Human Perception of Audio-Visual Synthetic Character Emotion Expression in the Presence of Ambiguous and Conflicting Information

Emily Mower [#1], Maja J Matarić [*2], Shrikanth Narayanan [#*3]

[#] *Department of Electrical Engineering, University of Southern California*
University Park, Los Angeles, California, USA 90089

[1] `mower@usc.edu`
[3] `shri@sipi.usc.edu`

[*] *Department of Computer Science, University of Southern California*
University Park, Los Angeles, California, USA 90089

[2] `mataric@usc.edu`

Abstract—Computer simulated avatars and humanoid robots have an increasingly prominent place in today's world. Acceptance of these synthetic characters depends on their ability to properly and recognizably convey basic emotion states to a user population. This study presents an analysis of the interaction between emotional audio (human voice) and video (simple animation) cues. The emotional relevance of the channels is analyzed with respect to their effect on human perception and through the study of the extracted audio-visual features that contribute most prominently to human perception. As a result of the unequal level of expressivity across the two channels, the audio was shown to bias the perception of the evaluators. However, even in the presence of a strong audio bias, the video data were shown to affect human perception. The feature sets extracted from emotionally matched audio-visual displays contained both audio and video features while feature sets resulting from emotionally mismatched audio-visual displays contained only audio information. This result indicates that observers integrate natural audio cues and synthetic video cues only when the information expressed is in congruence. It is therefore important to properly design the presentation of audio-visual cues as incorrect design may cause observers to ignore the information conveyed in one of the channels.

I. INTRODUCTION

Robots and animated characters are utilized in roles ranging from tutors to caregivers. These characters require well defined emotional models in order to interact naturally and socially with humans [1]–[3]. Accompanying these internal emotion models are external emotional behaviors (e.g., lips that smile, eyebrows that lower menacingly) [4]. The accuracy of these expressions strongly impacts the resulting social interaction between the human user and the synthetic agent [2]. An agent with a very detailed internal emotional model, but with a poorly designed emotional interface, may be limited in its ability to interact emotionally with a human user. For example, imagine a robot designed to empathize with an individual. Although this robot may have the ability to correctly recognize the state of a user, if, due to poor design, it is not able to express the socially recognized cues of empathy, the human may misinterpret the robot's actions. It is important to

have quantitative models of emotion perception to avoid such emotional expression design missteps.

The work presented in this paper is a quantitative analysis of human observers' emotional perception evaluations of video clips of simple emotionally animated facial expressions synchronized with emotional human vocalizations. The evaluation was designed to illuminate how observers utilize emotional cues in the presence of facial and vocal emotional mismatch and in the presence of expressivity mismatch (the human voice utilized in this study is more expressive than the animated face). This experimental structure provides a presentation framework over which many possible emotional stimulus combinations, not seen absent this mismatch, can be analyzed.

Human audio-visual expression can be both complementary and supplementary. Supplementary interactions occur when the information presented in each channel enhances the emotional experience. Complementary interactions occur when one channel contains significant information not present in the other channel, resulting in a combination that completes the experience. If the information contained within the two channels is supplementary, the experience of an audio-visual emotion would not differ greatly from that of an audio-only or video-only emotion display. However, there is evidence to suggest that human emotion perception relies on the complementarity of emotional displays. This necessitates the proper design of all channels that convey emotion information. For example, Mr. Bill [5], a Saturday Night Live Play-Doh character, has an extremely simple facial design combined with a high-pitched voice. It is difficult to determine the expressed emotion using either only the video (the face could indicate surprise, fear, irony) or only the audio (the degree of fear is difficult to pinpoint). However, when combined, the character of Mr. Bill is able to create recognizable expressions of both calmness and fear. This character presents one example of the perceptual difference between a complementary audio-visual emotion expression and that of an audio-only or video-only expression.

Complementary audio-visual emotion processing can be

studied through the creation of artificially generated stimuli. These stimuli can be separated into two types of presentations: *congruent* presentations, in which the emotions expressed in the facial and vocal channels match, and *conflicting* presentations, in which the emotions expressed across the two channels are mismatched. Conflicting and congruent presentations are important because they allow for the identification of audio-visual features that are emotionally salient in the presence of both emotionally relevant and emotionally ambiguous information. Features that are important during an emotionally congruent presentation are features that users integrate to determine the cohesive emotional expression. Features that impact emotion perception in the presence of ambiguous or discordant information are features that allow users to bypass emotional noise to interpret the emotional message of the utterance. Conflicting and congruent presentations allow researchers to determine how best to utilize the available audio-visual features given a synthetic interface that either possesses (as modeled by congruent presentations) or lacks (as modeled by conflicting presentations) the degrees of freedom in the individual channels necessary to express the emotion. These presentations also demonstrate the importance of proper multimodal cue integration as improper design results in increased user evaluation variance with respect to emotional perception.

Conflicting emotional presentations are not a purely artificial construction; they exist outside of the synthetic domain [6]. During human communication, the expressed audio-visual emotion may not represent the internal emotion state, or it may contain mismatched audio-visual information. Such inconsistencies may result from intended ambiguity (such as during emotion masking), effect (such as sarcasm), or physical limitations (such as a stroke). However, even in the presence of ambiguity, humans are able to detect underlying emotion states with varying degrees of accuracy by relying on facial, vocal, and/or body posture cues. The natural occurrence of congruent and conflicting displays in human-human communication suggests that the results obtained using this presentation method can be used to study human audio-visual joint emotional perception.

Synthetic interfaces, which are widely used in both research and entertainment domains, are also vulnerable to emotional ambiguity. Synthetic modalities, such as faces and voices, typically contain fewer degrees of freedom than their human counterparts. As a result, it may not be possible to always accurately express the intended emotions using a single modality (e.g., happiness can be expressed facially, but not when only the eyebrows are actuated). This ambiguity may result in unclear or conflicting information when the two modalities are combined.

Emotional ambiguity or emotional conflict may force observers to change the way they interpret emotional displays. This alteration can be used to study channel bias across multiple presentations. Channel bias analysis is the study of how observers weight the information presented across multiple channels to arrive at a final emotional assessment. This analysis is of particular importance when considering channels whose emotional information is modulated by multiple system constraints such as affect or speech goals. In these situations, the information presented in the individual channels may not fully express an intended emotion and it is the role of the observer to decide how to integrate this incomplete or incorrect information. The presence of channel bias is of particular relevance to this study, in which a natural human voice is combined with a synthetic animated face in congruent and conflicting presentations.

Despite the relative simplicity of some synthetic faces, animators have found effective ways to merge these faces with natural expressive voices (e.g., the animated series South Park). Even given the limitations of the simple animation, viewers are easily able to observe happiness and anger expressions in the characters. In research, the benefit to utilizing a simple animated face is that, in addition to providing a sufficient backdrop for emotional expression (especially when combined with an expressive voice), it allows for a reduced emotional search space. This decreases the number of feature combinations that evaluators must observe, and allows for a focus on a smaller number of features that are surmised to provide emotionally relevant information to observers. Furthermore, these simple animated faces either possess or can be made to possess some of the expressive limitations seen in many of today's robotic humanoid faces, which so far lack a large number of emotional degrees of freedom. Therefore, the use of these synthetic faces with reduced degrees of freedom may allow the results of these perceptual studies to extend to humanoid robot platforms.

Audio-visual human emotional perception has been studied using congruent and conflicting presentations [6]–[11]. Emotional feature sets have also been researched with respect to automatic emotion classification tasks [12]–[16]. However, the classification studies have dealt with feature set creation to optimize classification performance while the congruent and conflicting analyses have worked with either static images or single-word video presentations (rather than static images). The presented work proposes a method for analyzing this audio-visual interaction in the presence of emotionally ambiguous information.

The presented work is designed to provide a quantitative understanding of how humans integrate audio and video information during synthetic emotional presentations. This paper presents a summary of the results found in [17], [18] and an extension of the results found in [19]. It describes a methodology for the dimensional analysis of user perception over a broad input feature space. The goals of this study are to determine: 1) how audio and video information interact during the human emotional evaluation process and 2) the features that contribute to specific types of emotion perception. The audio features (pitch, energy, Mel Frequency Cepstral Coefficients, and speaking rate) and video features (eyebrow movement, eyebrow movement timing, eyebrow movement type, eyebrow angle, lip corner position, and eye shape) were tested using a combination of human (vocal) and synthetic (facial) stimuli. These stimuli were rated by evaluators along the dimensional scales of emotional valence, activation, and dominance ("VAD").

