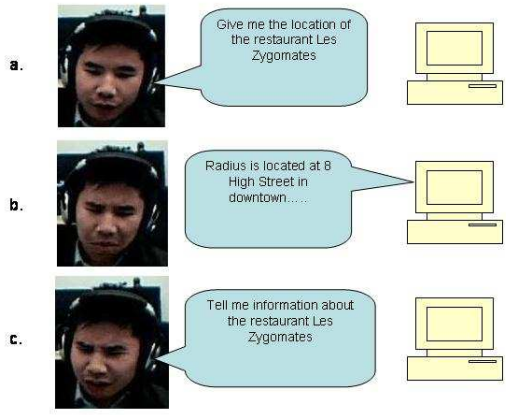

Figure 1: Illustration of communication errors. In a., the subject is making a query of a restaurant the first time. In b., the subject is listening to the response of the system. And in c., the subject repeats his query. The facial expression of the user in b. and c., as well as the tone of the user's voice in c. are the cues our algorithm exploits for the detection of communication errors.

In this paper we propose to detect communication errors multimodally by using visual and audio features. We focus on an environment where the user is interacting with a conversational information query system similar to the ones present in car navigation systems. We define a communication error as the instance where the system misrecognizes the driver speech, and makes an erroneous reply. This is illustrated in Figure 1. We detect the presence of this error when the user is speaking or when the user is listening to the system.

The rest of the paper is described as follows. First, we describe the features we extract from audio and visual streams to measure confusion. Then we give a description of our classification models and late fusion strategies, followed by our experiment setup and procedure. In the last section, we show comparative results of the different classifiers.

3 MULTIMODAL INPUT

3.1 Visual Component

In this section we describe our algorithm for estimating head pose and facial motion features from monocular image sequences. In our framework, head mo-

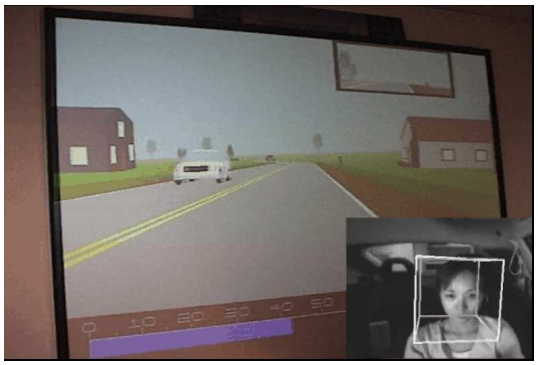

Figure 2: Audio-visual platform installed in a car simulator. The face of the driver is tracked using a camera installed near the rear view mirror: the white cube surrounding the driver's face corresponds to Π, the location and orientation of the pose estimate.

tion is decomposed into two distinct components. The first component consists of the 3D rigid motion of the head. The second component consists of the local motion generated by the non-rigid parts of the face (e.g. mouth, lips, eyes).

The head motion estimation algorithm consists of first estimating the rigid motion component of the head by using a robust algorithm. This rigid motion then is used to compensate the global motion of the head and to estimate the local motion of the non-rigid parts.

3.1.1 Head Motion and Pose Estimation

The algorithm for head motion and pose estimation is an implementation of the drift-free tracking technique described in (Morency et al., 2003). In contrast to the original work, which required stereo inputs, our implementation estimates head pose in monocular image sequences. In addition motion estimation is based on the robust estimator RANSAC (Fischler and Bolles, 1981) which allows large noise in the data (e.g. temporary occlusions of the face). The face tracking algorithm is initialized in a standard fashion using a frontal-view face detector (Viola and Jones, 2004).

Our algorithm provides the head pose in a 3D coordinate system and is represented by a 6-vector Π:

$$\Pi = (t_x, t_y, t_z, \phi_\alpha, \phi_\beta, \phi_\gamma) \quad (1)$$

where (t_x, t_y, t_z) is the location and $(\phi_\alpha, \phi_\beta, \phi_\gamma)$ the orientation (pan, tilt and swing angles) of the head.

