HUMAN PERCEPTION OF SYNTHETIC CHARACTER EMOTIONS IN THE PRESENCE OF CONFLICTING AND CONGRUENT VOCAL AND FACIAL EXPRESSIONS

Emily Mower[† ‡], *Sungbok Lee*[†], *Maja J Matarić*[‡], *Shrikanth Narayanan*[†]

University of Southern California
[†] Speech Analysis and Interpretation Laboratory, [‡]Interaction Laboratory
University Park, Los Angeles, California, USA 90089

ABSTRACT

Audio-visual emotion expression by synthetic agents is widely employed in research, industrial, and commercial applications. However, the mechanism through which people judge the multimodal emotional display of these agents is not yet well understood. This study is an attempt to provide a better understanding of the interaction between video and audio channels through the use of a continuous dimensional evaluation framework of valence, activation, and dominance. The results indicate that the congruent audio-visual presentation contains information allowing users to differentiate between happy and angry emotional expressions to a greater degree than either of the two channels individually. Interestingly, however, sad and neutral emotions which exhibit a lesser degree of activation show more confusion when presented using both channels. Furthermore, when faced with a conflicting emotional presentation, users predominantly attended to the vocal channel. It is speculated that this is most likely due to the limited level of facial emotion expression inherent in the current animated face. The results also indicate that there is no clear integration of audio and visual channels in emotion perception as in speech perception indicated by the McGurk effect. The final judgments were biased toward the modality with stronger expression power.

Index Terms— audio-visual emotion perception, facial emotion expression, McGurk effect

1. INTRODUCTION

The study of the perception of synthetic character emotion is important for two reasons. Firstly, a greater understanding of how humans perceive these displays may lead to more streamlined designs for avatar and robotic emotional displays. Secondly, it will further allow researchers to address the more fundamental question of how humans integrate information to arrive at final emotional assignments.

The goal of this study was to determine how participants made emotional decisions when presented with both conflicting and congruent information in an animated display composed of two channels: the facial expression and the vocal expression. Congruent information referred to an expression of the same emotion class across both the video and audio channel (e.g., angry face, angry voice). Conflicting information referred to different emotions expressed on the two channels (e.g. angry face, happy voice).

This study was motivated by the well-known work of McGurk and MacDonald [1] in which they found that in the presence of conflicting syllabic audio-visual information, the combined perception may result in a syllable perception different from that presented in either of the individual channels. In [2] this effect was interpreted using a Bayesian Discriminant function.

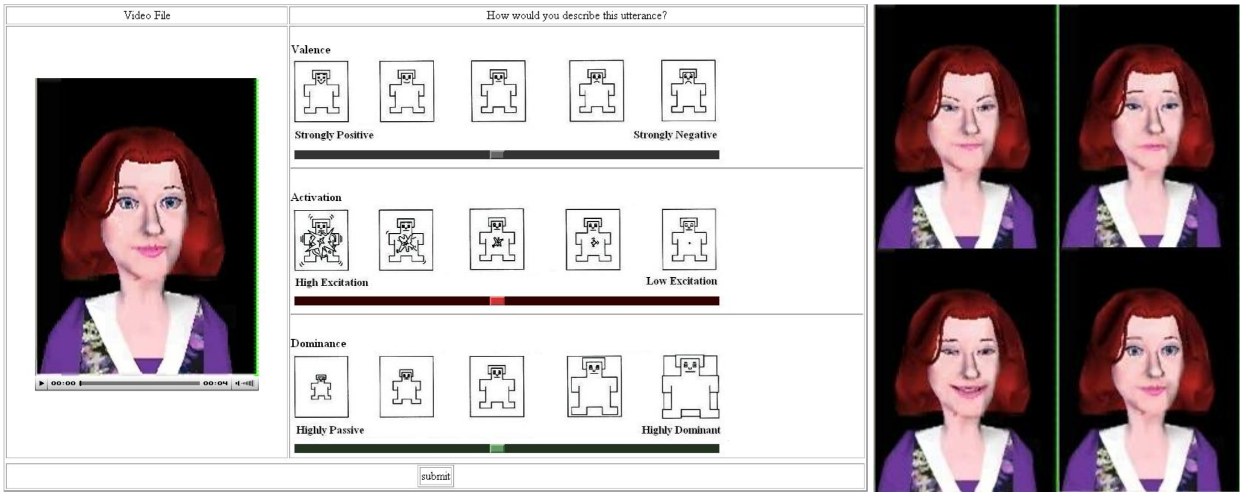
The question of perception integration in the presence of conflicting emotional cues has been studied most commonly through the use of still photographs [3, 4, 5, 6] and accompanying emotional vocalizations presented concurrently to the participant. The participant was then asked to identify the emotion to which either the combined or single channel (voice only or face only) utterance belonged using a discrete emotion category (e.g. happy vs. sad) also referred to as a forced-choice analysis. The results showed that the facial emotion expression more strongly biased the emotional perception of the user than the vocal emotion expression [3]. In another study, using film [7], researchers combined emotional music with neutral video content. They modeled the user's emotional perceptions using linear regression and found that the music accompanying the film clip had a stronger effect on the perception of the users than the visual content of the film clip.

