USING EMOTIONAL NOISE TO UNCLOUD AUDIO-VISUAL EMOTION PERCEPTUAL EVALUATION

Emily Mower Provost*, Irene Zhu*, and Shrikanth Narayanan‡

*Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA
‡Electrical Engineering, University of Southern California, Los Angeles, CA, USA
{emilykmp, ireeene}@umich.edu, shri@sipi.usc.edu

ABSTRACT

Emotion perception underlies communication and social interaction, shaping how we interpret our world. However, there are many aspects of this process that we still do not fully understand. Notably, we have not yet identified how audio and video information are integrated during the perception of emotion. In this work we present an approach to enhance our understanding of this process using the McGurk effect paradigm, a framework in which stimuli composed of mismatched audio and video cues are presented to human evaluators. Our stimuli set contain sentence-level emotional stimuli with either the same emotion on each channel ("matched") or different emotions on each channel ("mismatched", for example, an angry face with a happy voice). We obtain dimensional evaluations (valence and activation) of these emotionally consistent and noisy stimuli using crowd sourcing via Amazon Mechanical Turk. We use these data to investigate the audio-visual feature bias that underlies the evaluation process. We demonstrate that both audio and video information individually contribute to the perception of these dimensional properties. We further demonstrate that the change in perception from the emotionally matched to emotionally mismatched stimuli can be modeled using only unimodal feature variation. These results provide insight into the nature of audio-visual feature integration in emotion perception.

***Index Terms*—** Emotion perception, McGurk effect

1. INTRODUCTION

Emotion perception is central to human interactions. It modulates how we interpret our world, fundamentally shaping our behavior. However, we still do not understand the mechanism underlying the perception of emotion. Such knowledge is essential and can inform the design of human-centered multimedia content and interfaces. This paper presents an investigation into audio-visual perception using dynamic sentence-level stimuli with mismatched emotional cues, constructed using the audio-visual McGurk paradigm. The results provide insights into audio-visual feature reliance by modeling the relationship between features and perception and specifically emotion-specific audio-visual feature reliance.

Audio-visual feature reliance is challenging to estimate due to the inherent correlation that exists across modalities. This correlation renders it difficult to determine how individual modalities shape gestalt perception. The McGurk effect paradigm can be used to create new stimuli for which the cross-modal perceptual correlation is reduced. The paradigm is based on an audio-visual perceptual phenomenon discovered by McGurk and MacDonald in 1976 [1]. They found that when subjects were presented with conflicting audio-visual phoneme cues (e.g., the phoneme "ba" and the viseme "ga") the final perception was a distinct third phoneme (e.g., "da"). In the emotion domain, this paradigm has been used to investigate how audio and video interact during the emotion perception process.

In previous works, McGurk effect stimuli have been created using a combination of human audio and still images [2–5], human audio and single word video [6], and human audio and animated video [7]. Sentence-level dynamic human (both audio and video) McGurk Effect stimuli have not been studied due to stimuli creation challenges. The audio and video from separate audio-visual emotion clips must be correctly aligned at the phoneme-level to create clips that do no look "dubbed". Accurate alignment is crucial because improper synchrony creates perceptual artifacts that may disrupt the joint processing of the two modalities. The dynamic human stimuli more closely reflect real-life human displays of emotion.

This paper presents a novel study investigating the dynamic perception of human audio-visual emotional McGurk effect stimuli. The stimuli are a collection of utterances from a single actress read-speech database in which an actress read single sentences across four emotion categories: anger, happiness, neutrality, and sadness [8]. We extract the audio and video information, synchronize the video with the audio from conflicting emotions, and then merge the audio and new video to create a series of new clips with mismatched audio and video information (e.g., a happy face with an angry voice)[1]. The dimensional properties (valence and activation) of the clips were evaluated by a group of anonymous evaluators with a crowd sourcing approach using Amazon Mechanic Turk. Audio and video features were extracted from these emotionally matched and mismatched utterances and linear regression models were used to assess the relationship between perception and audio-video feature streams.

