UMEME: University of Michigan Emotional McGurk Effect Data Set

Emily Mower Provost, *Member, IEEE*, Yuan Shangguan, *Student Member, IEEE*, Carlos Busso, *Senior Member, IEEE*

Abstract—Emotion is central to communication; it colors our interpretation of events and social interactions. Emotion expression is generally multimodal, modulating our facial movement, vocal behavior, and body gestures. The method through which this multimodal information is integrated and perceived is not well understood. This knowledge has implications for the design of multimodal classification algorithms, affective interfaces, and even mental health assessment. We present a novel dataset designed to support research into the emotion perception process, the University of Michigan Emotional McGurk Effect Dataset (UMEME). UMEME has a critical feature that differentiates it from currently existing datasets; it contains not only emotionally congruent stimuli (emotionally matched faces and voices), but also emotionally incongruent stimuli (emotionally mismatched faces and voices). The inclusion of emotionally complex and dynamic stimuli provides an opportunity to study how individuals make assessments of emotion content in the presence of emotional incongruence, or emotional noise. We describe the collection, annotation, and statistical properties of the data and present evidence illustrating how audio and video interact to result in specific types of emotion perception. The results demonstrate that there exist consistent patterns underlying emotion evaluation, even given incongruence, positioning UMEME as an important new tool for understanding emotion perception.

Index Terms—Emotion perception, McGurk effect, Multimodal, Unimodal, Affect

1 Introduction

Emotion perception is fundamental to human communication. It underlies social communication [1]–[4], affects how we interpret our environment, and shapes how we understand the opinions and sentiments of others [5], [6]. A clear understanding of the emotion perception process has important implications for engineering including design principles for interactive agents and insights into multimodal classification algorithms. Further, this knowledge will provide insight into how typical, and by extension, atypical, emotion perception patterns can be quantified and compared, which will have impact in the assessment of mental health [7], [8].

However, emotion perception is challenging to unobtrusively interpret. The audio and video cues that accompany any given emotion state are partially redundant [9]. This redundancy renders it difficult to tease apart the individual effects of each modality. We address this through the design of novel stimuli that highlight the contribution of each modality, the University of Michigan Emotional McGurk Effect (UMEME) Dataset. UMEME contains stimuli with and without audio-visual emotional mismatch (e.g., an angry voice and a happy face) expressed dynamically over single sentences. We demonstrate that the UMEME stimuli can be used to understand how emotion is perceived and to understand how emotional mismatch can uncover the relationship between audio and video information in emotion perception.

Multimodal mismatch offers a structured approach to investigate the impact of individual modalities on emotion perception. This methodology is generally motivated by the McGurk Effect paradigm [10], an audio-visual perceptual phenomenon in which the vocal and facial channels convey two separate phonemes, yet evaluators report a third distinct sound. The classic example is the acoustic realization of "ba" and the lip movement (viseme) associated with the speech sound "ga." Together, the perception is the speech sound, "da" [9]. Listeners experience this effect even when aware of this mismatch, suggesting early audio-visual integration.

The emotion community has been motivated by this paradigm to investigate how emotion content is integrated. The stimuli have included matched and mismatched information from the facial and vocal channels [9], facial channel and context [11], and facial and body postural/positional information [12]. These stimuli are usually evaluated with either discrete emotion labels (e.g., happy or angry) [9], [13]–[17] or dimensional evaluation (e.g., valence, activation, and dominance) [11], [18]. This effect has also been studied using functional magnetic resonance imaging (fMRI) [11] and electroencephalogram (EEG) studies [12]. However, these studies have been primarily conducted using a combination of human audio and still images [9], [13], [15], [16], which makes it difficult to understand how humans integrate

- *E. Mower Provost and Y. Shangguan are with the Computer Science and Engineering Department, University of Michigan, Ann Arbor, MI, 48109. C. Busso is with the Electrical Engineering Department, University of Texas at Dallas, Richardson, TX, 75080.*
  E-mail: {emilykmp, juneysg}@umich.edu, busso@utdallas.edu
- *This work is supported by the National Science Foundation (NSF RI 1217183 and NSF RI 1217104).*

temporal information from multiple sources.