3.1.2 Facial Motion Features

Let δ be the rigid motion between the last two frames $t-1$ and t. The facial motion features are defined as

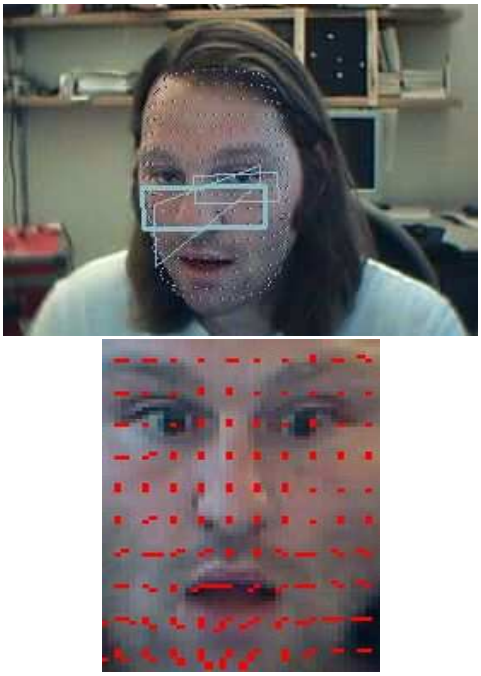

Figure 3: Our head motion estimation algorithm consists in estimating the rigid motion component (left image) and compensated optical flow on the face surface (right image).

the head motion-compensated optical flow, i.e. the optical flow between the images I_{t-1} and I_t 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), independent from the global head motion. In our framework, the vision-based features are defined as:

$$V_f = (\bar{u}(m_1), \ldots, \bar{u}(m_N)) \quad (2)$$

where $\bar{u}(m_k)$ is the head motion-compensated optical flow for a point m_k of the face.

3.2 Audio Component

We use three kinds of prosody features: the intensity E, the pitch, F_0, and the first formant frequency, F_1. The prosody feature vector A_f is then defined as:

$$A_f = (E, F_0, F_1) \quad (3)$$

These features are computed at every 10 msecs using the speech analysis software, PRAAT (Boersma, 2001). The intensity E is computed as: $E = log(\sum_{i=1}^{N}(x[i]-\bar{x})^2)$ where N is the window length and $x[i]$ is the ith sample in that window and $\bar{x}$ is the local average signal value. In our computation (and for the rest of this section) we used a window length of 40 samples. The pitch F_0 is estimated as the reciprocal

of the fundamental period as described in (Boersma, 1993). In our experiments, we set the search range of the pitch to be 75 - 1000 Hz. As for the computation of the first formant frequency, F_1, a segment of N samples is extracted for every time step of 1 msec. This segment is multiplied by a Gaussian-like window and the LPC coefficients are computed. This first formant is then extracted using these coefficients by the Burg algorithm described in (Childers, 1978).

In previous work (Sebe et al., 2004) syllable rate was used as a prosody feature. However, in our work, our audio data consists of spoken as well as non-spoken words, e.g. exclamations, gasps or humming, which we want to model for automatic problem detection. and our speech recognizer had a lot of difficulty computing an accurate syllable rate. Of the 219 utterances processed by the speech recognizer, 97 utterances have an incorrect number of hypothesized vowel phones. On average, these incorrectly recognized utterances have 2.73 syllables more than the hypothesized ones.

4 Multimodal Detection of System Errors

We explore different techniques to detect communication errors from sequences of audio-visual features estimated in Section 3.2. First, we describe unimodal classification models followed by the multimodal fusion strategies we tested.

4.1 Unimodal Classification Methods

We want to map an observation sequence $\mathbf{x}$ to class labels $y \in \mathcal{Y}$, where $\mathbf{x}$ is a vector of t consecutive observations, $\mathbf{x} = \{x_1, x_2, \ldots x_t\}$. In our case, the local observation x_t can be an audio feature A_f, or a visual feature, V_f.