The results demonstrate that the perception of the McGurk effect stimuli is different from that of the original stimuli as shown in [7]. The results further show that the McGurk effect perception can be accurately modeled with linear regression models. These models highlight the features that contribute to reported perception and can be extended to capture perceptual change, explaining causes behind changes in reported perception between the McGurk effect and original stimuli. The framework is a novel method to investigate the cross-channel bias in emotion perception by modeling perception change as a function of unimodal feature change and contributes to a novel human emotional McGurk stimuli set.

2. DESCRIPTION OF DATA

The stimuli used in this experiment can be separated into two sets: Original Audio-Visual Emotions (**OAV**) and Resynthesized Audio-Visual Emotions (**RAV**). This section describes the collection of the OAV clips and the creation of the RAV clips. The OAV clips were recorded from a single actress reading semantically neutral sentences. The use of read speech enables the collection of identical

[1] See the following website for example clips: http://eecs.umich.edu/~emilykmp/mcgurk/

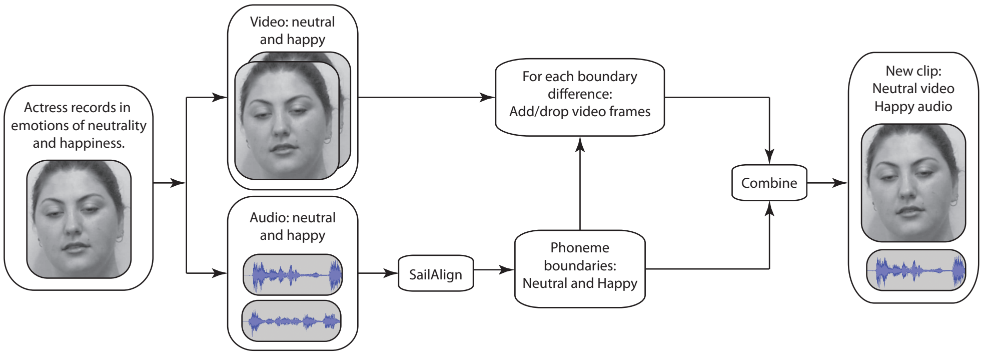

Actress records in emotions of neutrality and happiness.
Video: neutral and happy
Audio: neutral and happy
SailAlign
Phoneme boundaries: Neutral and Happy
For each boundary difference: Add/drop video frames
Combine
New clip: Neutral video Happy audio

Fig. 1. The method used to create the McGurk (RAV) stimuli. The audio and video are separated. The audio clips from two disparate emotions (e.g., happy and neutrality from the same utterances) are aligned at the phoneme level. The timing differences between the phoneme durations of the neutral and happy utterance are used to guide the addition or deletion of frames from the happy video. The audio from the neutral recording is combined from the video from the happy recording once the phoneme boundaries from the two emotions are aligned. This creates a multimodal clip with two different emotions expressed across the audio and video channels.

lexical content across emotions. This lexical continuity is crucial for seamless resynthesis of audio and video information from separate OAVs (e.g., proper lip movement). Each sentence was read four times, once for each emotion class (anger, happiness, sadness, neutrality) creating four OAV stimuli for each utterance in the set. There were a total of 18 sentences used in this study. This created a set of 72 OAV clips (18 utterances x 4 emotions/utterance). There are four OAVs associated with each $utterance_i, 1 \leq i \leq 18 : OAV_{i,k}, k \in \{angry, happy, neutral, sad\}$.

The RAV stimuli are created using the audio and video streams from the OAV stimuli. For clarity, the creation of the RAVs will be described using $utterance_i$. The audio and video are extracted from each $OAV_{i,k}$, resulting in four audio emotions and four video emotions per utterance. The audio and video emotions are permuted and recombined to create twelve mismatched audio-visual clips ($_4\mathrm{P}_2$), requiring phoneme-level temporal alignment between the utterances in $\forall k \; OAV_{i,k}$. The clips were aligned using SailAlign, developed at the University of Southern California [9]. The resulting phoneme boundaries were retouched by hand. The phonemes were then aligned using NIST's SClite scoring tool [10]. The alignment was used to provide warping instructions for the video stream. The audio channel was not warped because preliminary evidence suggested that artifact-free phoneme-level audio warping was very challenging. The frames from the video stream were extracted and up sampled by a factor of four to permit more seamless warping. We describe the warping process with respect to two example clips, anger and happiness (target RAV – "angry-audio with happy-video"). Once again, the video was warped while the audio remained unchanged, thus, the audio phoneme boundaries were the target in the video warping. For a given phoneme, if the audio duration was longer than the video duration, frames were added, otherwise frames were dropped (or if unchanged, there were no changes made). At the conclusion of this process, the video timing and duration matched that of the audio. The angry audio was combined with the aligned happy video to create the target RAV. This was repeated over the remaining 11 emotion combinations. After iterating over all utterances in the utterance set, there are a total of 216 RAV utterances (18 utterances x 12 recombinations/utterance, see Figure 1).

