JOINT–PROCESSING OF AUDIO-VISUAL SIGNALS IN HUMAN PERCEPTION OF CONFLICTING SYNTHETIC CHARACTER EMOTIONS

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

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

ABSTRACT

Expressive audio-visual synthetic characters are increasingly employed in research and commercial applications. However, the mechanism that people employ to interpret conflicting or uncertain multimodal emotional displays of these agents is not yet well understood. This study is an attempt to provide a better understanding of the interpretation of conflicting expressive displays in video and audio channels through the use of a continuous dimensional evaluation framework of emotional valence, activation, and dominance. The results indicate that when two conflicting emotions are presented to subjects using audio and video channels, the means of the dimensional evaluations of the resulting emotional judgments by the subjects is located in between the audio-only and video-only emotion perceptual centers. Furthermore, the deviation from the audio-only center is proportional to the distance between the audio and video centers. This indicates that the perceptual judgment of conflicting emotions involves the joint processing of both the audio and the video information irrespective of the perceptual bias toward the audio channel. In general the amount of interaction between audio and video channel seems proportional to the emotional disparity of the two channels in the continuous emotional space considered in this study.

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

1. INTRODUCTION

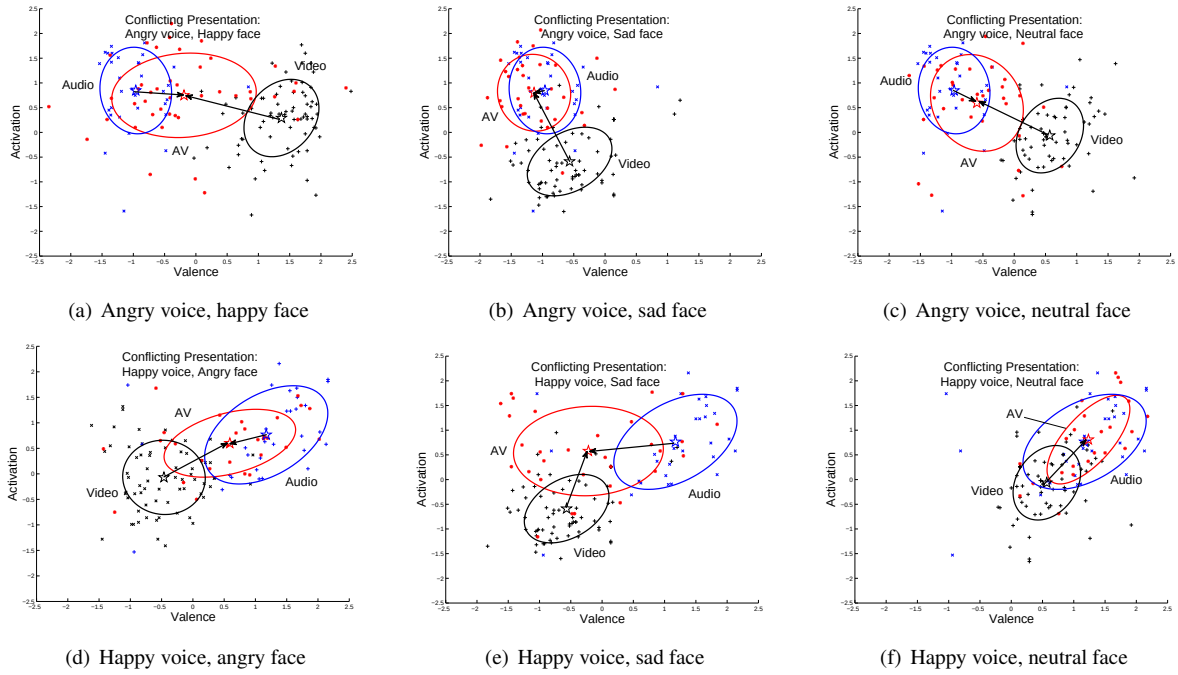
The study of the perception of synthetic character emotion in limited expression domains is important for its implications towards user comprehension and evaluation. In synthetic character emotion the audio and visual channels can be manipulated independently. In order to design effective emotional display strategies for computer and robot agents it is important to understand how the information portrayed across these two channels, in the presence of matching (congruent) or mismatching (conflicting) emotional expressions, interacts to affect the perception of the user across the spectrum of available emotions.