These data were analyzed using statistical techniques to

determine how the emotional evaluations varied as a function of presentation (audio-only, video-only, or audio-visual) and presentation type (conflicting or congruent). The data were analyzed using feature selection techniques to determine the audio-visual features utilized by evaluators during the emotional evaluations. The data were split using the presentation type (congruent vs. conflicting) to determine the features that contributed most to the explanation of the evaluators' variance within the two styles. Features that are prominent in the congruent presentations are those upon which evaluators rely when the information conveyed in the two channels can be easily integrated. Features that are prominent in the conflicting presentations are those that provide information allowing the evaluator to disambiguate between the information presented in the two channels. This implies that congruent salient features are important to consider when designing interfaces or emotional expressions with enough bandwidth to express a desired emotion across both channels while conflicting salient features are those that a designer must include when one of the expressive channels does not have the bandwidth necessary to properly express the desired emotion.

The perceptual evaluations of the observers were strongly biased by the audio information, due in part to the relatively high level of emotional expressivity when compared to that of the provided video information in the stimuli of the present study. Feature analyses indicated that even given this expressivity mismatch, both audio and video features were integrated in congruent emotional evaluations. However, when presented with conflicting emotional expressions, evaluators relied almost exclusively on the audio information. This work is novel in that it utilizes multiple sentences rather than a single word [8] or a single sentence [11], video segments rather than still images [6], [7], [9], [10], and a dimensional emotional evaluation (e.g, valence, activation, dominance) rather than a discrete emotional evaluation (e.g., angry, happy, sad, neutral) [6], [7], [9], [10]. The related work will be more fully discussed in Section II. Audio-visual evaluations of sentence-length emotional presentations may be more representative of the human perceptual process than those of either static images or single word utterances.

The remainder of this paper will explore human perceptions of a specific instantiation of a synthetic character's emotional expressions. Section II discusses previous research in this field. Section III describes the data used in this experiment. Section IV describes general trends observed in human perception. Section V quantifies the human perceptions with respect to the dimensions of emotion utilized in this study and the observed human biases with respect to channel attention (audio vs. video). Section VI describes the audio and video features with high predictive power with respect to the evaluators' dimensional emotion perception. Section VII summarizes the findings detailed in this paper. Finally, Section VIII provides conclusions and suggests future research directions.

II. RELATED WORK

This work was principally motivated by the McGurk effect [20]. The McGurk effect occurs when mismatched audio and video phonemes (called visemes) are presented to a human observer. Instead of perceiving either of the two presented phonemes, McGurk and MacDonald found that observers perceived a third, distinct phoneme. This finding has motivated many emotion research studies designed to determine if such an effect occurs within the emotion domain. Emotional McGurk evaluations are primarily conducted using either discrete emotion assignment (e.g., happy or angry) [6]–[11] or dimensional evaluation [17], [18], [21]. This effect has also been studied using fMRI [21] and EEG measurements [22]. The emotion presentations include congruent and conflicting information from the facial and vocal channels (e.g., [6]), facial channel and context (e.g., [21]), and facial and body postural/positional information (e.g., [22]).

In discrete choice evaluations, users are asked to rate the utterance by selecting the emotional label that best fits the data. Such evaluations allow researchers to determine at which point along an emotional continuum a given face or voice is of a sufficient emotional strength to bias the decision of the evaluators [7]. In a dimensional analysis evaluation, evaluators are asked to rate the presented stimuli according to the properties of those stimuli. Common properties (or dimensions) include valence (positive vs. negative), activation (calm vs. excited), and dominance (passive vs. dominant). One common dimensional evaluation technique utilizes Self-Assessment Manikins (SAM) [23]. This evaluation methodology presents the dimensions of valence, activation, and dominance using a pictorial, text-free display. This display method allows evaluators to ground their assessments using the provided end-points.

In [7], researchers combined still images with single spoken words in three distinct experiments. The first experiment presented images morphed from two archetypal happy and sad emotional images into a visual emotional continuum. These images were accompanied by either vocally happy or sad human utterances. The evaluators were presented with an audio-only, video-only, or combined audio-visual presentation and were asked to assign the combined presentation into one of the discrete emotional categories of "happy" or "sad." The researchers found that the evaluators were able to correctly recognize the emotion in the audio-clip 100% of the time. In the combined audio-visual presentation, they found that the voice altered the probability that an evaluator would identify the presentation as "sad." The researchers then repeated the experiment, this time asking the users to judge the face and ignore the voice. They found that the emotion presented in the audio channel still had an effect on the discrete emotion assignment. However, in this experiment, they found that the effect was smaller than that seen previously. In the final experiment, the researchers created a vocal continuum ranging from happiness to fear (to allow for a more natural vocal continuum). They asked the users to attune to the voice and to ignore the face. They found that users were still affected by the visually presented emotion.

A similar study was presented in [10]. In that work, evaluators were asked to rate 144 stimuli composed of still images and emotional speech using happy, angry, and neutral emotions. The evaluators were asked to respond as quickly

(a) Dimensional evaluation tool.

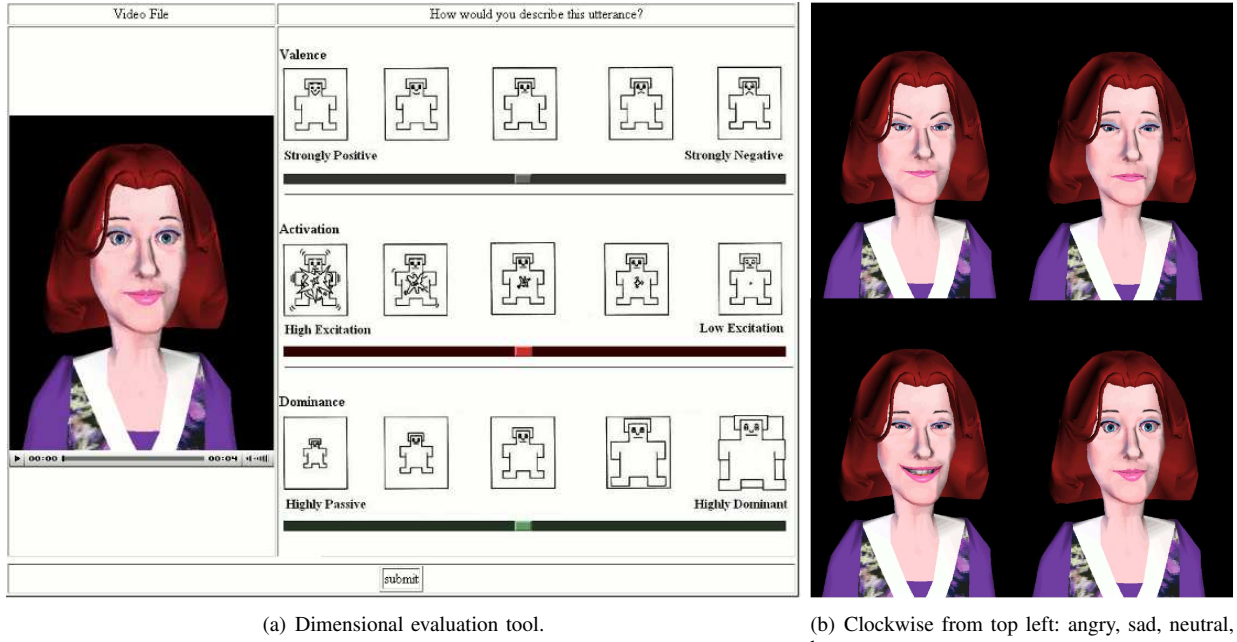(b) Clockwise from top left: angry, sad, neutral, happy.

Fig. 1. The online emotion evaluation interface (left) and frames of the four emotional presentations (right) used in this study.

as possible after viewing a stimulus presentation. In the first experiment, the researchers asked the evaluators to attune to the emotion presented in the facial channel. They found that the emotion presented in the vocal channel (the unattended emotion) affected the evaluators with respect to accuracy (discrete classification of facial emotion) and response time (faster for congruent presentations). When the evaluators were instead asked to attune to the facial channel, the researchers found that vocal channel mismatches decreased the facial emotion recognition accuracy and increased the response time. In the final experiment, users were presented with an emotional stimulus (a vocal utterance), a pause, and a second emotional stimulus (the facial emotion) used as a "go-signal." The researchers found that when the emotion presentations were separated by a delay, the channels no longer interacted in the emotion evaluation and the evaluators based their decisions on the vocal signal only.