The study presented here used self-report manikins [8] to analyze the emotional content of the clips. These manikins present a pictorial description of the dimensional axes of emotion primitives. They include categories of valence (positive vs. negative), activation (calm vs. excited), and dominance (passive vs. aggressive) heretofore referred to as VAD. These dimensions allowed for a continuous analysis of emotion integration rather than a discrete categorical analysis. To our knowledge, the current study is the first attempt to use the dimensional approach to analyze the combined perception of conflicting audio and video in a continuous framework. This continuous environment will allow for a more accurate understanding of how audio and video data interact in the human emotion perception process. This study has been designed to test how users interpret audio-visual synthetic emotional displays based on two hypotheses (Section 2) using self-reporting manikins. Section 3 will discuss the evaluation method. Section 4 will discuss the results. Section 5 will present the discussion.

2. HYPOTHESES

Hypothesis 1: The congruent combination of audio and visual information will result in a perception that is distinct from that of either the audio-only or video-only presentations.

Past experiments have shown that individuals incorporate both the audio and visual information of a clip to arrive at a final emotional decision [3, 4, 5, 6]. There exists a supplementary interaction between the speech and voice information such that the combination of the channels results in more information than either of the individual channels [9]. In the current perception experiment the clusters are expected to be more distinct in the VAD domain in the presence of congruent (consistent information provided by both channels) compared to clusters resulting from either audio or video only clips. Thus, the congruent audio-visual clusters should be more



(a) The web interface.

(b) The four emotional faces (clockwise from top left: angry, sad, neutral, happy).

Fig. 1. Emotion evaluation interface.

separated and less variant than those of the audio-only or video-only.

Hypothesis 2: In the conflicting presentation condition, the evaluator's judgment will be biased by the modality with the stronger perceptual effect.

This will result in more cluster overlap in cluster centers when the audio is held constant and varying facial emotions are presented vs. when the video is held constant and varying vocal emotions are presented. A similar effect was seen in [3] wherein the facial information was the dominant modality.

3. METHODS

3.1. Audio Data

The vocal utterances were recorded from a female professional actress [10]. The 280 sentences were recorded across four emotion classes (happy, angry, sad, and neutral) and were phonemically balanced and semantically neutral. The sentences were rated by four human evaluators using forced-choice evaluation. Sentences that were correctly rated by all four evaluators were chosen as potential sentences for the current study. The corpus was then semantically and emotionally balanced such that the same number of each utterance was used in each emotion class.

3.2. Video Data

A synthetic character was chosen to allow for consistent facial emotional expression even when the emotion expressed across the two channels (face and voice) did not match. This type of facial and vocal manipulation would be very difficult for a human actor and may have produced confounding effects.

The animated face was controlled using the CSLU toolkit [11]. This toolkit allows for a rapid creation of emotional synthetic avatars. The interface allows for the synchronization of an audio file with a target emotional facial expression (in this study: happy, angry, sad, and neutral). The emotions are specified using a slider bar ranging from 0.00 to 1.00 for each of the available emotions.

3.3. Evaluation Procedure

A total of 13 participants (10 male, 3 female) performed the evaluation using a web interface (Figure 1(a)). The participants were randomly presented with a clip from one of the following three categories: video and audio, audio only, and video only. There were a total of 139 videos, 36 audio files, and 35 video files. Each participant observed 68 distinct clips presented in random order. The order was randomized with respect to clip content and clip type (video, audio, audio-visual) such that the participants viewed approximately 50% audio-visual clips, 25% video clips and 25% audio clips. The participants were able to stop and start the evaluation as many times as they desired.

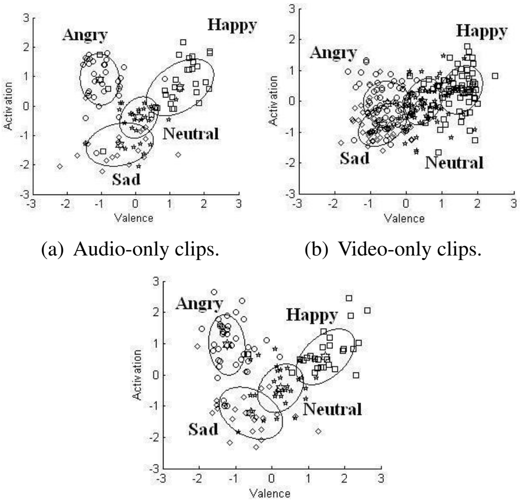
The evaluation displayed a flash video player and a rating scheme (Figure 1(a)). The clip was rated from 0 – 100 along the dimensions of valence, activation, and dominance (VAD). Each of the three dimensions were input using a slider bar underneath a pictorial display of the variation along the dimension. The rate scores were normalized using z-score normalization along all three dimensions.