At issue is the creation of dynamic stimuli. In order to understand natural emotion perception patterns, the audio and video information should be synchronized. However, the two sources of information are often produced across separate recordings. This challenge was first addressed by Fagel in 2006 using single-word stimuli [17]. Our previous work further increased the duration of these stimuli using animated facial information, which allows for easy construction of time-synchronized dynamic stimuli [18], [19]. More recently, we demonstrated that dynamic sentence-level audio-visual stimuli could be created using human audio and video information from a single actress [20]. The perceptual patterns associated with the stimuli provided insight into how acoustic cues ("audio") and visual displays ("video") are integrated during the perception of an actress's displays. However, there are open questions relating to the extension of these results beyond a single actress to emotion expressions across individuals. In this paper, we present the first dataset that includes dynamic, sentence-level, human audio-visual stimuli with emotionally mismatched content over multiple subjects.

The UMEME dataset is derived from a series of dyadic improvisations. In each improvisation, one of the actors was required to speak a specific target sentence. Each sentence was embedded into four emotional scenarios: anger, happiness, neutrality, and sadness. These audio-visual recordings form our "original" stimuli. We manipulate these stimuli to create a new set of stimuli with mismatched emotion content. This paradigm allows us to obtain the fixed lexical content needed to synthesize emotionally mismatched utterances while allowing for enhanced naturalness compared to read speech collection paradigms [21]. We evaluate the emotion content using Amazon Mechanical Turk [22], used in our previous work [20], and model the resulting perceptual patterns.

The results demonstrate that the perception of the emotionally mismatched stimuli is different from that of emotionally matched stimuli. The mismatch introduces higher variance in the evaluations. However, there are consistent patterns in the evaluations; the dimensional perception of emotionally mismatched audio-visual stimuli can be estimated based on the dimensional perception of their unimodal components. The result is that the evaluations can be used to understand the biasing effect of audio and video information, providing insight into the effect of individual channels on categorical and dimensional perception. UMEME is a novel dataset that has the potential to increase our understanding of the relationship between audio and video information and their effect on human emotion perception.

2 Related Work

The modeling of emotion expression and perception is an important and interesting computational problem. Human emotion is composed of incredibly rich and complex signals in which the producer's and perceiver's intentions, perceptions, and expression patterns factor into the perceived content of the spoken message. This coupled perception-production process increases the complexity of the frameworks needed to correctly model and capture the behavior of the underlying signal. Furthermore, emotion expression is inherently multimodal, affecting the producer's vocal, facial, body-position, and gestural patterns. Decoding human communication cues is a multi-level, multimodal mapping problem.

There are two psychologically grounded methods for the quantification of emotion. The **dimensional** view of emotion posits that emotion can be described as points existing on a continuum, captured by axes with specific semantic meanings. Different axes have been used in countless works, most often taking on the labels valence (positive vs. negative), activation/arousal (calm vs. excited) and dominance (passive vs. aggressive) [23]–[28]. The **categorical** (or discrete) view of emotion is based on the assumption that there exists a set of emotions that can be considered "basic." A basic emotion is defined as an emotion that is differentiable from all other emotions. Ekman elucidates the properties of emotions that allow for the differentiation between the basic emotions [29]. The set of basic emotions can be thought of as a subset of the space of human emotion, forming a "basis" for the emotional space. More complex, or secondary, emotions can be created by blending combinations of the basic emotions. For example, the secondary emotion of jealousy can be thought of as the combination of the basic emotions of anger and sadness [30].

2.1 Multimodal Perception

There has been much interest in exploring multimodal emotion perception, specifically audio-visual perception [31]–[36]. Calvert and colleagues demonstrated that during perceptual tasks individuals integrate multimodal information to reduce perceptual ambiguity and increase the identification of stimuli [31].

Swerts and colleagues investigated how the emotional displays produced by blind and sighted people were perceived by an evaluator population [37]. Their results demonstrated that the emotions of sighted individuals were recognized more accurately given audio-visual or video-only information. However, the emotions of blind individuals were recognized more accurately in the audio-only condition. The results suggest both the inherent multimodality of emotion expression and the differing roles that the modalities play in conveying emotion. Audio-visual analyses have also been applied to understanding perception in populations with known perceptual deficits. Williams and colleagues demonstrated that children with autism relied on multimodal information in speech processing tasks [38]. Smith and Bennetto demonstrated that individuals with autism were less accurate at speech processing than their typically developing peers, potentially due to deficits in audio-visual integration [39].