Interactions between emotional channels have also been studied using emotional faces paired with contextual movies [21]. In that study, the evaluators were presented with four seconds of a movie and were then shown a static image. The contextual emotions included positive, negative, and neutral. The emotional faces included happy, fear, and neutral. These combined presentations were rated using SAMs. The researchers found that faces presented with a positive or negative context were rated significantly differently than faces presented in a neutral context. Furthermore, the fMRI data showed that pairings between faces and emotional movies resulted in enhanced BOLD responses in several brain regions.

The McGurk effect has also been studied with respect to body posture and facial analyses [22]. In this study, researchers paired emotional faces with emotional body positions (fear and anger for both) to analyze the interplay between facial and postural information in emotion evaluation. They found that evaluators were able to assess the emotion state (using

a discrete choice evaluation) of the stimulus most quickly and accurately when viewing congruent presentations. The results showed that the analysis time for faces-only was faster than for bodies-only. These results suggest that facial emotional assessment is biased by the emotion embedded in body posture.

The expression of emotion has also been studied in a more localized manner. One such method utilizes a "Bubble" [24]. This method is designed to identify regions of interest that correspond to task-related performance by only permitting users to view certain areas of the stimulus. The stimulus is covered by an opaque mask. Regions are randomly shown to the evaluators by creating Gaussian "bubbles," which allow users to glimpse regions of the masked stimulus. Given an infinite number of trials, all window combinations will be explored. The "Bubble" method allows for a systematic evaluation of stimuli components, but produces results that are difficult to translate into system design suggestions.

These past studies suggest that the video and audio channels interact during human emotion processing when presented synchronously. However, due to the discrete nature of the evaluation frameworks, it is difficult to determine how the perception of the targeted emotions change in the presence of conflicting information. The present work uses the dimensional evaluation method reported in [23] to ascertain the nature of the audio-visual channel interactions. This study also differs from previous work in its use of video clips rather than static photographs. Human-computer and human-robot interactions usually include dynamic faces and voices. The inclusion of dynamic facial stimuli in this study makes the results more transferable to the interactive design domain.

This work was also motivated by the analysis by synthesis approach to emotion research. In this approach, stimuli are synthesized using a broad range of parameters to determine combinations that result in enhanced performance. In [25],

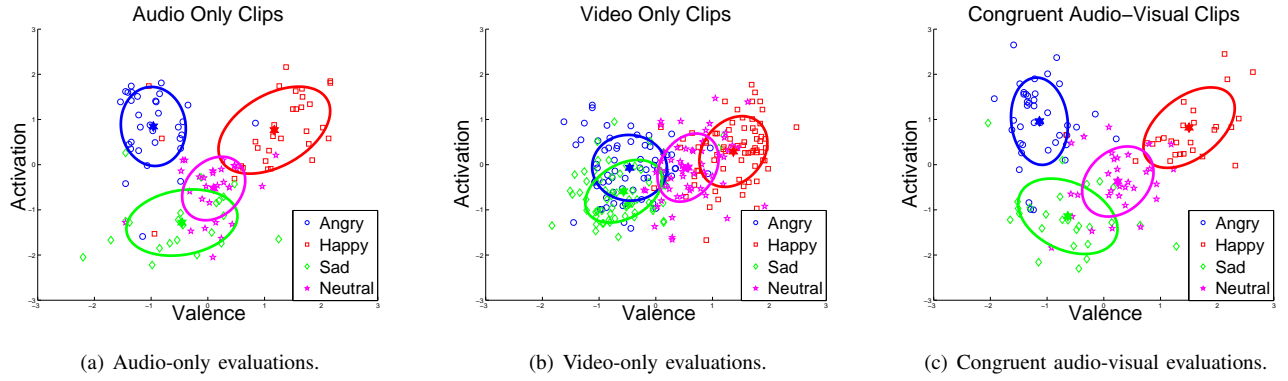(a) Audio-only evaluations.   (b) Video-only evaluations.   (c) Congruent audio-visual evaluations.

Fig. 2. The valence (x-axis) and activation (y-axis) dimensions of the evaluations; the ellipses are 50% error ellipses.

automatic emotion recognizers were used to decrease the input feature space. This reduction allowed for a tractable user-evaluation over an initially large input parameter space. The work presented in this paper represents a first step in the analysis by synthesis approach for audio-visual emotional expressions. The feature selection analysis (Section VI) suggests audio-visual features that should be included in future emotional synthesis presentation experiments. This reduced audio-visual feature subset will decrease the necessary number of synthesized utterances and will allow for a more concentrated focus on the creation of emotional models that represent the relationship between specific feature variation and human emotion perception.

III. DATA DESCRIPTION

The stimuli in this experiment include emotional audio (vocal) and video (facial) information. The emotions expressed across the two channels (the face and the voice) are either expressions of the same emotion state (congruent presentation) or are of differing emotion states (conflicting presentation). This type of experimental construct permits an analysis of the relative importance of audio-visual features across broader levels of combinations and an analysis of how the evaluators used the available features during their emotional assessments over an expanded set of emotional feature combinations.

One of the challenges of this study is creating stimuli that are free from artifacts. Purely human data present challenges in that it may be difficult for actors to express an angry vocal signal with a happy facial expression. It would be undesirable to present stimuli to evaluators containing residual facial information resulting from unintended vocal emotional expressions, and vice versa. As a result, we used an animated facial display. Despite its expressivity limitations, this interface allowed for simple and artifact free synchronization between the audio and video streams.

A. Audio-Visual Stimuli

The vocal prompts utilized in this experiment were recorded from a female professional actress [26]. The actress recorded semantically neutral utterances across each of the following emotion states: happy, angry, sad, and neutral. The sentences were then rated by four evaluators using a forced-choice evaluation framework (happy, angry, sad, neutral, and other).

Sentences that were consistently and correctly rated by all the evaluators across all four emotion classes were used in the study. The resulting set was composed of nine distinct sentences recorded across all four emotions, for a total of 36 distinct vocal utterances.

The video prompts created for this experiment were designed using the CSLU toolkit [27]. This toolkit allows a novice animator to quickly and reliably create animations of targeted facial emotions that are synchronized with an input speech signal. The toolkit has sliders (representing the strength of emotion) for happy, angry, sad, and neutral emotions (Figure 1). Each vocal utterance (of 36 total) was combined with each of the four facial emotions (happy, angry, sad, and neutral) to create a total of 144 audio-visual clips.

B. Evaluation Procedure

The stimuli were evaluated by 13 participants (ten male, three female) using a web interface (Figure 1(a)). The stimuli included audio-only, video-only, and audio-visual clips. These clips were randomly presented to the evaluators. There were a total of 139 audio-visual clips, 36 audio-only clips, and 35 video-only clips (one of the sad utterances was inadvertently, but inconsequentially, omitted due to a database error). Each participant evaluated 68 clips. The clip presentation order was randomized with respect to clip type (angry, happy, sad, neutral) and to clip content (audio, video, audio-visual). Each evaluator observed approximately 50% audio-visual clips, 25% audio clips and 25% video clips. The evaluators were allowed to stop and start the evaluation as many times as they desired.

The evaluation included both a flash video player and a rating scheme (Figure 1(a)). Each clip was rated from 0 – 100 along three dimensions, valence, activation, and dominance (VAD) using a slider bar underneath a pictorial display of the variation along the dimension. These scores were normalized using z-score normalization along all three dimensions for each evaluator. Z-score normalization was used to mitigate the effect of the various rating styles of the evaluators and thus make the evaluations more compatible.

IV. ANALYSIS OF VAD RATINGS FOR EMOTIONAL CLUSTERS

The VAD ratings of the evaluators were plotted along the dimensions of valence and activation (Figure 2) to observe

(a) "Angry" vocal emotion held constant, facial emotion varied.

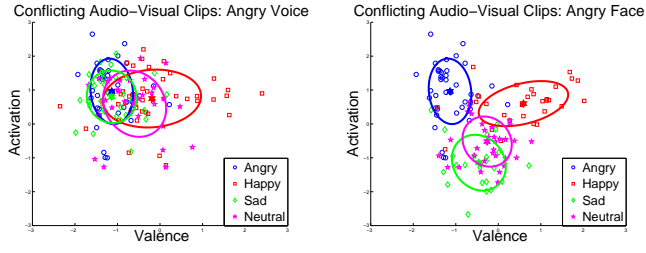(b) "Angry" facial emotion held constant, vocal emotion varied.