The goal of this study was to determine how participants interpret conflicting emotional displays of a computer simulated avatar with a human recorded voice. In this paper, conflicting displays will refer to the expression of different emotions on the two available display channels (e.g., angry face, happy voice). This study was motivated by the 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.

Perception integration has been studied with respect to conflicting emotional cues using still photographs [3, 4, 5, 6] presented with emotional vocalizations. The participants identified the emotion to either one or both of the channels (voice only or face only) belonged using a discrete emotion category (e.g. happy vs. sad). The results showed that emotional expression in the facial channel biased the emotional perception of the user with respect to the vocal channel [3]. In another study, using film [7], researchers presented neutral video content and emotional music to model the user's emotional perceptions. They found that the music accompanying the film clip had a stronger effect on the perception of the users than the visual content.

In an earlier study by the authors [8] it was shown that the audio data biased the emotional perception of the evaluators with respect to the video data. In the study, participants evaluated audio-visual emotion displays using a dimensional analysis consisting of the categories: valence (positive vs. negative), activation (calm vs. excited), and dominance (passive vs. aggressive) heretofore referred to as VAD. The audio-bias effect was studied using Discriminant Analysis and ANOVA post-hoc analyses. The previous study, however, did not address the dimensional shift resulting from conflicting presentations. This understanding is important when the categorical interpretability of the synthetic character's emotion cannot be guaranteed. Furthermore, this type of experimentation will provide a further quantitative understanding of the nature of the joint-processing of audio-visual emotional information.

This paper will address specifically the effect of conflicting presentations on the dimensional emotion perception of evaluators by investigating how users interpret audio-visual synthetic emotional displays with unequal levels of emotion expression of the audio and visual channels through the analysis of conflicting presentation evaluations. A "conflicting presentation" is a synthetic emotional expression consisting of two different emotions expressed across the facial and vocal channels. This effect will be measured with respect to the emotional evaluations of the human raters. The presented study will analyze this effect using the VAD ratings. The effect of the conflicting emotional presentation will be measured by analyzing the shift in the VAD cluster means observed in the evaluations of the audio-visual presentations. These shifts will be described with respect to the evaluations of the original audio-only and video-only presentation conditions. This analysis will demonstrate the effect that conflicting presentations have on the joint processing of audio-visual signals. In section 2 the data and analysis method are described and the results are presented in section 3. A discussion is presented in section 4.



(a) Angry voice, happy face (b) Angry voice, sad face (c) Angry voice, neutral face

(d) Happy voice, angry face (e) Happy voice, sad face (f) Happy voice, neutral face

Fig. 1. These graphs indicate the shift from the audio-only and video-only presentation evaluations to the combined audio-visual presentation evaluation for presentations with angry and happy vocal information with respect to valence (x-axis) and activation (y-axis).

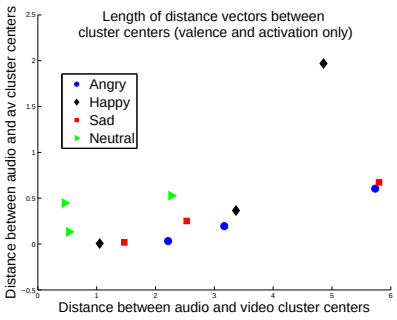


Fig. 3. A comparison between the norm squared of the valence and activation distance vector between the means of the audio-only and video-only evaluations compared to the distance between the means of the audio-only and audio-visual evaluations over all 12 conflicting presentations (Table 1).