4. LISTENING TEST RESULTS

4.1. General Observations

As shown in Figure 2 the separation between the clusters was higher in the audio-only than in the video-only evaluation (Figure 2). Discriminant analysis shows that there existed a higher level of confusion in the video-only evaluation (classification rate: 71.3%) than in the audio-only evaluation (classification rate: 79.3%). This result suggests that the emotions presented in the audio data were more highly differentiable than in the video data, probably 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 to 80.8% (Table 1). 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.

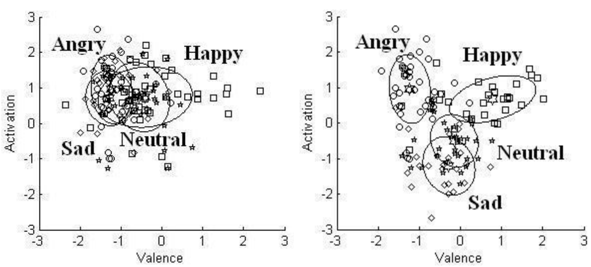


(a) Audio-only clips.

(b) Video-only clips.

(c) Congruent audio-visual clips.

Fig. 2. A depiction of the valance and activation ratings of the (a) audio-only, (b) video-only, and (c) audio-visual VAD ratings.



(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 conflicting audio-visual presentation.

The audio signal exhibited a stronger influence than the video signal on the emotional perception of the participants. This trend is especially clear in the conflicting audio-visual presentation scenario (Figures 3(a) and 3(b)). There were a total of 16 audio-visual combinations across the four emotion categories. These combinations were further grouped by (a) the emotion expressed in the audio channel and (b) the emotion expressed in the video channel. When the emotion expressed in the audio was held constant and the facial emotion varied across the four emotion categories, the cluster centers (describing the facial emotion) all shifted towards the cluster center of the vocal emotion (compare figures 2(c) and 2(a)). This trend was seen across all vocal emotion categories. A similar trend was observed when the facial emotion was held constant and the vocal emotion varied across the four emotion categories (compare figures 2(b) and 3(b)). However, this shift was much smaller than the shift of grouping (a).

Discriminant analysis of the two groupings produced a similar result (Table 3). The confusion between clusters was much higher in group (a) where the audio emotion was held constant than in group (b). This was seen across all four emotion categories and suggests that in the presence of conflicting emotional information, participants used the vocal channel to a larger degree than the facial channel to disambiguate the emotion state. This result is in conflict with previous studies which suggested that the facial information provides a stronger emotional effect than the vocal information. However, note that in these studies natural human faces were used in contrast with the animated face used in the current study which may explain the limited facial effect observed in the current study.

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

Aud	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%).

Vid	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%).

AV	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 1. Discriminant analysis classification (A=angry, H=happy, S=sad, N=neutral) for (a) audio-only, (b) video-only evaluations, and (c) audio-visual.

Both the audio-only and video-only evaluations resulted in four distinct clusters in the VAD space (Table 2).

4.2. Test of Hypotheses

The VAD ratings were significantly different in the three presentation conditions (audio only, video only, and audio-visual). The cluster means of the congruent audio-visual clusters were compared to those of the audio-only and video-only clusters. This hypothesis will be supported if the clusters in the audio-visual domain are significantly different from those in the audio-only or video-only domain.

The emotional evaluations were analyzed in groups containing either the audio-visual, the audio-only, or the video-only presentation. A one-way ANOVA analysis, performed using SPSS, of the cluster means (the independent variables were z-normalized valence, activation, and dominance, the dependent variables were either emotion class or presentation type as appropriate for all ANOVA analyses presented here) showed that the group means were significantly different for angry across all three dimensions ($F(2,130) > 25.273, p < 0.001$), for happy across the activation dimension ($F(2,129) = 7.84, p = 0.001$), for sad across the activation and dominance dimensions ($F(2,116) > 5.769, p \leq 0.004$), and for neutral across the valence and activation dimensions ($F(2,127) > 6.453, p \leq 0.002$). This indicated that the clusters were distinct in the three presentation conditions, supporting hypothesis one.

Once it was established that the audio-visual clips provided a unique emotion experience, it was necessary to determine if distinct emotional clusters existed in this new space. In the audio-visual domain, an ANOVA post hoc analysis showed that the four clusters are distinct in at least two dimensions at the $\alpha = 0.01$ level of significance (Table 2).