Fig. 3. Comparison between the emotion perceptions resulting from congruent and conflicting audio-visual presentations.

the effect of audio-visual information on emotion perception vs. that of either video-only or audio-only information. The dominance plots are not shown due to space constraints. This visualization allows for a graphical depiction of the relationship between the emotion states and their VAD ratings.

As reported in [17], the separation between the clusters was higher in the audio-only evaluation (Figure 2(a)) than in the video-only evaluation (Figure 2(b)). Discriminant analysis shows that there exists a higher level of confusion in the video-only evaluation (classification rate: 71.3%) than in the audio-only evaluation (classification rate: 79.3%) (test of proportions, $\alpha \leq 0.1$). This result suggests that the emotions presented in the audio data were more highly differentiable than in the video data, possibly due to the limited expression in the animated face used in this analysis.

A discriminant analysis of the congruent audio-visual data showed that the average classification accuracy increased non-significantly (test of proportions, $\alpha \leq 0.1$) to 80.8% (Table I). The congruent angry and happy classification rates increased when compared to the video-only and audio-only classification rates. However, the neutral and sad classification rates decreased. This suggests that the audio and video data were providing emotionally confounding cues to the participant with respect to the sad and neutral emotion classes. The confusion between these two classes in the congruent audio-visual case was in between that of the audio-only (higher level of confusion) and video-only (comparatively lower level of confusion). This indicates that the information added to the audio channel by the video channel improved the classification accuracy and supported the complementarity of the audio and video channels. However, this does not indicate that the perceptual weighting of the two channels is equivalent.

V. BIASES IN EVALUATION AND LESSONS LEARNED

The design of audio-visual emotional interfaces requires both a knowledge of how observers interpret specific facial and vocal features and how observers weight the audio and video channels during the perceptual process. This weighting process is dependent on the relevance of the emotional information contained in the channels, and on the affective bandwidth of the channels. The affective bandwidth of a channel is defined as, "... how much affective information a channel lets through [2]." The bandwidth of the channel is a function of

TABLE I
DISCRIMINANT ANALYSIS CLASSIFICATION (A=ANGRY, H=HAPPY, S=SAD, N=NEUTRAL) FOR (A) AUDIO-ONLY, (B) VIDEO-ONLY EVALUATIONS, AND (C) AUDIO-VISUAL. THE VERTICAL COLUMNS ARE THE TRUE LABELS, THE HORIZONTAL ROWS ARE THE CLASSIFIED LABELS.

(a) Confusion matrix for audio-only (ave. = 79.3%).

	A	H	S	N
A	90.6	3.1	0	6.3
H	3.2	80.6	6.5	9.7
S	0	0	72.7	27.3
N	6.5	2	19.4	71.0

(b) Confusion matrix for video-only (ave. = 71.3%).

	A	H	S	N
A	70.0	1.4	11.4	17.1
H	0	76.7	1.4	21.9
S	11.3	2.8	71.8	14.1
N	5.9	20.6	7.4	66.2

(c) Confusion matrix for congruent audio-visual (ave. = 80.8%).

	A	H	S	N
A	95.0	0	2.5	2.5
H	3.6	89.3	0	7.1
S	7.7	0	69.2	23.1
N	9.7	9.7	16.1	64.5

TABLE II
CLASSIFICATION ACCURACY IN THE PRESENCE OF CONFLICTING AUDIO-VISUAL INFORMATION, (A) ANGRY VOICE HELD CONSTANT, (B) ANGRY FACE HELD CONSTANT. THE VERTICAL COLUMNS ARE THE TRUE LABELS, THE HORIZONTAL ROWS ARE THE CLASSIFIED LABELS.

(a) Confusion matrix for angry voice held constant (ave. = 40.5%).

	A	H	S	N
A	55.0	5.0	12.5	27.5
H	14.0	48.8	16.3	20.9
S	36.4	9.1	39.4	15.2
N	28.1	46.9	12.5	12.5

(b) Confusion matrix for angry face held constant (ave. = 70.5%).

	A	H	S	N
A	87.5	2.5	7.5	2.5
H	10.3	75.9	3.4	10.3
S	8.0	0	72.0	20.0
N	11.4	8.6	34.3	45.7

the physical limitations (e.g., number of degrees of freedom) and the emotional relevance of the channel (e.g., the voice is the primary source for activation differentiation but alone cannot sufficiently convey valence [28]). An understanding of the audio-visual perceptual process would allow designers to tailor the information presented to maximize the emotional information conveyed to, and recognized by, observers.

Within this experiment, the natural audio channel dominated the perception of the users [17], [18]. This bias can be observed graphically (Figures 3(a) and 3(b)). These graphs show that when the emotion presentations are grouped by the emotion presented in the audio channel (an angry voice), the shift in perception is stronger than when the presentations are grouped by the emotion presented in the video channel (an angry face). This audio bias has also been verified using discriminant analysis (Table II). The presence of an audio bias suggests that when evaluators were presented with ambiguous or conflicting emotional information, they used the natural vocal channel to a larger degree than the synthetic facial channel to determine the emotion state.

The contribution of the audio and video information to user audio-visual emotion perception can be seen by evaluating the

TABLE III
CLUSTER SHIFT ANALYSIS WITH RESPECT TO THE VAD DIMENSIONS (WHERE ΔV_{audio} REPRESENTS THE SHIFT IN VALENCE MEAN FROM THE AUDIO-ONLY EVALUATION TO THE AUDIO-VISUAL EVALUATION). ENTRIES IN BOLD DESIGNATE EVALUATIONS OF THE AUDIO-VISUAL PRESENTATIONS THAT ARE SIGNIFICANTLY DIFFERENT, WITH $\alpha \leq 0.05$, FROM THAT OF EITHER THE VIDEO-ONLY OR AUDIO-ONLY PRESENTATIONS (PAIRED T-TEST). ENTRIES WITH A STAR (*) DESIGNATE EVALUATIONS THAT ARE SIGNIFICANTLY DIFFERENT WITH $\alpha \leq 0.001$.

Audio Emotion	Video Emotion	ΔV_{audio}	ΔA_{audio}	ΔD_{audio}	ΔV_{video}	ΔA_{video}	ΔD_{video}
Angry	Happy	**0.77***	-0.10	**-0.46**	**-1.56***	**0.46***	**0.97***
	Sad	-0.17	-0.045	- 0.21	**-0.57***	**1.39***	**1.69***
	Neutral	**0.36**	-0.25	-0.20	**-1.17***	0.66	**1.29***
Happy	Angry	**-0.58**	-0.16	0.17	**1.05***	**0.67***	**-0.24**
	Sad	**-1.39***	-0.19	- 0.37	**0.35**	**1.17***	**0.31**
	Neutral	0.06	0.045	0.19	**0.67***	**0.87***	**0.47**
Sad	Angry	-0.001	0.14	**0.92**	0.005	**-1.08***	**-0.64**
	Happy	**0.72**	**0.40**	**0.43**	**-1.11***	**-1.17***	**-0.50**
	Neutral	0.46	0.19	**0.64**	**-0.56***	**-1.02***	- 0.23
Neutral	Angry	**-0.37***	0.006	0.28	**0.20**	-0.45	-0.43
	Happy	**0.68***	0.25	0.05	**-0.58***	**-0.56***	-0.03
	Sad	**-0.62***	-0.24	-0.30	0.05	-0.18	0.09

audio-visual cluster shifts of the conflicting presentations. In this analysis, the cluster center of the audio-visual presentation (e.g., angry voice – happy face) was compared to the audio-only and video-only cluster centers (e.g., angry voice and happy face) using paired t-tests implemented in MATLAB. All reported results refer to a significance of $\alpha \leq 0.05$.

As reported in [18], the audio biased the activation audio-visual emotion perception of the users. In 10 of the 12 conflicting audio-visual presentation types, the cluster means of the audio-visual presentation were significantly different than the video-only cluster mean presentations. These same presentations were significantly different from the audio-only presentation in only 1 of the 12 conflicting presentations (Table III). This result suggests that the audio information biased the evaluations of the users in the activation dimension.

The valence dimension was not as strongly biased by the audio information as the activation dimension. In the valence dimension, 10 of the 12 conflicting audio-visual presentation clusters had means significantly different than those of the video-only presentations and 8 of the 12 conflicting audio-visual presentation clusters had means significantly different from the audio-only presentations (Table III). This suggests that in the valence dimension, the evaluators integrated both the audio and video information when making emotional assessments.