To detect communication errors, learning the sequential dynamics of these observations is important. Hidden Markov Models (HMMs) (Rabiner, 1989) are well known generative probabilistic sequence models that capture sequence dynamics; Hidden Conditional Random Fields (HCRFs) (Quattoni et al., 2004; Wang et al., 2006) are discriminative analogs that have been recently introduced for gesture recognition. We compare both techniques in our experiments below; experiments with classifiers taking a single observation as input previously demonstrated poor results, and were not included in our experiments.

Hidden Markov Models (HMM) - We trained a HMM model for each communication state. During evaluation, test sequences were passed through each of these models and the model with the highest likelihood was selected as the recognized communication state. This is a generative, sequential model with hidden states. More details of this model are described in (Rabiner, 1989).

Hidden Conditional Random Fields (HCRF) - The HCRF is a model that has recently been introduced for the recognition of observation sequences (Quattoni et al., 2004). Here we describe the HCRF model briefly:

A HCRF models the conditional probability of a class label given an observation sequence by:

$$P(y \mid \mathbf{x}, \theta) = \sum_{\mathbf{s}} P(y, \mathbf{s} \mid \mathbf{x}, \theta) = \frac{\sum_{\mathbf{s}} e^{\Psi(y, \mathbf{s}, \mathbf{x}; \theta)}}{\sum_{y' \in \mathcal{Y}, \mathbf{s} \in S^m} e^{\Psi(y', \mathbf{s}, \mathbf{x}; \theta)}} \quad (4)$$

where $\mathbf{s} = \{s_1, s_2, \ldots, s_m\}$, each $s_i \in S$ captures certain underlying structure of each class and S is the set of hidden states in the model. If we assume that $\mathbf{s}$ is observed and that there is a single class label y then the conditional probability of $\mathbf{s}$ given $\mathbf{x}$ becomes a regular CRF. The potential function $\Psi(y, \mathbf{s}, \mathbf{x}; \theta) \in \Re$, parameterized by θ, measures the compatibility between a label, the observation sequence, and the configuration of the hidden states.

In our paper, the local observations are the visual features, V_f, or the audio features, A_f. We trained a single two-class HCRF. Test sequences were run with this model and the communication state class with the highest probability was selected as the recognized error state.

For the HMM model, the number of Gaussian mixtures and states was set by minimizing the error on training features. For the HCRF model, the number of hidden states was set in a similar fashion.

4.2 Multimodal Fusion Strategies

We have a choice between early or late fusion when combining the audio and visual modalities. In early fusion, we can model the audio and visual features in a single joint feature space, and use the joint feature for training a single classifier. In late fusion, we can train a classifier on each modality separately and merge the outputs of the classifiers. As illustrated in Figure 1, our communication error detection has two different modes: in b. we use visual features only for error detection and in c. we use both audio and visual features. The single mode in b. requires us to train a classifier using a single input stream. In addition, training classifiers based on individual streams is a simpler process. As such, we choose late fusion techniques, i.e. fusing the outputs of two classifiers. We use two common late-fusion strategies as described in (Kittler et al., 1998).

Let the feature input to the j-th classifier, $j = 1,...,R$ be x_j, and the winning label be h. A uniform prior across all classes is assumed.

PRODUCT rule: $h = \arg\max_k \prod_{j=1}^{R} P(w_k|x_j)$.

With the product rule, we multiply the probabilities of the visual feature classifier and the audio feature classifier, and select the winning class based on the highest scoring multiplication.

SUM rule: $h = \arg\max_k \sum_{j=1}^{R} P(w_k|x_j)$.

With the sum rule, we add the probabilities of the visual feature classifier and the audio feature classifier, and select the winning class based on the highest scoring sum.