2.1. Evaluation

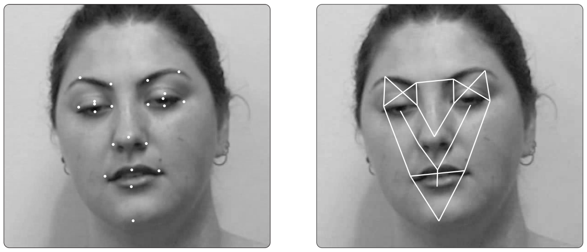
The data were evaluated through crowd sourcing via Amazon's Mechanical Turk (AMT) [11]. AMT has been effectively used in many domains [12] including: transcription tasks [13], labeling tasks [14], and in emotion labeling tasks [15]. AMT provides an interface that researchers can use to solicit a large number of rapid assessments from a population of evaluators. In this study, the evaluator population was restricted to people within the United States (N = 117).

The full dataset (OAV, RAV, audio-only, and video-only clips) was evaluated using AMT. The evaluators were asked to label the primary emotion and secondary emotion (if such an emotion was present/observable) from the set {angry, happy, neutral, sad, other}. They also assessed the dimensional properties of the clips including valence (positive vs. negative) and activation (calm vs. excited) using a five-point Likert scale and transcribed the clip to ensure that they could view the clips. The evaluators were paid $0.10 per evaluation.

The average standard deviation of the dimensional perception of the OAV clips was 0.634 ± 0.275 (valence) and 0.676 ± 0.208 (activation). They evaluations are not significantly different across dimensions (paired t-test, α = 0.05). The average standard deviation for the evaluation of the RAV clips was 0.683 ± 0.235 (valence) and 0.844 ± 0.262 (activation). This difference is significant. The standard deviation for the RAV clips is higher than that of the OAV clips. This may reflect the increased emotional complexity due to the conflicting emotional sources.

The data were smoothed using two techniques: (1) evaluator selection via the weighted Kappa statistic and (2) dropping the highest and lowest evaluation for each clip. The Kappa statistic is a measure of the agreement between two evaluators that takes chance agreement into account. We use a weighting function that is exponential with respect to the difference between two evaluations. This permits a comparison of ordinal evaluations, important when using Likert scales. The weighted Kappa statistic was calculated over every pairing of the 117 distinct evaluators who evaluated the same clip. All evaluators with an average Kappa below the 10% quantile were discarded as noise. The second smoothing technique accounts for periodic inattention or evaluation mistakes. There was an average of 5.31 ± 1.06 evaluators for each OAV clip and 5.36 ± 1.17 evaluators for each RAV clip. The standard deviation of the OAV clips after smoothing was 0.425 ± 0.268 (valence) and 0.423 ± 0.249 (activation), not statistically significantly different across dimensions (paired t-test, α = 0.05). The standard deviation of the RAV clips after smoothing was 0.455 ± 0.278 (valence) and 0.571 ± 0.290. The values are statistically significantly different, perhaps indicating a relative difficulty in assessing activation from RAV stimuli.

We then assessed whether the evaluators correctly perceived the actress's target emotion. We assigned the primary emotion label for each utterance using a majority voting over all of the reported emo-



(a) Estimated facial marker locations.

(b) Euclidean dist. between points.