2. DATA ANALYSIS

2.1. Data Description

The data used in this experiment consist of audio-visual, audio-only, and video-only clips. The audio files were recorded from a female professional actress [9] and contained seven distinct sentences (nine in total) repeated over four emotions (angry, happy, sad, neutral). The animated face was designed using the CSLU toolkit [10] and was synchronized with the audio data to produce four emotional facial displays (angry, happy, sad, and neutral). The audio and video information were presented to evaluators over a web interface. The 13 participants (10 males and 3 females) rated the valence, activation, and dominance of the 139 audio-visual files, 36 audio-files, and 35 video-only files using a slider scale from 0 – 100. The scores were then normalized with respect to the evaluator using z-score normalization along each dimension. The data are described more fully in [8].

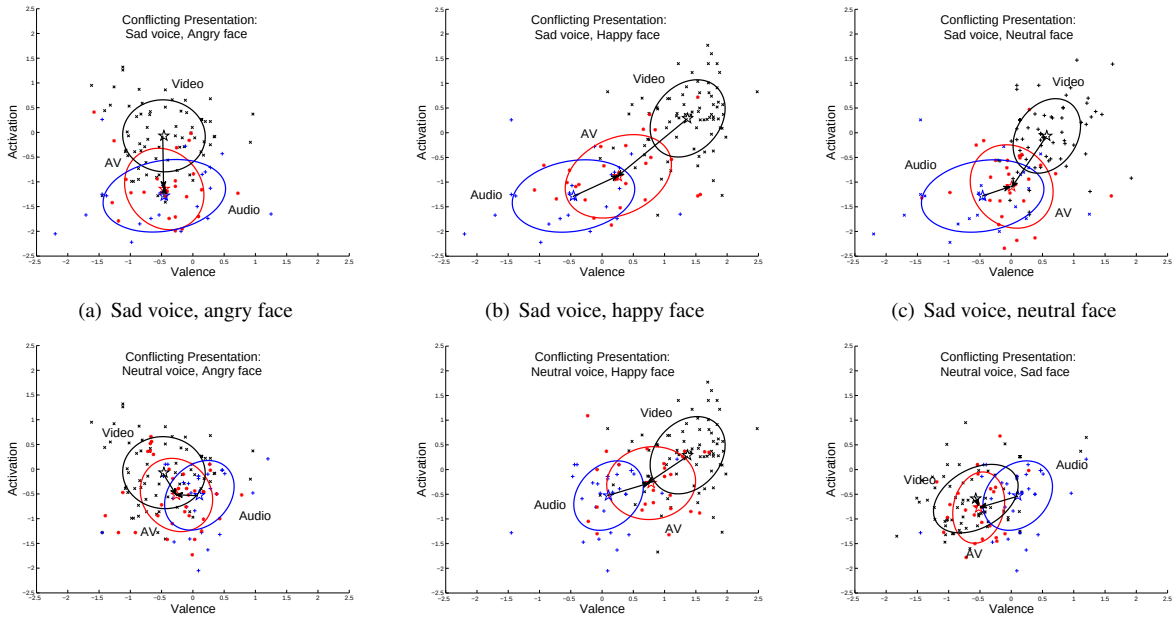
Our initial study with this data provided evidence supporting the joint processing of audio and visual cues in emotion perception [8]. The results showed that in the presence of disparate emotional expressivity, users evaluate the emotional expression of the synthetic character in accordance with the more expressive modality. We showed that the audio signal (human recorded) has a stronger effect on the perception of emotion given the implementation tested than the video signal (animation of limited expression). Discriminant analysis showed that there existed four distinct clusters in the congruent audio-visual, audio-only, and video-only presentation conditions in the VAD space. This analysis indicated that the four emotional displays were evaluated as separate emotions in the described presentation conditions and that the participants were able to distinguish between emotional displays in the congruent audio-visual presentation.