2.2 Emotion Perception Via the McGurk Effect

The McGurk presentation paradigm is a popular method of investigating how individuals integrate audio-visual information during perception. Collignon and colleagues investigated audio-visual integration patterns using dynamic stimuli composed of emotionally mismatched facial and vocal behavior (non-linguistic utterances) under varying audio-visual noise conditions [40]. They found that given incongruence, the visual modality biased perception. Further, they found that when Gaussian white noise is added to the three color channels, thus rendering the video channel less reliable, the evaluations are biased by audio information. However, the relationship between the modalities is not well-understood given linguistic utterances. In our preliminary work, we developed a new set of emotional McGurk Effect stimuli [20]. These stimuli are sentence-length and include dynamic facial and vocal information. We found evidence supporting the link between strength of modality and perceptual effect as a function of evaluation task [20]. There have also been studies investigating the effect of incongruence across the lexical and vocal information [41], [42].

This effect has also been studied using still images mismatched with human audio. de Gelder combined these still images with single spoken words and demonstrated the effect of multimodal information on perception given mismatch, even when individuals were asked not to attune to a certain modality [43], also demonstrated in [40], [44]. Hietanen and colleagues demonstrated that this effect could be mitigated given the addition of a delay; the channels no longer interacted in the emotion evaluation and the evaluators based their decisions on the vocal signal only [44]. Interactions between emotional channels have also been studied using emotional faces paired with contextual movies [45] and mismatched facial and postural stimuli [46]. Cassell and colleagues explored the link between speech and gesture using a narrative accompanied by gestures that were either matched or mismatched to the speech content [47]. They found that gesture mismatch increased the number of errors in retelling accuracy, suggesting that listeners jointly integrate both the speech and gestural cues [47].

Emotional mismatch has also been investigated using the emotional Stroop test. The Stroop test requires that the participant attend to the colors of words rather than the content of the words. In the emotional Stroop (e-Stroop), the words have emotional significance. A slower response to the emotional words compared to neutral words indicate that the emotional content of the words affects performance [48].

3 Description of UMEME Database

The UMEME dataset is composed of two distinct types of data: human-produced ("original") and artificially created ("reconstructed") expressions. When discussing the audio-visual emotion displays we will speak of original audio-visual (OAV) clips and reconstructed audio-visual (RAV) clips. The RAV clips are created from the audio-visual information of the OAV clips. In the remainder of this section, we describe the collection of the OAV clips and the creation of the RAV clips.

3.1 Original Recording

The OAV clips were recorded from semantically neutral utterances embedded within emotionally charged improvisational scenarios. This collection paradigm results in stimuli with the same lexical content across each of the four emotion classes, which will be instrumental in allowing us to artificially create emotionally mismatched audio-visual displays. The stimuli were generated from 15 semantically neutral sentences: (1) How can I not; (2) I'm quite sure that we will find some way or another; (3) Ella Jorgenson made the pudding; (4) The floor was completely covered; (5) They are just going to go ahead regardless; (6) It has all been scheduled since Wednesday; (7) I am going shopping; (8) A preliminary study shows rats to be more inquisitive than once thought; (9) That's it the meeting is finished; (10) I don't know how she could miss this opportunity; (11) It is raining outside; (12) Your dog is insane; (13) She told me what you did; (14) Your grandmother is on the phone; and (15) Only I joined her in the ceremony.

Each sentence was embedded into four emotional contexts (anger, happiness, neutrality, and sadness). We chose to embed the utterances in improvised scenarios rather than rely upon read speech (data recorded by an individual reading an utterance in different emotions) to enhance the naturalness of the expression (see [21] for a discussion of the naturalness of the utterances). For example, the sentence, "I am going shopping," was used to admonish a roommate who ate all the snacks in the house prior to a party (emotion: anger) and as an expression of celebration following a promotion at work (emotion: happiness). Each utterance was performed in a dyadic interaction with a "scene partner" acting as a friend to whom the actor was recounting his/her experiences. Please see [21] for more detail regarding the OAV recording process.