Fig. 2. The facial feature configuration.

tion labels. The majority vote included only those evaluators whose Kappa statistic was above the 10% quantile threshold. In the OAV presentations, the evaluator assignment matched the goal of the actress in 69 of the 72 clips (95.83%). The evaluations based only on audio information matched the actress target in 61 of the 72 clips (84.72%). The largest confusion occurred between the target class of sadness and perceived class of neutrality and between the target class of happiness and the perceived class of anger. This perceptual auditory confusion is commonly reported [16]. The video-only evaluator assignment matched the actress target in 70 of 72 clips (97.22%).

2.2. Feature Extraction and Selection

The video features are distances between points on the face, estimated using BoRMaN (Boosted Regression with Markov Networks for Facial Point Detection) [17] (Figure 2). There are 22 points automatically extracted in each frame. The video feature extraction is time intensive and our frames required manual adjustment. We mitigated this issue by: (1) extracting features from down sampled OAV stimuli (which had been previously up sampled) and (2) extracting features only from the OAV stimuli. First, the feature points were extracted from the OAV stimuli at every fourth frame (the frames were previously up sampled by a factor of four). The values of the remaining frames were assigned using interpolation. Second, the important difference between RAV and OAV stimuli is timing, the content of the RAV stimuli is a derivative of the content of the OAV stimuli. In Section 2 we discussed a method to obtain the RAV video stream from the OAV video stream. We applied the same warping parameters to the extracted frame-level feature points to obtain the RAV features. We use Euclidean distances between points, rather than the points themselves, to increase the interpretability of the results. Preliminary evidence also supported enhanced regression estimates when using distances rather than points (Figure 2).

The raw audio features were extracted using openSMILE [18]. The audio features are from the Interspeech 2010 Paralinguistic Challenge [19] and include: root-mean-square of the energy in the signal frames, Mel-Frequency cepstral coefficients (MFCC) and logs of the power of each Mel-Frequency band (MFB, 1-12), the zero-crossing rate of the time signal, the probability of voicing, the LPC coefficients and spectral pair frequencies from the coefficients, and the fundamental frequency (F0) [18].

The final feature set contained utterance-level statistical functionals of the Euclidean distances and audio features. The statistics include: mean, standard deviation, lower quantile, upper quantile, and quantile range. The final feature set contained 185 audio features and 105 video features (290 features total).

3. METHODS

3.1. Statistical Analysis of Perceptual Evaluation

The studies in this section are motivated by two hypotheses. Hypothesis 1: Audio-visual valence/activation perception is different from audio-only or video-only perception. Hypothesis 2: RAV valence/activation perception is different from OAV valence/activation perception. The analyses assess how audio and video interact during the perception of emotionally clear (OAV) and emotionally noisy (RAV) stimuli. Hypothesis 1 is supported by prior work, which showed that multimodal perception is different from that of unimodal perception [7] and that multimodal models can more accurately capture the target emotions than unimodal models [20]. Hypothesis 2 is supported by prior work, which showed that dynamic RAV perception was different from OAV perception (e.g., valence/activation perception of RAV “angry-video, sad-audio” is different than the perception of anger and/or sadness) [7].

Hypothesis 1 and 2 are tested using ANOVA (analysis of variance) and secondary t-tests. The ANOVA analyses assess whether the means of the valence and activation perception differ across presentation condition (OAV: audio-only, video-only, audio-visual and RAV: video-only, audio-only vs. audio-visual). If the ANOVA analysis suggests that the means of reported perception are different, a t-test is performed to identify the specific presentation conditions that have different valence and activation perceptions (e.g., RAV “angry-audio, sad-video” valence perception differs from the OAV sad utterance). In all analyses $\alpha = 0.05$. This section uses techniques similar to those presented in [7]. However, the current results refer to a novel stimuli set and provide initial information used in the novel perception regression analyses.

3.2. Regression

The regression analyses provide insight into (1) the relationship between feature expression and audio-visual perception and (2) the relationship between feature change and audio-visual perception change: how perception shifts as a function of feature-level change. The advantage of working with McGurk stimuli is that RAV stimuli have unimodal feature-level similarity to OAV stimuli. For example, a RAV utterance with a neutral face and a happy voice has the same audio cues as the happy OAV utterance and has video cues that are similar to the neutral OAV utterance. This provides an ideal opportunity to study how audio-visual feature change affects perception change. There are two regression studies presented. The first predicts reported **perception** using audio and video feature values. The second predicts **perception change** using a change in presented features. The results of these studies will provide insight into how feature changes, in the presence of perpetual noise, provide evaluators with perceptually relevant information. In all analyses, the results are presented via leave-one-utterance-out cross-validation. There are four and twelve utterances for each cross-validation fold for the OAV and RAV utterances, respectively.