VI. AUDIO-VIDEO FEATURE ANALYSIS

Channel bias analysis is an evaluation of how individuals utilize the available information in a global sense. However, this type of analysis does not immediately suggest how the evaluators integrate the information embedded within the two channels. The goal of this section is to identify the features in the audio and video channels that contribute most, in terms of explanation of variance, to the human perceptual evaluations. In this analysis, the input feature set was reduced using information gain feature selection. The resulting feature set was then validated (in terms of variance explanation) using support vector machine (SVM) classification.

In our earlier work [19], the audio-visual features (Table IV) were analyzed over the entire data set (both congruent and

conflicting evaluations). These results identified the audio-visual features with emotional salience over the combined congruent-conflicting database. However, given a prefabricated device or avatar, or the opportunity to design such a device, it is important to know which features contribute most in the separate presentations of congruent and conflicting expressions. If it is known, for example, that the eyebrow movement timing within a synthetic face contributes to the valence evaluation of a user in the presence of congruent emotional presentations, then the designer will know that he/she must consider this feature when designing emotionally congruent expressions. If instead the designer realizes that due to constraints in one or both channels, he/she will be unable to properly create a desired emotional experience, he/she needs to know which features he/she can utilize to present recognizable information in this ambiguous emotional setting. In this scenario, activation information could be conveyed reliably by utilizing a vocal prompt with a properly designed pitch range. This feature relevance disambiguation motivated the separation of the database into two smaller databases, a congruent database and a conflicting database.

A. Feature Sets

1) Audio-Visual Features: The audio feature set was composed of 20 prosodic and 26 spectral utterance-level features (averaged over the entire sentence). The prosodic features were composed of energy, pitch, and timing statistics. The spectral features were composed of the mean and standard deviation of the first 13 Mel Frequency Cepstral Coefficients (MFCC) [29]. These features are summarized in Table IV. The audio features chosen have all been used for either design guidelines or synthetic speech modulation [25], [30], which supports their design relevance. Please see [26] for more details regarding the acoustic properties of the audio data.

The video features used in this study were based on the Facial Action Coding System (FACS), developed by Ekman and Friesen. FACS was designed to catalogue the human facial muscle movements [31]. Like the audio features, the video features were chosen for their design relevance.

Design-centered user modeling requires knowledge of the features that contribute to user perception. Since the video

TABLE IV
A SUMMARY OF THE AUDIO AND VIDEO FEATURES USED IN THIS STUDY.

Stream Type	Feature Class	Measures
Audio	Pitch	mean, standard deviation, median, min, max, range, upper quartile, lower quartile, quartile range
	Intensity	mean, standard deviation, max, upper quartile, lower quartile, quartile range
	Rate	pause to speech ratio, speech duration mean and standard deviation, pause duration mean and standard deviation
	MFCC	1 − 13, mean and standard deviation
	Prior Knowledge: Binary Emotion	angry voice, happy voice, sad voice, neutral voice
	Prior Knowledge: Mean Statistics	valence, activation, dominance of each emotion class
Video	Eyebrow Movement	none, downward, upward, downward upward, upward downward, downward upward downward, upward downward upward
	Eyebrow Movement Type	none, once, twice, thrice
	Eyebrow Movement Timing	none, first third, second third, final third
	Eyebrow Angle	flat, inner raised, inner lowered, outer raised, outer lowered
	Lip Corner Position	neutral, raised, lowered
	Eye Shape	eyes wide, top soft, top sharp, bottom soft, bottom sharp
	Prior Knowledge: Binary Emotion	angry face, happy face, sad face, neutral face
	Prior Knowledge: Mean Statistics	valence, activation, dominance of each emotion class

features used in this experiment can be illustrated through physically realizable action units, any of these features that are identified as important could, given physical constraints, be included in a synthetic character's facial display. This method of facial feature analysis highlights salient, relevant facial patterns from the set of facial actions.

A simplified subset of the FACS action units were used given the limitations of the toolkit through which the video files were created. The input video stream utilized in the stimuli presentation is a simple realization (Figure 1). Therefore, the complexities in the facial movements as described by FACS are not directly applicable. The video features employed in this study are summarized in Table IV. These features include eyebrow (movements, types, and angles), eye shape, and lip corner position features. Other areas of the face were not analyzed because they were static with respect to emotion presentation for this data. Please see Table VI-A1 for a more detailed analysis of the representation of the video features within the data.

2) Prior Knowledge Features: Prior knowledge refers to the features that indicate the emotional content of the channels. In this study there were prior knowledge features included in both the audio and the video features sets. These prior knowledge audio-visual features included average value statistics for the individual audio channel and video channel (i.e., the average VAD ratings of the audio-only and video-only components of the clip) and indicator variables that encode the presence or absence of an emotion in the audio and video channels (e.g., angry video- y/n, happy audio- y/n). From a design perspective these features describe the relevance of general emotion descriptors with respect to subsequent emotion perception. An audio label of "happy" does not fully describe the properties of the clip. Such a clip may present excited behavior, or a more tranquil happiness. If such an emotion label is a feature which predicts the outcome of an evaluator, this would indicate that evaluator behavior could be predicted using general knowledge of the channel behavior, rather than a detailed knowledge of

TABLE V
THE REPRESENTATION OF THE VIDEO FEATURES WITHIN THE DATA.

Video Feature Type	Feature Values and Number of Occurrences within the database
Eyebrow Movement	Downward (22) Downward, upward (7) Downward, upward, downward (3) Upward (25) Upward, downward (46) Upward, downward, upward (2) None (34)
Eyebrow Movement Type	Once (49) Twice (51) Thrice (5) None (34)
Eyebrow Movement Timing	First third (22) Second third (2) Final third (25) First and second thirds (13) First and final thirds (12) Second and final thirds (26) First, second, and final thirds (5) None (34)
Eyebrow Angle	Inner raised, outer lowered (35) Inner lowered, outer raised (34) Flat (70)
Lip Corner Position	Raised (35) Lowered (69) Neutral (35)
Eye shape	Top sharp, bottom sharp (35) Top soft, bottom sharp (34) Top soft, bottom soft (35) Eyes wide (35)

its components.

B. Method

1) Class Definition: This study was designed to identify the audio-visual features that contribute most to the evaluators' emotional perception. The contribution of the features was assessed using Information Gain. The feature set was reduced using a minimum gain threshold. The reduced feature set was validated using Support Vector Machine (SVM) classification, a classification tool developed by Vapnik [32].

The evaluations of the presented audio-visual, audio-only, or video-only clips were rated dimensionally (valence, activation, and dominance) on a scale from 0-100. These evaluations were preprocessed using z-score normalization across all three VAD dimensions to allow for inter-evaluator comparisons. The emotional VAD space was also preprocessed using a binary discretization based on the neutral VAD centroids. This binarization was done to account for the simplicity of the video information, since the perception of this channel did not vary widely across emotion class. After discretization, each evaluator rating was composed of a 3-dimensional binary vector representing the VAD rating with respect to the neutral centroid. For example, the VAD evaluation vector for subject i and clip j would be represented as: $eval_{i,j} = [V_{i,j}A_{i,j}D_{i,j}] = [0, 1, 1]$ for a clip rated as possessing negative valence and positive ("high") activation and dominance.

In [19], the feature selection results were presented for the data set consisting of all audio-visual presentations. In this study, the presentations were categorized as either congruent or conflicting. The feature selection and validation methods were applied separately to the two categories. This type of separation identifies features that are salient when the information presented across the two channels is emotionally equivalent (congruent presentations) or if the information is emotionally ambiguous/discordant (conflicting presentations).