The OAV clips were recorded from 12 actors (6 male, 6 female) from the School for Arts and Humanities at the University of Texas at Dallas (UT Dallas). Each utterance was recorded in each of the four emotion scenarios, resulting in four OAV clips per utterance. Four of the actors (2 male, 2 female) recorded the first 10 utterance scenarios. The remaining eight actors recorded scenarios associated with all fifteen sentences. One female actor's data were not used due to difficulties in post-processing. This resulted in a set of 145 utterances over the eleven speakers with 580 emotional realizations (145 utterances * 4 emotions). Each utterance is associated with a set of four OAV's: angry, happy, neutral, sad.

3.2 Creation of McGurk Effect Clips

The RAV clips contain emotionally mismatched audio and video information extracted from the OAV clips.

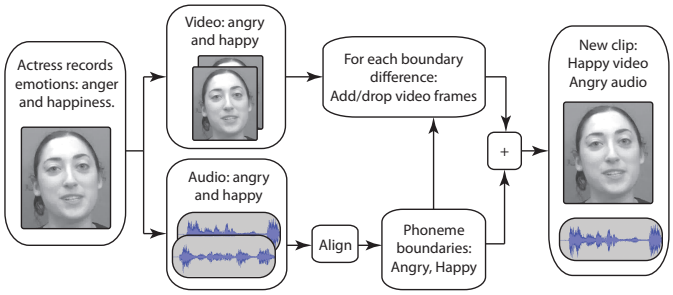

Fig. 1: This figure describes the method used to create the RAV clips. The audio is extracted from the OAV clips (same lexical content, spoken by the same actor). The phoneme boundaries are found using forced alignment and differences in timings are used to provide guidance for the video warping. The audio is combined with the warped video, resulting in the RAV clip.

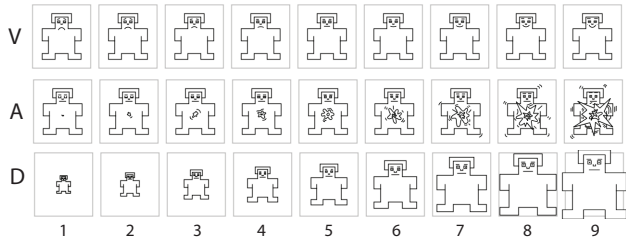

Fig. 2: Self Assessment Manikins (SAM) used to assess valence (V), activation (A), and dominance (D). The values are from 1 (left) to 9 (right).

There are 12 RAV clips derived from each set of four OAV clips; one emotion is assigned to the audio channel and another to the video channel ($_4P_2$ = 12). RAV clips are created separately for each actor.

The first step in RAV creation is to find the timing differences between the two OAV files that contribute the audio and video information (Figure 1). We first extract the audio and video information from the two OAV clips. This results in four files: two video files ("Original Video", OV) and two audio files ("Original Audio", OA). We will use these four components to create two new RAV files. Using the example from Figure 1, the RAV set would include: (1) angry audio with happy video and (2) happy audio with angry video.

We achieve audio-visual synchronization by warping the video from one OAV file to match the timing of the other OAV file. Thus, when the video from the first and audio from the second are combined, they will appear time synchronized. We identify the differences in timing by using the phoneme-level transcripts from the two extracted OA files. We obtain these transcripts by force aligning the OA clips to their known transcripts using SailAlign [49], with manual corrections as necessary. SailAlign is a toolkit developed by Katsamanis and colleagues at the University of Southern California (USC). The system requires a known transcript. It then uses Hidden Markov Models to identify the timing boundaries that best fit the associated wave file given the known transcript. The output is a set of hypothesized word- and phoneme-level boundaries. However, the output of the system cannot directly be applied to the warping process due to allowable differences in pronunciations that result in slightly different phoneme-level transcripts. This occurs even when the lexical content and speaker are fixed (only the emotion changes). We align the phoneme-level output using NIST's SClite scoring tool [50], which takes two transcripts as input and finds the best alignment between them using dynamic programming. The goal is to identify areas of match and mismatch. We then use a heuristic approach to create associations between the phonemes in each emotion and use this alignment information to identify the OA timing differences that will be used to produce a series of video warping instructions.