4. ANALYSIS OF REPORTED PERCEPTION

The combination of OAV, unimodal, and RAV stimuli offer insight into how audio and video modalities interact during perception. We plot the perception of the OAV and RAV stimuli to provide insight into how emotional mismatch shapes perception. Figure 3 presents the valence and activation perception of audio-only, video-only, and audio-visual stimuli. The points are labeled by the actress’s target emotion (e.g., angry). The figures demonstrated that in all presentation conditions there is separation between anger, happiness, and sadness/neutrality. However, the emotions of sadness and neutrality are confused in all three presentations, most notably the audio-only (see Section 2.1). Figure 4 presents perceptual results of the RAV evaluation. The figure is broken into two presentations: stimuli grouped by emotion on the audio channel (left) and stimuli grouped by the emotion on the video channel (right). For example, the angry audio subplot shows that the combination of angry audio and any video affect creates a perception that is more negative and more excited than the original OAV perception. Similarly, the angry video

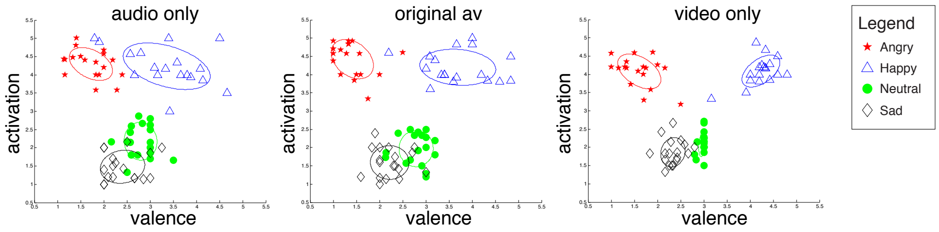


Fig. 3. The perception of the OAV clips. The red stars are angry, the blue triangles are happy, the green squares are neutral, and the black diamonds are sad. The ellipses contain 50% of the class data. Valence is low (1) to high (5) and activation is calm (1) to excited (5).

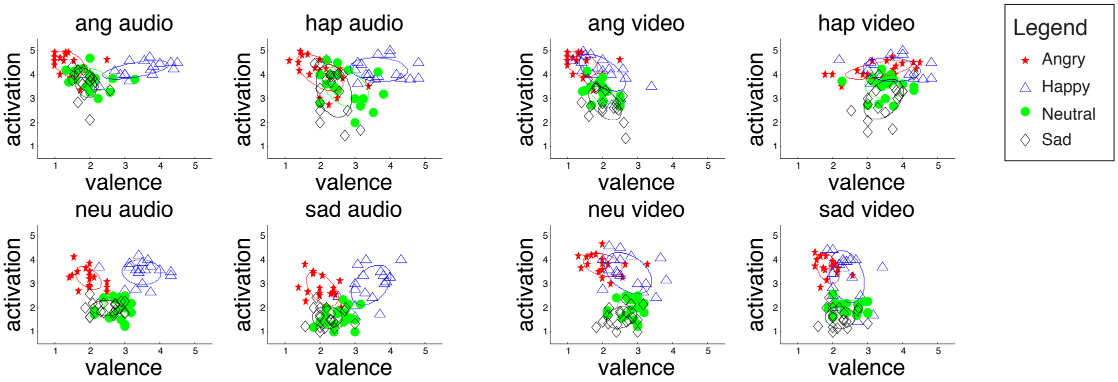


Fig. 4. The perception of the RAV clips grouped by consistent audio (left) or video (right) emotions. The red stars are angry, the blue triangles are happy, the green squares are neutral, and the black diamonds are sad. The ellipses contain 50% of the class data.