5 Experiments and Results

5.1 Data Collection

To evaluate the performance of the different classifiers and fusion strategies, we collected an audio-visual database where the facial expressions and the audio cues would correspond to the actual conversational state of the subject. There were several design issues we had to consider to minimize bias of our data collection experiment. These issues were approached in a similar fashion to the database collected for natural facial expressions (Sebe et al., 2004). In particular, the subjects could not know that they were being tested for their communication state [1]. Such knowledge could influence their communication state and invalidate the results.

We set up a conversational kiosk with a hidden camera and microphone array. This kiosk contained a web-based navigation query speech interface consisting of a display showing a Google map with restaurant icons. Subjects did not know that they were involved in an experiment about communication error detection: they were told their task was to test the navigation system and report what they liked or disliked about it at the end of the experiment. They were given a list of restaurants to query for information. The subjects had to make the queries in sequential order, and repeat the query in any way they wished (e.g. repeating the same question, or using a different phrase) when the system did not respond correctly. They could only proceed to the next query when the system displayed the correct restaurant information on the display. The purpose of this sequential query was to create a need to solve the communication error when it occurs. The audio and video of the user were recorded throughout the whole experiment under brightly lit and low noise conditions. The video was recorded at 15 Hz and the audio at 44kHz. A total of six subjects performed the study. All six subjects were male and were between 20-30 years of age.

From the database described above, all the sequences were manually labeled and segmented for training our classifiers described in Section 4.1. A total of 227 error sequences and 84 error-free ones were collected. For each human subject, 90% of his/her data were picked at random and used for training, while the remaining ones were used for testing.

5.2 Testing and Evaluation

5.2.1 Visual Features Classification

Facial motion features, V_f, described in Section 3.1 are used as observations for training and testing. Figure 4 shows the results of the classifiers described in Section 4.1 using visual features. From this figure, HCRF performs better than HMMs for visual feature classification.

5.2.2 Prosody Features Classification

Using prosody features, A_f, from Section 3.2 as observations, we trained two classifiers described in Section 4.1. Figure 4 shows the ROC curves for the different classifiers. From this figure, both HCRFs and HMMs perform poorly for prosody feature classification. This is due to the use of only three acoustic characteristics as our prosody features and shows that such features are not very indicative of communication problems.

5.2.3 Audio-Visual Classification

We compared the performance of HMMs and HCRFs in the late fusion experiments. Figure 5 shows the ROC curve of the combining the various classifiers using the SUM and PRODUCT rule. The classifiers show a significant improvement using the late fusion strategies, despite a poor performance when only prosody features are used. Interestingly, this concurs with findings in the area of audio-visual speech perception (Massaro, 1987; Summerfield, 1987), indicating that humans fuse information from different modalities at a relatively late stage. In addition, fusion of HCRF classifiers performed better than fusion

[1] At the end of the experiment, we procured agreement for the use of the audio-visual footage from the subjects for our experiments.

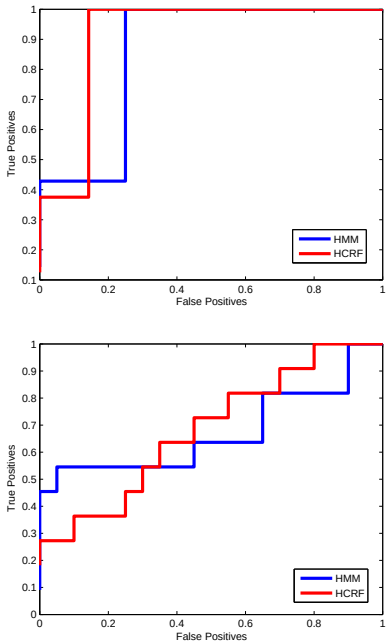

Figure 4: (top) ROC curves showing the performance of the different classifiers of visual features. (bottom) ROC curves showing the performance of the different classifiers using prosody features. From this figure, HMMS and HCRFs do not perform well on the prosody features, while HCRFs perform much better than HMMs on visual features.