The second step is to warp the OV files (Figure 1), resulting in warped video files ($\widehat{OV}$). We use video warping, rather than audio warping, because our initial experiments demonstrated that this style of warping was less susceptible to perceptual artifacts. We present the perceptual effects of video warping in Section 5.3. We create the initial warped video file by extracting the video frames from the original video clip and up sampling. We adjust the timing of the $\widehat{OV}$ clip by adding and dropping frames in accordance with the phoneme-level timing differences observed between the two OA clips in step one. Continuing our example, consider the OA and OV files extracted from the angry and happy OAV files (goal: the creation of an RAV with angry audio and happy video). For clarity, we will add the subscripts $angry$ and $happy$. For each phoneme in OA_{angry} we take one of three actions: (1) if the phoneme duration of OA_{happy} is the same, the video information for $\widehat{OV}_{happy}$ is not changed; (2) if the phoneme duration for $OA_{angry} < OA_{happy}$, video frames are deleted from $\widehat{OV}_{happy}$; and (3) if the phoneme duration for $OA_{angry} > OA_{happy}$, video frames are added to $\widehat{OV}_{happy}$. We distributed the added or removed video frames evenly over the duration of the OV phoneme. For example, if three video frames must be added to a phoneme that is 16 video frames in length, then the first extra video frame (a copy of the previous frame) is inserted at position 4, the second at position 8, and the third at position 12. Initial black frames are added to the RAV clip if the OA clip contains longer initial silence than the OV clip. Once this process has terminated, the warped video file, $\widehat{OV}_{happy}$, has the same phoneme-level durations as OA_{angry}. The two files are combined to create the new audio-visual synchronized RAV. This process is repeated over all OAV clips for a total of 1,740 RAV files (145 utterances * 12 mismatch conditions), see Figure 1. The initial dataset contained the 1,740 RAV files, the 1,740 warped video files, the 580 audio files, 580 video files, and 580 OAV files for a total of 5,220 files.

4 EVALUATION

We conducted evaluations of the UMEME data in two stages: (1) to identify the RAV stimuli with perceptual artifacts and (2) to label the emotion content of all clips without perceptual artifacts. The evaluations were conducted on Amazon's Mechanical Turk [22] ("MTurk"). MTurk, and crowd sourcing platforms more generally, have seen increasing use in the fields of emotion modeling [20], [21], [36], [51], [52], sentiment analysis [53], analysis of social behavior [54], transcription [55], [56], perceptual evaluation [57], [58], and assistive technology for the blind [59]. These platforms allow for the rapid collection of many evaluations from a large disparate user population [60].

There were a total of 135 unique evaluators for the perceptual artifact evaluation task and 426 for the emotion evaluation task. There were 20 evaluators who performed both tasks. The evaluators who participated in both evaluations contributed 2,398 evaluations out of the 43,744 non-rejected emotion evaluations (5.48% of the total emotion evaluations).

4.1 Perceptual Artifact Evaluation

The full set of 5,220 clips was manually reduced to remove clips that had clear perceptual artifacts. This resulted in 4,390 stimuli (1,325 RAV, 1,325 warped video-only, and 580 OAV, OA, OV). We then conducted the first MTurk study to identify additional RAV clips containing perceptual artifacts after the video warping. There were a total of 135 unique evaluators. Evaluators could participate in an unlimited number of evaluations and the clips to evaluate were assigned randomly. The mean number of clips evaluated by each evaluator was 42.13 with standard deviation 70.57. The median number of evaluations was 15.

Each evaluator was asked to assess the synchronization on a five-point Likert scale ($1 =$ very poor to $5 =$ perfect), naturalness on a five-point Likert scale ($1 =$ very poor to $5 =$ perfect), and smoothness on a binary scale ($1 =$ disconnected to $2 =$ smooth). Three evaluators rated each stimulus. The ratings in each dimension were summed (e.g., if three evaluators noted that the clip was smooth, the rating was a six). Utterances with a synchronization score less than 8 and a smoothness score less than 6 were eliminated. This resulted in 3,704 clips (990 RAV, 990 warped video-only ($\widehat{OV}$), and 580 OAV, OA, and 564 OV).