Video Emotion	Audio Emotion							
	Valence				Activation			
	A	H	N	S	A	H	N	S
A		✓$_{A*V}$	✓$_{A*V}$	✓$_{AV*}$			✓$_{A*V*}$	✓$_{A*V*}$
H	✓$_{A*V*}$	✓$_{AV}$	✓$_{A*V*}$	✓$_{A*V*}$			✓$_{A*V*}$	✓$_{A*V*}$
N	✓$_{AV*}$	✓$_{A}$		✓$_{V*}$	✓$_{AV*}$	✓$_{A*V*}$		✓$_{V}$
S	✓$_{V*}$	✓$_{A*}$	✓$_{A}$		✓$_{A*V*}$	✓$_{A*V*}$		✓$_{V}$

Table 1. The perceptual differences of the OAV and RAV clips from the unimodal valence/activation perception (A, H, N, S stand for angry, happy, neutral, sad, respectively). OAV perception is on the diagonal. RAV perception is on the off diagonal. ANOVA indicated overall group differences (uni/multimodal presentation). T-tests were performed given significant differences between group means (ANOVA, $\alpha = 0.05$). Subscripts of A or V indicate that the audio-only or video-only group means, respectively, were different from the multimodal group mean at the $\alpha = 0.05$ level (t-test). The asterisk indicates a significance of $\alpha \leq 0.001$.

subplot shows the same effect for angry video. It is interesting to note that happy video has a markedly stronger impact on perception change than does happy audio (Figure 4). This highlights the perceptual relationship between happiness and video information [21].

We use Analysis of Variance (ANOVA) analyses to quantify perception change. We first analyze the group mean (if the perception from the three different presentation conditions differ). If there is a statistically significant difference we use t-tests to determine if the audio-visual perception differs from (1) the video-only perception and (2) the audio-only perception. The results suggest that the perception of OAV valence differs significantly from that of the unimodal presentation of happiness (both audio and video). The perception of OAV activation differs significantly only from the unimodal presentation of sadness. This finding does not support hypothesis 1 (the perception of OAV presentations is significantly different from that of the unimodal presentations). This may be due to the clarity of the emotional information discussed at the end of Section 2.1. The findings are summarized in Table 1. The second hypothesis asserts that the perception of valence/activation from the RAV stimuli is significantly different from that of either of the audio-only or video-only perception. The results support this hypothesis in general. In all twelve RAV presentations, the valence perception is statistically significantly different from that of the audio and the video presentations at $\alpha = 0.05$. The RAV activation perception is statistically significantly different from the audio-/video-only presentations in nine out of twelve presentations at $\alpha = 0.05$, suggesting that the audio and/or video information contribute to the multimodal perception of the RAV clips in both activation and valence dimension. This supports the second hypothesis that the RAV perception is different from that of either unimodal component.

5. REGRESSION ON REPORTED PERCEPTION

The regression studies approximate audio-visual feature reliance by indicating types of features and combinations of features that predict

Stimuli	Audio Cues		Video Cues		AV Cues	
	val	act	val	act	val	act
OAV	0.726	0.927	0.795	0.888	0.858	0.981
RAV	0.060	0.494	0.571	0.094	0.709	0.798

Table 2. Average adjusted R^2 value when regressing on mean reported perception of OAV and RAV clips using unimodal (e.g., audio-only or video-only) or multimodal features.

reported emotion perception. Models that are well correlated with perception indicate that the features used in the model may be important to audio-visual perception. In all cases, the reported perception is modeled using stepwise linear regression (LR). LR has been used widely in the behavior modeling community to predict evaluator perception [22]. In the first study, the dependent variable is either the average reported OAV/RAV valence or activation. In the second study (Section 6), the dependent variable is the change in reported perception from the original OAV presentation to the RAV presentation. In all cases, the independent variables are entered into the model at a significance value of $\alpha \geq 0.95$ and removed at $\alpha \leq 0.90$

The first study models the reported valence and activation perception of the OAV and RAV presentations to learn whether feature-level information can be used to estimate perception. Three LR models are constructed for the OAV stimuli using: (1) audio-only, (2) video-only, and (3) audio-visual features. The audio-only and video-only LR models are correlated with the valence rating and highly correlated with the activation rating (Table 2). Valence is correlated more highly with the video-only model and activation is correlated more highly with the audio-only model. The correlation between these modalities and dimensional perception has also been observed in the emotion classification literature [7,23]. However, the strength of both unimodal models highlights a potential problem in the analysis of audio-visual feature reliance. Since evaluators attune to both modalities when viewing multimodal clips it becomes challenging to develop a strong causal relationship between multimodal cue presentation and resulting perception. The accuracy of all three models highlights the cross-channel redundancy of the audio and video information suggesting that the audio and video cues in the OAV presentations may not be providing complementary information to the evaluators. Given this redundancy it becomes challenging to establish the connection between feature presentation and perception.