2) Feature selection: The goal of the feature selection analysis is to determine the audio-visual features that contribute to the explanation of variance within the audio-visual perceptual evaluations. The data were separated into two groups: evaluations of congruent data ("congruent database") and evaluations of conflicting data ("conflicting database"). The feature selection techniques were applied to the two data subsets separately and the results from the two were compared. The feature set was reduced using the Information Gain Attribute Selection algorithm, implemented in Weka, a Java-based data mining software package [33]. Information gain feature selection techniques have been used previously in salient emotional feature selection [34]. Information gain is a measure of the difference between the entropy of set X, $H(X)$ (e.g., valence) and the conditional entropy between X and attribute Y, $H(X|Y)$ (e.g., valence given the presence of a lowered eyebrow) is known (Equation 1) [35]. Features selected by this algorithm contributed a gain of at least 0.1 with respect to the target class (discretized valence, activation, and dominance). The goal of this feature analysis was to determine the audio-visual features that contributed most to the combined audio-visual perception of the users when observing the congruent and conflicting presentations.

$$Gain(X,Y) \equiv H(X) - H(X|Y) \qquad (1)$$

Features with an information gain above the specified threshold were used to create the reduced feature sets for the congruent and conflicting data sets across the three VAD dimensions. The classification performances of the reduced feature sets were compared to the classification performances of the full feature sets using SVM. SVM is a classification algorithm that transforms input data into a higher-dimensional space to find an optimal separating hyperplane. SVM has been

TABLE VI
THE AUDIO-VISUAL FEATURES SELECTED IN THE **CONGRUENT DATABASE**. FEATURES IN BOLD ARE FEATURES THAT WERE SELECTED ACROSS THE VALENCE, ACTIVATION, AND DOMINANCE DIMENSIONS OF THE *Congruent* DATABASE. FEATURES IN BOLD-ITALICS ARE FEATURES THAT WERE SELECTED IN THE $Congruent_{VAD}$ AND $Conflicting_{AD}$ DATABASES.

| Dim	Relevant Features
Val	PRIOR KNOWLEDGE: angry (face, voice), happy (face, voice), neutral (face, voice) AVERAGE CHANNEL RATINGS: audio **VA**D, video **VAD** VIDEO: eye shape (**specific**, bottom sharp, bottom soft, wide eyes), eyebrow angle (general, flat, inner lowered, outer raised), eyebrow mvmt., **eyebrow mvmt. timing**, lip position PITCH: quartile (low, high) INTENSITY: ***max***, std, quartile (low, ***high***, ***range***) RATE: speech duration std MFCC: mean (**1**, 4, 6, 7, 10, **11**, *12*), std (**4**, **9**, 10)
Act	PRIOR KNOWLEDGE: happy (face, voice), sad (face, voice) AVERAGE CHANNEL RATINGS: audio **VA**, video **VAD** VIDEO: eye shape (**specific**, bottom soft, top sharp, top soft), eyebrow angle (general, inner raised, outer lowered), **eyebrow mvmt. timing**, lip position PITCH: mean, median, max, range, std, quartile (low, high, range) INTENSITY: mean, ***max***, quartile (***high***, ***range***) RATE: pause duration mean, pause to speech ratio MFCC: mean (**1**, 2, 3, 5, 6, 8, 10, **11**, *12*, 13), std (1, 3, **4**, 5, 6, 8, **9**, 10, 12, 13)
Dom	PRIOR KNOWLEDGE: angry (face, voice), sad (face, voice) AVERAGE CHANNEL RATINGS: audio **VA**D, video **VAD** VIDEO: eye shape (**specific**, top sharp, top soft), eyebrow angle (inner lowered, inner raised, outer lowered, outer raised), eyebrow mvmt., **eyebrow mvmt. timing** PITCH: max, std, quartile range INTENSITY: mean, ***max***, std, quartile (***high***, ***range***) MFCC: mean (**1**, 3, 5, 8, **11**, *12*), std (**4**, 5, 6, 7, 8, **9**, 11, 13)
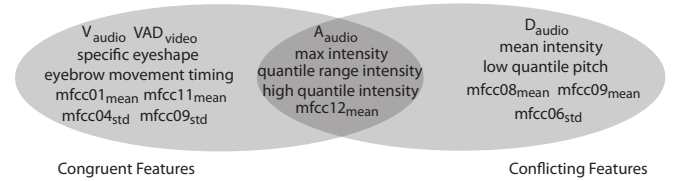
Fig. 4. Comparison between the features selected in the congruent and conflicting databases.

employed for emotion classification tasks [14], [36]–[38]. It was implemented here using Weka.

C. Feature Selection Results

Feature reduction techniques (information gain with a threshold of 0.1) were applied to congruent and conflicting audio-visual databases. Within the congruent database both audio and video features were selected across all three dimensions (Table VI). In [19], no video features were selected for the combined congruent-conflicting database in the dimensions of either activation or dominance. In the conflicting database, only audio features were selected across the three dimensions, reasserting the presence of an audio bias (Table VII).

In the congruent database, there were a total of fifteen audio-visual features selected across the three dimensions ($Congruent_{VAD}$) (Table VI and Figure 4). These features included an eye shape feature (describing the emotional shape

TABLE VIII

THE CLASSIFICATION RESULTS (SVM) OVER THE TWO DATABASE DIVISIONS (CONGRUENT, CONFLICTING) AND THREE DIMENSIONS (VALENCE, ACTIVATION, DOMINANCE) USING FEATURE SETS REDUCED WITH THE INFORMATION GAIN CRITERION DISCUSSED IN SECTION VI-B2.

Presentation	Dimension	Classification Accuracy (Prior) (%)		Classification Accuracy (No Prior) (%)		Baseline
		Full Feature Set	Reduced Feature Set	Full Feature Set	Reduced Feature Set	
Congruent	Valence	88.71	89.52	88.71	89.52	54.84
	Activation	84.68	86.29	84.68	86.29	58.87
	Dominance	78.23	78.23	77.42	78.23	58.87
Conflicting	Valence	71.23	–	71.27	–	58.03
	Activation	84.79	84.51	83.67	84.23	52.96
	Dominance	64.79	67.32	63.66	67.89	55.21

TABLE VII

THE AUDIO-VISUAL FEATURES SELECTED IN THE **CONFLICTING** DATABASE. FEATURES IN BOLD ARE FEATURES THAT WERE SELECTED ACROSS THE ACTIVATION AND DOMINANCE DIMENSIONS OF THE $Conflicting$ DATABASE. FEATURES IN BOLD-ITALICS ARE FEATURES THAT WERE SELECTED IN THE $Congruent_{VAD}$ AND $Conflicting_{AD}$ DATABASES.

Dim	Relevant Features
Val	none over the threshold of 0.1
Act	PRIOR KNOWLEDGE: angry voice, sad voice AVERAGE CHANNEL RATINGS: audio V***AD*** PITCH: mean, median, max, min, range, std, quartile (**low**, high, range) INTENSITY: **mean**, ***max***, std, quartile (low, **high**, ***range***) RATE: pause duration (mean, std), speech duration MFCC: mean (1, 3, 4, 5, 6, 7, **8**, **9**, 10, 11, ***12***), std (1, 2, 3, **6**, 8, 9, 10, 11, 13)
Dom	AVERAGE CHANNEL RATINGS: audio ***AD*** PITCH: **quartile low** INTENSITY: **mean**, ***max***, quartile (**high**, ***range***) MFCC: mean (**8**, **9**, ***12***), std (5, **6**)

of the eye), an eyebrow timing feature (describing if movement occurred within the first, second, third, or multiple thirds of the utterance), intensity features (including the utterance length maximum and quantile, representing 25%–75% of the energy, max and range), MFCC mean and standard deviation features, and prior knowledge average statistics (describing the mean valence and activation of the audio clip and the mean valence, activation, and dominance of the video clip). These features are presented in Table VI in bold font.

In the conflicting database only audio features were selected. There were a total of eleven features selected across both the activation and dominance dimensions ($Conflicting_{AD}$) (Table VII and Figure 4). There were no features selected in the valence dimension (all of the features presented an information gain under the threshold of 0.1). The features selected across the activation and dominance dimensions included a pitch feature (lower quantile, 25% of the full range), intensity features (mean, max, upper and range quantile features), MFCC mean and standard deviation features, and prior knowledge average statistics (describing the mean activation and dominance of the audio clip). These features are presented in Table VII in bold font.

There were a total of five features represented over the three dimensions of the congruent database and the activation and dominance of the conflicting database. These features included intensity features (max, upper and range quantile), an MFCC mean feature, and a prior knowledge average statistic (describing the mean activation of the audio clip). These

features are presented in Tables VI and VII in bold italic font. This feature set is an audio-only feature set. The visualization of the features selected across all three dimensions in the congruent database and two dimensions in the conflicting database can be seen in Figure 4.

D. Validation: SVM Classification

The SVM classification results indicate that there exists a subset of the full emotional audio-visual feature space that explains the variance of the user evaluations with the same level of accuracy as the full feature set across the activation and dominance dimensions in both congruent and conflicting presentations and across the valence congruent presentations using 20-fold cross-validation (Table VIII). There did not exist such a feature subspace in the conflicting database for the valence dimension (no features were selected) and as a result, the classification performance for this dimension was not included in the results table. The performance was analyzed using the difference of proportion test with $\alpha \leq 0.01$. This description holds for all statistical analyses presented in this section unless otherwise noted.