4.2 Emotion Evaluation

The purpose of the second MTurk study was to assess the emotion content of the clips remaining after the first MTurk study. The set of clips included: RAV, OAV, OA, OV, $\widehat{OV}$. The evaluators were asked to assess the emotion using the valence (negative vs. positive), activation (calm vs. excited), and dominance (passive vs. dominant) dimensions using Likert scales from one to nine. Each position was associated with a Self Assessment Manikin (SAM, Figure 2) [61], a powerful approach to describe emotions using iconic images. SAMs provide interpretations of the emotional dimensions without assigning linguistic terms (e.g., anger, sadness) and have been used successfully in the evaluation of different emotional corpora [62], [63]. Their use has been shown to improve the reliability and inter-evaluator agreement of the evaluations [64]. The evaluators also assessed the primary emotion of the clip from the set: angry, happy, neutral, sad, and other and an unlimited number of secondary emotions from the set: accepting, angry, annoyed, anticipating, disgusted, embarrassed, excited, fearful, happy, nervous, neutral, pitying, regretful, relaxed, sad, surprised, other, and none. They rerated the synchrony of the clips by answering a binary question (yes/no). Finally, the evaluators were asked to identify the stimulus type from the set: audio, video, or audio-visual. If an evaluator answered this question incorrectly, the evaluation was rejected. If an evaluator answered three stimulus type questions incorrectly in a row, the evaluator was rejected. There were a total of 426 non-rejected unique evaluators and 44,128 evaluations. Only 384 evaluations were rejected.

The evaluation was structured to enable analyses of user-specific emotion perception patterns and changes in these patterns over time. In an evaluation session, an evaluator was assigned a set of 60 stimuli (they could quit after evaluating 30 stimuli, the mean number of evaluations was 55.10 ± 10.76 per session). The software occasionally assigned users more than 60 stimuli if the evaluators refreshed the page before it finished loading. This occurred in 8.68% of the recorded sessions (68 out of 783 total sessions over all evaluators). Each evaluator was allowed to participate once per day. The evaluators could participate in multiple sessions with the requirement that the sessions be separated by at least 24-hours (the mean of the total number of evaluations per evaluator was 103.59 ± 78.00).

The set of 60 stimuli included two components: (1) a "family" of clips and (2) randomly chosen clips from the remaining set of stimuli. A clip family is defined for all pairs of utterances and actors (e.g., sentence 5 spoken by female 1). It includes all possible combinations of audio information and video information (four OAV, 12 RAV, four OA, four OV, and 12 $\widehat{OV}$) for a maximal set size of 36 stimuli. There exist certain families with less than 36 clips after the first MTurk experiment; the average family size is 25.54 ± 7.13 clips. There are 11 families with 36 clips (maximal families) and one family with only eight clips (minimal family). The presentation order of the 60 clips was randomized.

The challenge with an evaluation task is to separate differences in the opinions of individual evaluators from evaluation noise. In this experiment, we use weighted kappa to identify evaluators whose evaluations are likely noise given the evaluations of other individuals. We have used this measure in our prior work [20]. It is important to note that we do not perform evaluator normalization in this paper. For each evaluator, we

	Valence	Activation	Dominance	Overall
None-All	1.20 ± 0.36	1.59 ± 0.35	1.70 ± 0.38	1.50 ± 0.42
K-All	1.13 ± 0.35	1.51 ± 0.34	1.62 ± 0.37	1.42 ± 0.41
S-All	0.68 ± 0.31	0.99 ± 0.33	1.10 ± 0.37	0.92 ± 0.38
KS-All	0.61 ± 0.32	0.89 ± 0.35	1.00 ± 0.39	0.83 ± 0.39
OA	0.60 ± 0.35	0.90 ± 0.35	0.99 ± 0.38	0.83 ± 0.40
OV	0.62 ± 0.30	0.90 ± 0.35	0.98 ± 0.39	0.83 ± 0.38
$\widehat{OV}$	0.62 ± 0.33	0.91 ± 0.34	1.01 ± 0.39	0.85 ± 0.39
OAV	0.58 ± 0.31	0.83 ± 0.34	0.96 ± 0.39	0.79 ± 0.38
RAV	0.62 ± 0.31	0.91 ± 0.34	1.02 ± 0.38	0.85 ± 0.39