2.2. Dimensional Shift Analysis

This paper will present the shift in the VAD cluster means observed in the evaluations of the audio-visual presentations when compared to the evaluations of the original audio-only and video-only presentation conditions. A graphical analysis of the shift (Figures 1 & 2) is presented to determine the location of the evaluated audio-visual cluster center with respect to the audio-only and video-only evaluated cluster centers. The black arrows on the graphs indicate both the direction and degree of cluster center movement. The ellipses are 50% error ellipses. The following analysis will demonstrate the effect that the conflicting presentations have on the joint processing of the the audio-visual signals. The statistical tests discussed below are all paired t-tests, performed using MATLAB.

3. RESULTS

The activation differentiation between the emotional clusters was much higher in the audio-only presentation than the video-only presentation [8]. It is therefore expected that the audio signal would



(a) Sad voice, angry face (b) Sad voice, happy face (c) Sad voice, neutral face

(d) Neutral voice, angry face (e) Neutral voice, happy face (f) Neutral voice, sad face

Fig. 2. These graphs indicate the shift from the audio-only and video-only presentation evaluations to the combined audio-visual presentation evaluation for presentations with sad and neutral vocal information with respect to valence (x-axis) and activation (y-axis).

bias the evaluation of the conflicting audio-visual presentation with respect to the video signal since the video signal contained very little activation information. Table 1 indicates that in 10 of the 12 audio-visual presentations, the activation was significantly different from the video-only presentation with significance $\alpha \leq 0.05$, and in 9 of the 10 presentations with $\alpha \leq 0.001$. The activation of the audio-visual presentations were different from the audio-only activation evaluations in only one case, with significance $\alpha \leq 0.05$.

The only activation evaluation of the audio presentation that differed significantly from that of the audio-visual presentation was the conflicting sad voice – happy face presentation. In this presentation the emotions have a valence, activation, and dominance mismatch. It is therefore possible, that the perceived activation of this presentation condition was affected by the conflicting information with respect to the valence and the dominance.

The valence of the presentation conditions were differentiated in both the video-only and audio-only presentation conditions [8]. It was therefore expected that the valence evaluation in the audio-visual presentation conditions would be affected by both the audio and video channel information. Additionally, due to the observed channel bias, it was expected that the audio information would affect the evaluations more strongly. Table 1 shows that the valence evaluation of the audio-visual presentation conditions differed significantly ($\alpha \leq 0.05$) from the audio-only presentation in eight of the 12 conflicting presentation conditions and from the video-only presentation conditions in 10 of the 12 presentation conditions.

Previous work [11] has shown that facial information carries more valence differentiation than vocal information, indicating that humans tend to use facial displays to disambiguate between the valence of an emotional presentation. Therefore, although it was shown in the previous analysis of this emotional data [8] that the audio information more strongly biased the emotional evaluations, the valence bias of the audio information was mitigated by our reliance on facial information. This may explain why the valence information of the audio-visual clips was distinct from the audio-only clips in 2/3 of the presentation conditions while the activation information was different in only 1/12 of the presentation conditions.

The audio-visual presentations with significantly different valence means included emotional combinations with logical valence mismatches. For example, the angry voice – happy face presentation resulted in a significant valence mismatch when compared to both the audio-only and video-only presentations.

The dominance of the presentation conditions were differentiated much more strongly in the audio data than in the video data. Table 1 indicates that the dominance information of the audio-visual presentation evaluation differed significantly from the video-only presentation condition in nine of the 12 conflicting presentation ($\alpha \leq 0.05$). The dominance of the audio-visual presentation condition differed from that of the audio-only presentation condition in only four of the 12 conflicting presentations. The presentation conditions resulting in significant dominance cluster shifts occurred between emotion combinations with different levels of dominance (sad voice – angry face).