The performance of the SVM classifier was affected by presentation type. The classification accuracies for the valence and dominance dimensions were significantly lower in the conflicting presentations than in the congruent presentations for both the full and the reduced feature sets (dominance, reduced feature set: $\alpha \leq 0.05$). The classification accuracies were not affected by prior knowledge across either dimension or presentation type (congruent vs. conflicting). This suggests that the information contained within the prior knowledge features (semantic labels) is also contained within the audio and video features. In all presentations (except for conflicting valence and both full feature sets for conflicting dominance), the SVM performance beat the baseline (chance) with a significance of $\alpha \leq 0.01$.

VII. DISCUSSION

The emotional experience provided by an audio-visual presentation is distinct from that of either an audio-only or video-only experience. The emotional perception of the evaluators is affected by both the presented audio and video information. This results in a combined audio-visual perception that is often distinct from that of either of the individual channels. The emotional context of the audio-visual experience in our experiments was biased heavily by the natural audio information. However, despite this bias, the synthetic video information

contributes to the emotional evaluations of the users. Feature analysis further supports the presence of an audio bias through the selection of mostly audio features.

Audio-visual presentations offer a unique emotional experience. Angry and happy audio-visual presentations are more differentiable than those of either audio-only or video-only presentations. However, the sad and neutral audio-visual presentations were less differentiable. The results with happiness and anger are not surprising as humans generally rely on audio for the activation information and additional modalities, such as video, for the valence differentiation [28]. For example, anger is differentiated from sadness by its activation (anger and sadness have similar valence) and from happiness by its valence (anger and happiness have similar activation). The overlap between the sadness and neutrality is surprising. Sadness is generally described by low valence and low activation, both of which should distinguish such a presentation from a neutral presentation. However, in the final audio-visual presentation, the two clusters are recognized with a lower accuracy rate (using discriminant analysis) than in either the audio- or video-only presentations. In the audio-only presentation, sadness and anger were not jointly misclassified. However, in the audio-visual presentation, 7.7% of the sad presentations were classified as anger and 2.5% of the angry presentations were misclassified as sad. This sadness-neutrality overlap was also observed in [14].

As stated earlier, the video channel did not provide as much activation information as did the audio channel. It is therefore possible that the increased confusion between sadness and neutrality in the audio-visual presentation was the result of a reliance placed on a channel without the emotionally discriminative power necessary for such a differentiation. The increased confusion in the audio-visual presentations suggests that evaluators were influenced by the emotion conveyed on both channels, reinforcing the need for proper emotional design. If a user is reliant on a channel for information that it cannot provide, the final emotional perception may not be the perception that the designer anticipated.

The final user evaluations are affected by the amount of information conveyed across each channel. The evaluations within this study were biased by the natural audio information. This bias can be seen with respect to the graphical presentation of results (Figure 3), the discriminant analysis results (Table II), and the cluster shift analysis (Table III).

Feature selection techniques further verified an audio bias. In the conflicting presentations, evaluators had to decide to which information they should attend when making their assessments. The results suggest that, in the presence of conflicting information, evaluators relied on the audio channel for emotional disambiguation. The VAD cluster centers of the conflicting presentations were shifted with respect to the audio-only cluster centers (Table III) but the shift was not large enough to alter the binary class assignment of the dimensions (e.g., positive vs. negative valence). This combined with the feature selection results suggests that in the conflicting presentations, the audio information affects the gross emotional assessment behavior while the video information affects the fine emotional assessment. Within the congruent database,

the feature selection results suggest that in the presence of congruent information, evaluators utilize both the audio and video information. The presence of video features in all three dimensions in this database analysis suggests that evaluators strongly attune to video features only when the video features provide information that matches the audio channel.

The results of the SVM analysis support trends observed in audio-visual evaluation. The SVM classification results for the activation dimension do not change significantly between the congruent and conflicting presentation types (Table VIII). This is expected since humans tend to rely on audio for activation detection. Evaluators relied primarily on the audio channel when presented with conflicting information. Therefore, when asked to determine the activation, there is evidence that the users were well prepared in either presentation type because the variance in the evaluations did not increase enough to decrease the SVM classification performance.

Humans tend to utilize facial information for valence disambiguation. In the audio-visual presentations, the evaluators integrated the audio and the video information when making their valence assessment. However, as previously stated, when observing conflicting emotional presentations, evaluators tended to rely on the audio information. Therefore, when the evaluators attempted to analyze the valence dimension based primarily on the audio rather than the video signal, the variance in the evaluations increased and the performance of the SVM classification algorithm decreased.

The results of the SVM classification on the full and reduced feature sets suggest that it is possible to identify a feature subset with emotional explanatory power in the congruent presentations and in the activation and dominance dimensions of the conflicting presentations. SVM classification performance on this feature subset was not significantly different from that of the full feature set (except for the valence dimension, over which no features were selected). This result indicates that the selected feature subset could adequately explain the variance in the user evaluations.

The VAD ratings of the dominance dimension are affected by both the audio and video channels (Table III). It is therefore expected that the results of the SVM classification would lie in between that of the activation and valence dimensions with respect to performance decrease. The SVM classification results support the hypothesis that the variance in the evaluation of dominance is affected by both the audio and video channels.

Classification tasks have perviously been used for perceptual experiments. In [25], speech synthesis parameters were selected using classification techniques. The features that were selected in this process were used to synthesize and modify speech. The creation of this feature subset allowed the researchers to minimize the utterances to be rated by evaluators. In future studies, this same technique will be applied to determine which combinations of facial and vocal channel capabilities should be utilized for emotion recognition applications. This presentation framework will allow for the study of how various audio-visual combinations affect human emotional perception.

In [39], the authors presented an analysis of audio-visual emotional modulation. The data were segmented by utterances

and compared across four emotion classes (angry, happy, sad, neutral). The utterances were further segmented by phoneme boundaries. The phonemes of the emotional utterances were compared to a neutral baseline. The data suggested that phonemes that contained little emotion modulation (the feature values of the emotional utterance were not significantly different than those of the neutral utterance) were accompanied by more facial movement than those phonemes that were more strongly emotionally modulated. This recasts the emotion production problem as an emotional bit allocation problem in which the information is transmitted across the two channels based on the channel bandwidth available. In the work presented in the current paper, such emotional subtleties were not included due to the limited nature of the video channel and the nonlinearities inherent in a fusion between audio and video channels. Future experiments will utilize audio-visual data recorded using motion capture and microphone recording devices [40]. This will allow for a closer study of the channel modulation, fusion, and perceptual integration resulting from the natural expression of audio-visual emotional utterances.

VIII. CONCLUSION

This work demonstrated an analysis of an animated character emotional presentation. It showed that emotional experiences are not consistent across presentations (audio, video, and combined) and have varied representation across the dimensions of valence, activation, and dominance. Consequently, the designer must be able to anticipate channel bias, that is, the channel upon which the user will rely when making an emotional assessment, to accurately design recognizable and consistently interpretable emotion displays. Finally, this work illustrated the video and audio features that are utilized during emotional evaluations of both congruent and conflicting presentations.

This work was limited by the expressivity constraints on the video channel. Given the expressivity inequalities and the single instantiation of the emotional interface, it is difficult to generalize these results broadly without follow-up investigations. Future studies will include a four-by-four factorial design including both synthetic and human faces and voices. These four combinations will begin to illustrate the interaction between audio and video information across varying levels of emotional expressivity.

Future studies will also include human data manipulated using dynamic time warping. This method can be used to align the phoneme duration of a target sentence (e.g., angry) given the phoneme duration of another sentence (e.g., neutral). This manipulation will permit a combination of audio and video data streams with the same lexical content, but different emotional realizations, allowing for a comparison across purely human audio and video features.

This paper presented both an analysis of channel interaction and a quantitative analysis of the features important to emotional audio-visual perception. A reduced feature set was created and validated using SVM classification. However, more analysis is required to determine what impact these features have on emotion perception, rather than on emotion classification. Future work will explore this avenue using analysis by synthesis techniques to compare the emotional salience of features as a function of their identified relevance.

REFERENCES

[1] J. Gratch and S. Marsella, "A domain-independent framework for modeling emotion," *Cognitive Systems Research*, vol. 5, no. 4, pp. 269–306, 2004.

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

[3] W. Swartout, J. Gratch, R. Hill, E. Hovy, S. Marsella, J. Rickel, and D. Traum, "Toward virtual humans," *AI Magazine*, vol. 27, no. 2, pp. 96–108, 2006.