The RAV clips present an opportunity to break the cross-channel correlation while investigating multimodal emotion perception. In these clips the audio and video channels are synchronized but uncorrelated with the emotion expressed in the opposite channel. The same three LR models are constructed for the RAV stimuli. The single-modality (audio-only and video-only) LR models of valence perception demonstrate that in the case of emotional noise (RAV stimuli) valence perception may be heavily biased by the video information. The single-modality LR models of activation perception demonstrate that activation may be more heavily influenced by the audio information (Table 2). This reinforces numerous findings in the emotion research community that point to the perceptual relationship between audio and activation and video and valence [23].

The multimodal LR models of valence and activation perception for RAV stimuli demonstrate that the correlation is higher than that of either single-modality model (Table 2). This points to multimodal feature interactions between the modalities. This potential multimodal interaction is important because the single-modality results suggest a single-modality perceptual bias (valence-video and activation-audio). The multimodal model suggests that despite the emotionally independent generation of the feature streams, both modalities are integrated during perception and contribute to dimensional evaluation. The increased correlation of the both the valence and activation multimodal LR models suggests that there are inherent cross-modal perceptual interactions.

Change	Dimension	Error	Adj. R^2
video	valence	0.565 ± 0.416	0.863 ± 0.012
	activation	0.575 ± 0.423	0.727 ± 0.012
audio	valence	0.595 ± 0.469	0.619 ± 0.016
	activation	0.545 ± 0.401	0.860 ± 0.007

Table 3. Average error and average adjusted R^2 from LR models that predict the change in perception using feature change cues.

6. REGRESSION ON CHANGE IN PERCEPTION

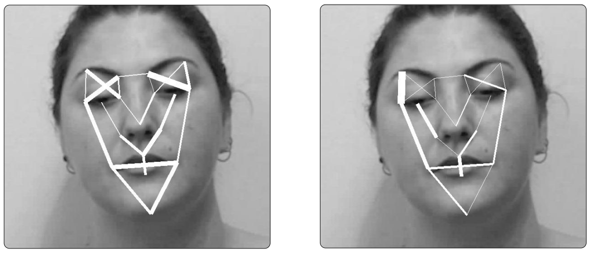
The RAV stimuli can also be used to determine the effect of single-modality change (delta) on multimodal perception. The RAV stimuli are significantly different from the OAV stimuli only over a single dimension (e.g., OAV = happy audio and video, RAV = happy audio and angry video, they differ based on the video features). Therefore, the RAV stimuli provide an opportunity to systematically investigate how major changes in audio-visual perception can be explained by changes in the presentation of a single-modality. The previous RAV LR results suggested that perception is heavily biased by a single modality (Table 2). In the second study the LR models demonstrate how single-modality change can explain perpetual differences, even when modalities are less strongly correlated with the perceptual dimension of interest. The LR models in this section estimate the change in perception (OAV to RAV) based on the change in the unimodal audio/visual feature values. This experiment hypothesizes that perception change can be explained by feature change.

The results demonstrate a similar dimensional-modality interaction. LR models with independent variables of change in video features are more correlated with valence perception than those with change in audio features. The opposite is true for activation perception, the LR models based on audio feature change are more strongly correlated with reported perception than those based on video feature change (Table 3). Previously (Table 2), the results demonstrated that audio features did not contribute to valence perception and video features did not contribute to activation perception in RAV stimuli. However, the results suggest that these features can be used to predict changes in perception. This discrepancy is important because it demonstrates that while the features of the perceptually "weaker" modality do not strongly inform resulting perception, changes in these features can be used to predict how perception will change. This result, in conjunction with the statistical analyses of perception change (Table 1), demonstrates that the "weaker" modality features contribute to the statistically significant change in perception.