The strength of the audio bias was inversely proportional to the distance between the audio and video evaluation cluster means in the angry, happy, and sad emotional conditions. That is, as the norm squared of the vector composed of the Euclidean distance between the valence and activation of the audio-only and video-only evaluations increased, the observed evaluation bias of the audio channel decreased (Figure 3). In [8], it was found that the evaluations of the combined audio-visual presentation were more strongly affected by the audio channel data than by the video channel data. However, figure 3 indicates that the strength of this bias decreases as the perceptual mismatch between the two channels (audio and video) increases in the angry, happy, and sad emotional categories. The relatively small change in the distance between the audio-av distance may be due to the disparity between the available levels of expression in the audio and video channels. It is possible that in the presence of con-

Audio Emotion	Video Emotion	ΔV_{audio}	ΔA_{audio}	ΔD_{audio}	ΔV_{video}	ΔA_{video}	ΔD_{video}
Angry	Happy	**0.7714***	-0.09718	**-0.4646**	**-1.5579***	**0.45866***	**0.96733***
	Sad	-0.17391	-0.045341	- 0.20647	**-0.56823***	**1.3898***	**1.6927***
	Neutral	**0.36438**	-0.25	-0.2025	**-1.1652***	0.66096	**1.293***
Happy	Angry	**-0.58187**	-0.16454	0.17355	**1.0532***	**0.66894***	**-0.24329**
	Sad	**-1.3909***	-0.18653	- 0.37045	**0.34862**	**1.1665***	**0.31177**
	Neutral	0.061445	0.044937	0.19442	**0.66567***	**0.87384***	**0.47293**
Sad	Angry	-0.0012364	0.13656	**0.92138**	0.0045429	**-1.0755***	**-0.64009**
	Happy	**0.71672**	**0.39851**	**0.4289**	**-1.108***	**-1.1733***	**-0.50079**
	Neutral	0.46351	0.19279	**0.64068**	**-0.56153***	**-1.0239***	- 0.22544
Neutral	Angry	**-0.36616***	0.0062765	0.2776	**0.19986**	**-0.45186**	**-0.43471**
	Happy	**0.68133***	0.25342	0.052632	**-0.58316***	**-0.56441***	-0.027907
	Sad	**-0.62431***	-0.24345	-0.29488	0.046185	-0.182	0.091862

Table 1. 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 signify that the evaluation of the audio-visual presentation differs significantly, with $\alpha \leq 0.05$, from that of either the video-only or audio-only presentations (paired t-test). Entries with a star (*) indicate that the evaluations are significantly different with $\alpha \leq 0.001$.

flicting information expressed on channels with equal levels of emotional expression, the relationship between audio-video cluster mean distance and audio-av cluster mean distance would have a more linear relationship.

4. DISCUSSION

This study provides further evidence supporting the joint processing of audio and visual cues in human emotion perception. This result is most apparent when observing the cluster shift behavior of the conflicting audio-visual presentation conditions. This data indicate that when the emotion states expressed across the available communication channels are in conflict, the final observed emotion state may differ dimensionally from the emotions expressed in either of the two channels. Furthermore, this work indicates that the integration of the emotional cues results in a different experience than observing the cues individually. This has been shown previously in [3, 6] regarding facial prominence only.

One of the limitations of this study in terms of generality was the limited level of expression inherent in the animated face. We found previously that users tuned to the audio channel more predominantly than the video channel when making their emotional assessments. Since the two channels did not have a similar level of expression this may have led to the perceived importance of the audio signal. However, this unequal level of expression was an ideal platform for the investigation of intended consequences of emotional mismatch in limited expression domains. Since many robots and computer avatars are designed within this domain, this research provides emotional evaluation information that can be used in future designs to create more emotionally consistent expressions.

The next step of the evaluation will be to utilize a human voice-human face presentation 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 quantitative 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.

5. ACKNOWLEDGEMENTS

This research was supported in part by funds from the National Science Foundation.

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] Emily Mower, Sungbok Lee, Maja J Matarić, and Shrikanth Narayanan, "Human perception of synthetic character emotions in the presence of conflicting and congruent vocal and facial expressions," *ICASSP*, April 2008.

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

[10] 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.

[11] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," *Proceedings of the 6th international conference on Multimodal interfaces*, October 2004.