The second hypothesis asserted that participants would tune more to the audio channel than the video channel due to the enhanced level of emotional expression within the audio channel. In this analysis there were two separate foci: the evaluations in which the facial emotion was constant, and the evaluations in which the vocal emotion was held constant. When the vocal information was held constant and the facial information varied between the four

(a) Audio-only

	A	H	S	N
A	–	VD	AD	VAD
H	VD	–	VAD	VA
S	AD	VAD	–	AD
N	VAD	VA	AD	–

(b) Video-only

	A	H	S	N
A	–	VAD	AD	VD
H	VAD	–	VAD	VA
S	AD	VAD	–	VAD
N	VD	VA	VAD	–

(c) Congruent Audio-Visual

	A	H	S	N
A	–	VD	AD	VAD
H	VD	-	VAD	VA
S	AD	VAD	–	VAD
N	VAD	VA	VAD	–

Table 2. ANOVA post hoc analysis (A=angry, H=happy, S=sad, N=neutral) of the (a) audio-only, (b) video-only, and (c) congruent cluster means: VAD indicates that the cluster means are significantly different ($\alpha = 0.01$) in the valence, activation, and dominance dimensions.

(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

Table 3. Classification accuracy in the presence of conflicting audio-visual information, (a) angry voice held constant, (b) angry face held constant

emotion states, the clusters were not as distinct as in the congruent audio-visual case (Table 4.1). The means of the clusters across the four emotion states (angry face - angry voice to angry face - neutral voice) were distinct only across the valence and dominance dimensions ($F(3, 144) > 5.152, p \leq 0.002$). When the face was held constant and the voice was varied across the four emotion classes, the cluster means were significantly different across all three dimensions ($F(3, 125) > 34.239, p < 0.001$). This indicates that the vocal information of the clip provided a stronger influence on the emotion perception of the participant than did the facial information, supporting hypothesis two.

To further determine the effects of conflicting audio-visual data, the data were classified using Discriminant analysis. This analysis showed that the voice had a stronger effect on the classification accuracy than did the face. There were two classifications in this analysis. The first analysis studied the classification accuracy when the emotion expressed on the face was held constant (angry) and the emotion expressed on the voice varied across the four emotion states. The second analysis studied the classification accuracy when the emotion expressed in the voice was held constant (angry) and the emotion in the face varied across the four emotion states. The average classification accuracy for the first case was 70.5% and the average classification accuracy for the first case was 40.5% (See figure 3). This further indicated the strength of the vocal information in the emotional evaluation procedure, again supporting hypothesis two.

5. DISCUSSION

This study provides evidence supporting the joint processing of audio and visual cues in emotion perception. This was most stridently recognized when comparing the cluster results from the audio-only and video-only data to the congruent audio-visual clusters.

This study also provided evidence indicating that the combination of audio and visual cues does not always result in a combined emotional rating between the ratings of the individual channels. It would seem that the integration of these cues results in a different experience than observing the cues individually. This has been shown previously in [3, 6] regarding facial, but not vocal prominance.

One of the limitations of this study was the limited level of expression inherent in the animated face. Users tuned to the audio more predominantly than the video when making their emotional assessments. We believe that this is due in part to the highly expressive vocal information. Since the two channels did not have a similar level of expression this may have led to the perceived importance of the audio signal. In previous studies [3] it was found that the facial information strongly influenced the perception of the audio information when photographs of human faces were used.

The next step of the evaluation will be to utilize a more expressive animated face to analyze the interplay between the facial and vocal channel with an enhanced level of facial expression. The use of continuous domain analysis provides a novel tool for understanding the relationship between the level of expression and the relative strength of the emotional bias. Our further work will also analyze a synthetic voice combined with the current animation to determine if a combination of two channels with similar levels of expression will result in facial information having a more prominent role in the evaluation of the emotional display.

6. REFERENCES

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

[2] D.W. Massaro and D.G. Stork, "Speech recognition and sensory integration: a 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech," *American Scientist*, vol. 86, no. 3, pp. 236–242, May – June 1998.

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

[4] D.W. 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.

[5] B. de Gelder, K.B.E. 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.

[6] J.K. Hietanen, J.M. 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.

[7] Rob Parke, Elaine Chew, and Chris Kyriakakis, "Multiple regression modeling of the emotional content of film and music," *Audio Engineering Society*, 2007.

[8] M. Grimm and K. Kroschel, "Evaluation of Natural Emotions Using Self Assessment Manikins," *Proc. IEEE WSh. ASRU*, pp. 381–385, 2005.

[9] D.W. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, Mit Press, 1998.

[10] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," *Proc. Eurospeech, Lisbon, Portugal*, pp. 497–500, 2005.

[11] 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," *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 3221–3224, 1998.