The LR models of RAV perception change can also be used to highlight perceptually important features. Figure 5 presents the video features that contribute to valence or activation perception. The results demonstrate that the majority of the facial features contribute to the assessment of change in valence perception. The results further demonstrate that changes in the mouth movement features are an important predictor for change in activation perception. The audio features show that MFCC, MFB, and the LSP features contribute to both valence and activation perception (Figure 6).

7. CONCLUSIONS

This paper presents a novel set of stimuli designed to enhance our understanding of audio-visual emotion perception. The stimuli contain both emotionally matched (OAV) and emotionally mismatched (RAV) sentence-level audio-visual emotion displays. We collected dimensional evaluations of the datasets and used these evaluations to build regression models that predict reported perception and change



(a) Change in valence perception. Eye related features are tied to valence perception.

(b) Change in activation perception. Mouth related features are tied to activation perception.

Fig. 5. Prominent features for change in perception (unimodal LR models). Line thickness is proportional to the number of times a feature (change in feature from OAV clip) was included in a model predicting change in valence (left) or activation (right) perception.

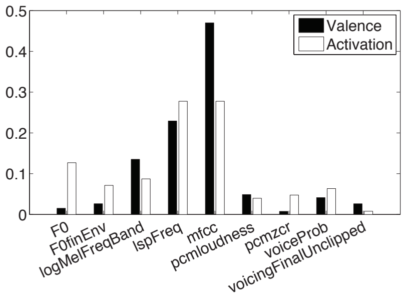


Fig. 6. Audio features used in predicting change in valence perception (black) and activation perception (white) for the unimodal (change) LR models of RAV perception. The y-axis corresponds to percentage of feature set grouped by the feature types (x-axis).

in perception (from the emotionally matched stimuli to the emotionally mismatched stimuli). These models provide insight into the how audio-visual cue interaction during audio-visual perception.

These audio-visual perceptual results have important implications in the design of multimodal affective interfaces. These interfaces require a detailed and quantitative understanding of how audio and video displays of emotion shape the perception of human users. Absent this knowledge it is challenging to produce natural automatic emotional displays. The knowledge gained from McGurk effect studies provide this description and insight into how audio and video bias gestalt perception. Future work includes the collection of additional data from a wider variety of speakers to understand how the trends observed in this study extend to novel speakers.

8. ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation (NSF RI 1217183).

References

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

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

[3] 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.

[4] 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.

[5] 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.

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

[7] E. Mower, M. Matarić, and S. Narayanan, "Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information," *IEEE Trans. on Multimedia*, vol. 11, no. 5, pp. 843–855, 2009.

[8] 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.

[9] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein, and S. Narayanan, "Sailalign: Robust long speech-text alignment," in *in Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Jan. 2011.

[10] "Speech recognition scoring toolkit," 1997.

[11] "Amazon mechanical turk," http://www.mturk.com/, Accessed: July 2012.

[12] M. Buhrmester, T. Kwang, and S.D. Gosling, "Amazon's mechanical turk a new source of inexpensive, yet high-quality, data?," *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011.

[13] M. Marge, S. Banerjee, and A.I. Rudnicky, "Using the amazon mechanical turk for transcription of spoken language," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 5270–5273.

[14] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. IEEE, 2008, pp. 1–8.

[15] R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 254–263.

[16] 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.

[17] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2729–2736.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR - introducing the munich open-source emotion and affect recognition toolkit," in *ACII*, Amsterdam, The Netherlands, Sept. 2009, pp. 25–29.

[19] B. Schuller and L. Devillers, "Incremental acoustic valence recognition: an inter-corpus perspective on features, matching, and performance in a gating paradigm," in *InterSpeech*, Makuhari, Japan, Sept. 2010, pp. 801–804.

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

[21] E. Mower, M. Matarić, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2011.

[22] M. Black and S.S. Narayanan, "Improvements in predicting children's overall reading ability by modeling variability in evaluators' subjective judgments," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012.

[23] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," *Affective Computing and Intelligent Interaction*, pp. 415–424, 2011.