TABLE 1: A description of the standard deviation of the reported valence, activation, and dominance perception. The entries include the mean and standard deviation of the standard deviations associated with each utterance. KS stands for Kappa and Sort. Sort refers to dropping the highest and lowest evaluations. The top half describes the standard deviation for all clips, the bottom half uses KS and describes the standard deviation for each data type.

count the number of times the evaluator agrees with the dimensional assessment for valence, activation, and dominance of all evaluators who observed the same stimuli. We penalize off diagonal terms in proportion to their distance from the diagonal. For example, if one evaluator noted a valence of 5 and another evaluator noted a valence of 9, this difference would be penalized more heavily than one evaluation of 5 and another of 6. The penalty for the difference between the judgment of any two evaluators (defined as $eval_i$ and $eval_j$) was defined as: $2^{|eval_i - eval_j|}$. We calculate the average weighted kappa for each evaluator and drop all evaluators with an average weighted kappa in the lowest 10th percentile. This reduced our evaluator population to 383 evaluators. We further smooth the evaluation of each dimension (valence, activation, and dominance separately) by removing the two highest and lowest ratings for each utterance. This allows us to smooth out noisy evaluations when the noise is not systematic for a given evaluator. For example, a given person may not be paying attention during a single evaluation or may confuse the scales for a single evaluation. If this process resulted in zero evaluations for a given clip, only the top and bottom evaluations were dropped. This results in 6.64 ± 2.18 evaluations per utterance (from an original 11.91 ± 2.13 evaluations per utterance). The reduction in the standard deviation of evaluations can be seen in Table 1, which describes the effect of each component of the evaluator smoothing. 'None' refers to retaining the full set of evaluations, 'K' (kappa) refers to smoothing achieved by rejecting the lower 10th percentile of evaluators, and 'S' (sort) refers to smoothing by rejecting the upper and lower evaluations. 'KS' refers to both kappa and sort smoothing.

5 Categorical Emotion Perception

In this section, we examine the categorical labels assigned to the UMEME stimuli. The results of this analysis will provide insight into how evaluators make assessments of categorical emotion in the presence of emotionally consistent information. The label of a stimulus is the emotion assigned by a majority of the individual evaluators. The labels include: angry ('ang'), happy ('hap'), neutral ('neu'), and sad ('sad'). Evaluators occasionally used the label other ('oth') for the assignment of OA (2.93%), OV (2.66%), $\widehat{OV}$ (3.03%), OAV (3.10%) and RAV (6.26%) stimuli. Stimuli without a majority voted label are assigned to class 'xxx.'

Perceived emotion → Target emotion ↓		ang	hap	neu	sad	oth	xxx
OAV	ang	0.66	0.02	0.19	0.00	0.06	0.07
	hap	0.01	0.79	0.13	0.00	0.02	0.04
	neu	0.03	0.04	0.78	0.05	0.03	0.07
	sad	0.01	0.01	0.27	0.62	0.01	0.08
OA	ang	0.50	0.03	0.36	0.01	0.03	0.08
	hap	0.20	0.31	0.37	0.01	0.04	0.06
	neu	0.01	0.03	0.86	0.03	0.03	0.04
	sad	0.05	0.00	0.59	0.26	0.02	0.08
OV	ang	0.54	0.03	0.30	0.05	0.02	0.07
	hap	0.01	0.85	0.08	0.01	0.03	0.02
	neu	0.05	0.06	0.72	0.06	0.04	0.08
	sad	0.04	0.01	0.23	0.64	0.02	0.07

TABLE 2: The agreement between evaluator judgment and actor target for the OAV, OA, and OV stimuli. The class 'xxx' represents no majority agreement.

5.1 Original Audio-Visual (OAV)

The OAV clips can the thought of as having two separate labels: (1) the actor targets and (2) the evaluator assessments. The actor targets were assigned by the scenario description and the evaluations were assigned using a majority vote over the primary labels from the MTurk evaluations. In a perfect world, the actor targets and evaluator assessments would match, indicating that the emotions that the actors intended to produce were perceived by the evaluator population. However, many studies have identified the mismatch that exists between actor target and evaluator perception (for example, see [65]). We found that the perception of the evaluators agreed with the target 71.21% of the time. See Table 2 for emotion-specific details.