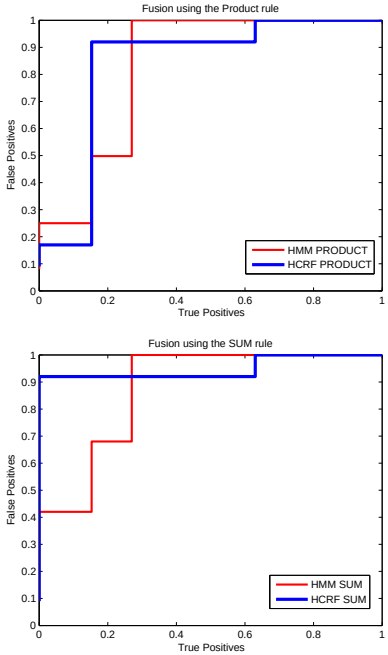

Figure 5: (top) ROC curve showing the performance of different classifiers using PRODUCT rule for fusion.(bottom) ROC curve showing the performance of different classifiers using SUM rule for fusion. Both curves show that despite poor performance from the audio stream, late fusion improved the performance significantly. In addition, HCRFs perform better than HMMs after fusion.

of HMM classifiers. Using the SUM rule to combine the HCRF classifiers produced the best result.

6 Conclusions

In this paper, we presented experiments evaluating different classification and fusion methods for detecting communication errors in a conversational system. Authentic audio-visual data of human-dialogue interactions with the conversational system was collected and labeled according to the presence of communication errors, and used to train and test the automatic system.

Features extracted from the audio included different prosody characteristics, such as pitch and intensity and the first formant frequency. From the visual channel, the user's global head motion and local face motion were extracted. Different strategies for classification from these cues were evaluated, as well as strategies for fusion of the two modalities. Despite the problems associated with asynchrony of audio and visual features, we used simple late fusion strategies of our HCRF and HMM classifiers and showed that they have improved error detection. To summarize, we find that communication errors in a dialogue-based system can be detected with a better accuracy using a HCRF with audio-visual input and a fusion strategy using the SUM rule.

REFERENCES

Barkhuysen, P., Krahmer, E., and Swerts, M. (2004). Audiovisual perception of communication problems. In *Speech Prosody*.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonis-to-noise ratio of a sampled sound. In *IFA*.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Chen, L., Huang, T. S., Miyasato, T., and Nakatsu, R. (1998). Multimodal human emotion/expression recognition. In *International Conference on Face and Gesture Recognition*.

Childers, D. G. (1978). *Modern Spectrum Analysis*. IEEE Press.

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

Hirschberg, J., Litman, D., and Swerts, M. (2001). 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*.

Kittler, J., Hatef, M., Duin, R., and Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239.

Litman, D., Hirschberg, J., and Swerts, M. (2001). Predicting user reactions to system error. In *ACL*.

Massaro, D. (1987). *Speech Perception By Ear and Eye*. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Morency, L., Rahimi, A., and Darrell, T. (2003). Adaptive view-based appearance models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 803–810.

Oviatt, S. L. and VanGent, R. (1998). Error resolution during multimodal human-computer interaction. In *SpeechCommunication*.

Quattoni, A., Collins, M., and Darrell, T. (2004). Conditional random fields for object recognition. In *Neural Information Processing Systems*.

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

Sebe, N., Lew, M., Cohen, I., Sun, Y., Gevers, T., and Huang, T. S. (2004). Authentic facial expression analysis. In *International Conference on Automatic Face and Gesture Recognition*.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd, B. and Campbell, R., editors, *Hearing by Eye*, pages 3–51. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Wang, S., Quattoni., A., Morency, L.-P., Demirdjian, D., and Darrell, T. (2006). Hidden conditional random fields for gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zeng, Z., Tu, J., Liu, M., Zhang, T., Rizzolo, N., Zhang, Z., Huang, T. S., Roth, D., and Levinson, S. (2004). Bimodal hci-related affect recognition. In *International Conference on Multimodal Interfaces*.