[4] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1–2, pp. 119–155, 2003.

[5] I. Dreamsite Productions. The official Mr. Bill homepage. [Online]. Available: http://www.mrbill.com/

[6] B. DeGelder and P. Bertelson, "Multisensory integration, perception, and ecological validity," *Trends in Cognitive Sciences*, vol. 7, no. 10, pp. 460–467, October 2003.

[7] B. de Gelder, "The perception of emotions by ear and by eye," *Cognition & Emotion*, vol. 14, no. 3, pp. 289–311, 2000.

[8] D. Massaro, "Fuzzy logical model of bimodal emotion perception: Comment on" The perception of emotions by ear and by eye" by de Gelder and Vroomen," *Cognition & Emotion*, vol. 14, no. 3, pp. 313–320, 2000.

[9] B. de Gelder, K. Böcker, J. Tuomainen, M. Hensen, and J. Vroomen, "The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses," *Neuroscience Letters*, vol. 260, no. 2, pp. 133–136, 1999.

[10] J. Hietanen, J. Leppänen, M. Illi, and V. Surakka, "Evidence for the integration of audiovisual emotional information at the perceptual level of processing," *European Journal of Cognitive Psychology*, vol. 16, no. 6, pp. 769–790, 2004.

[11] S. Fagel, "Emotional McGurk Effect," in *Proceedings of the International Conference on Speech Prosody*, vol. 1, Dresden, 2006.

[12] 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*, pp. 32–80, January 2001.

[13] C. M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–302, 2005.

[14] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th International Conference on Multimodal Interfaces*, State Park, PA, October 2004, pp. 205–211.

[15] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *IEEE International Conference on Multimedia & Expo (ICME)*, Los Alamitos, CA, USA, 2005, pp. 474–477.

[16] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, 2007.

[17] E. Mower, S. Lee, M. J. Matarić, and S. Narayanan, "Human perception of synthetic character emotions in the presence of conflicting and congruent vocal and facial expressions," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, April 2008, pp. 2201–2204.

[18] ——, "Joint–processing of audio-visual signals in human perception of conflicting synthetic character emotions," in *IEEE International Conference on Multimedia & Expo (ICME)*, Hannover, Germany, 2008, pp. 961–964.

[19] E. Mower, M. J. Matarić, and S. Narayanan, "Selection of emotionally salient audio-visual features for modeling human evaluations of synthetic character emotion displays," in *International Symposium on Multimedia (ISM)*, Berkeley, California, December 2008.

[20] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[21] D. Mobbs, N. Weiskopf, H. C. Lau, E. Featherstone, R. J. Dolan, and C. D. Frith, "The Kuleshov Effect: the influence of contextual framing on emotional attributionseffect: the influence of contextual framing on emotional attributions," *Social Cognitive and Affective Neuroscience*, vol. 1, no. 2, pp. 95–106, 2006.

[22] H. K. M. Meeren, C. C. R. J. van Heijnsbergen, and B. de Gelder, "Rapid perceptual integration of facial expression and emotional body language," *Proceedings of the National Academy of Sciences*, vol. 102, no. 45, pp. 16 518–16 523, 2005.

[23] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49 – 59, 1994.

[24] F. Gosselin and P. G. Schyns, "Bubbles: a technique to reveal the use of information in recognition tasks," *Vision Research*, vol. 41, no. 17, pp. 2261 – 2271, 2001.

[25] M. Bulut, S. Lee, and S. Narayanan, "Recognition for synthesis: automatic parameter selection for resynthesis of emotional speech from neutral speech." in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, April 2008, pp. 4629 – 4632.

[26] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "An acoustic study of emotions expressed in speech." in *International Conference on Spoken Language Processing International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, South Korea, 2004, pp. 2193–2196.

[27] S. Sutton, R. Cole, J. de Villiers, J. Schalkwyk, P. Vermeulen, M. Macon, Y. Yan, E. Kaiser, B. Rundle, K. Shobaki *et al.*, "Universal speech tools: the cslu toolkit," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, November – December 1998, pp. 3221–3224.

[28] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 90–100, March 2006.

[29] M. Grimm and K. Kroschel, "Rule-based emotion classification using acoustic features," in *Conf. on Telemedicine and Multimedia Communication*, Kajetany, Poland, October 2005, p. 56.

[30] M. Nicolao, C. Drioli, and P. Cosi, "Voice GMM modelling for FESTI-VAL/MBROLA emotive TTS synthesis," in *International Conference on Spoken Language Processing*, Pittsburgh, PA, USA, September 17–21 2006, pp. 1794–1797.

[31] P. Ekman, W. Friesen, and J. Hager, "Facial action coding system (FACS): Manual and investigator's guide," *A Human Face, Salt Lake City, UT*, 2002.

[32] V. Vapnik, *Statistical Learning Theory*. Wiley, New York, 1998.

[33] I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham, "Weka: Practical machine learning tools and techniques with Java implementations," in *Proceedings of ANNES International Workshop on emerging Engineering and Connectionnist-based Information Systems*, vol. 99, Dunedin, New Zealand, 1999, pp. 192–196.

[34] P. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, no. 1–2, pp. 157–183, 2003.

[35] T. M. Mitchell, *Machine Learning*. McGraw Hill, 1997.

[36] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 568–573, 2005.

[37] Y. Lin and G. Wei, "Speech emotion recognition based on HMM and SVM," *Proc. of Int. Conf. on Machine Learning and Cybernetics*, vol. 8, pp. 4898–4901, August 2005.

[38] P. Rani, C. Liu, and N. Sarkar, "An empirical study of machine learning techniques for affect recognition in human–robot interaction," *Pattern Analysis & Applications*, vol. 9, no. 1, pp. 58–69, May 2006.

[39] C. Busso and S. Narayanan, "Joint analysis of the emotional fingerprint in the face and speech: A single subject study," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Chania, Greece, October 2007, pp. 43–47.

[40] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, November 5 2008.

PLACE PHOTO HERE

Emily Mower received her B.S. in electrical engineering from Tufts University, Boston, MA in 2000 and her M.S. in electrical engineering from the University of Southern California (USC), Los Angeles, CA in 2007 . She is currently pursuing her Ph.D. degree in electrical engineering as a member of the Signal Analysis and Interpretation Laboratory (SAIL). She has been awarded the National Science Foundation Graduate Research Fellowship (2004-2007), the Herbert Kunzel Engineering Fellowship from USC (2007-2008), and the Intel Research Fellowship (2008-2009). Her research interests include developing quantitative models describing human emotion perception, which will inform the design of emotional multi-modal synthetic characters. She has also studied emotion recognition using physiological data.

PLACE PHOTO HERE

Maja J Matarić

PLACE PHOTO HERE

Shrikanth (Shri) Narayanan is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), and holds appointments as Professor of Electrical Engineering and jointly as Professor in Computer Science, Linguistics and Psychology. He received his Ph.D. in Electrical Engineering from UCLA in 1995. Prior to USC he was with AT&T Bell Labs and AT&T Research, first as a Senior Member, and later as a Principal member, of its Technical Staff from 1995-2000. At USC he is a member of the Signal and Image Processing Institute and directs the Signal Analysis and Interpretation Laboratory.

Shri Narayanan is an Editor for the Computer Speech and Language Journal (2007-present) and an Associate Editor for the IEEE Transactions on Multimedia. He was also an Associate Editor of the IEEE Transactions of Speech and Audio Processing (2000-04) and the IEEE Signal Processing Magazine (2005-2008). He served on the Speech Processing technical committee (2005-2008) and Multimedia Signal Processing technical committees (2004-2008) of the IEEE Signal Processing Society and presently serves on the Speech Communication committee of the Acoustical Society of America and the Advisory Council of the International Speech Communication Association.

Shri Narayanan is a Fellow of the Acoustical Society of America, a Fellow of IEEE, and a member of Tau-Beta-Pi, Phi Kappa Phi and Eta-Kappa-Nu. He is a recipient of an NSF CAREER award, USC Engineering Junior Research Award, USC Electrical Engineering Northrop Grumman Research Award, a Provost fellowship from the USC Center for Interdisciplinary research, a Mellon Award for Excellence in Mentoring, an IBM Faculty award, an Okawa Research award, and a 2005 Best Paper award from the IEEE Signal Processing society (with Alex Potamianos). Papers with his students have won best paper awards at ICSLP'02, ICASSP'05, MMSP06, and MMSP'07. He has published over 300 papers and has seven granted U.S. patents.