5.2 Unimodal Expression

In this section, we assess how the perception of categorical emotion differs when the presentation is unimodal, instead of multimodal. It is expected that the perception of the evaluators will agree less strongly with the actor target in the unimodal presentation scenario because the evaluators will have access only to a subset of the emotional modulations produced by the actors. Emotion classification literature provides evidence that audio reliably conveys activation information and video reliably conveys valence information [66]. Consequently, when presented with a subset of the audio-visual information it is expected that confusion will arise between emotions characterized by similar valence or activation.

The results demonstrate that the confusion between actor target and human evaluation increases when only audio information (48.28% agreement) or only video information (68.44% agreement) is presented to the evaluator population. During audio-only presentations, there is increased confusion between actor targets of happiness and evaluator perception of anger (the reverse is not true, Table 2). This confusion is between emotions with similar activation (both are active), but different valence (one is negative, the other positive). As expected, it is difficult to differentiate between valence using only the audio modality [67]. There is also an increased confusion between all emotions and neutrality compared to the OAV presentations (Table 2). Neutrality is differentiated from the other emotions across activation (anger and happiness) and valence (sadness). Consequently, the sadness-neutrality confusion is expected. However, the anger-neutrality and happiness-neutrality confusion suggests that video information, in addition to audio information, is needed to separate these emotions. In fact, the results demonstrate that given video information, the confusion between happiness and neutrality decreases to 8%, suggesting that video information is sufficient to differentiate between these classes of emotion. However, the confusion between anger and neutrality remains high (29.58%) suggesting that both audio and video information are needed to make this assessment.

5.3 Unimodal Warped Video Expression

During the construction of the RAV stimuli we add and drop frames from the video signal to match the timing of the audio signal. However, different emotions generally exhibit different durational characteristics [68]. Perception studies have investigated the relationship between changes in the timing of emotional speech and perception (e.g., speech synthesis studies [69]). In this section, we investigate how video warping changes the perception of categorical emotion.

In general, the perception of the warped emotion is similar to that of the unwarped emotion. In 79.74% of OV stimuli perceived as angry, happy, neutral, and sad, the associated $\widehat{OV}$ stimuli have the same evaluated emotion label after warping. The greatest change in perception occurs after warping sad expressions, which become confused with anger, neutrality, or become ambiguous ("xxx"). Angry utterances become confused with neutrality. Neutral expressions become more generally confused after warping (Table 3).

We hypothesize that the effect of the video warping procedure depends on the degree of emotional subtlety in the clip, positing that emotionally subtle expressions will be more greatly affected by the warping compared to more stereotypical expressions of emotion. We characterize emotional subtlety using the disagreement between actor target and evaluator assessment, reasoning that clips whose evaluations agree with the actor targets are less subtle than clips whose evaluations disagree with the actor targets. We refer to these clips as "agreement" and "disagreement" clips, respectively. We find that the video warping changes the perception of the disagreement clips more often than that of the agreement clips for the emotions of happiness, neutrality, and sadness (observe differences between the diagonals for the agreement and disagreement clips in Table 4). In the agreement stimuli, 85.41% of the $\widehat{OV}$ stimuli are perceived as having the same label as the OV stimuli. If we only consider OV and $\widehat{OV}$ stimuli labeled as angry, happy, neutral, or sad, this percentage increases to 91.61%. In the disagreement stimuli, only 53.02% of the $\widehat{OV}$ stimuli are perceived as having the same label as the OV stimuli. If we only consider $\widehat{OV}$ stimuli labeled as angry, happy, neutral, or sad this percentage increases to 83.08%. This suggests that subtle variations in phoneme-level timing has a larger effect on the perception of emotion in subtle displays, compared to less subtle displays.

		$\widehat{OV}$					
		ang	hap	neu	sad	oth	xxx
OV	ang	0.82	0.00	0.09	0.01	0.01	0.07
	hap	0.00	0.91	0.03	0.00	0.00	0.05
	neu	0.05	0.02	0.80	0.02	0.04	0.07
	sad	0.08	0.01	0.12	0.66	0.02	0.12

TABLE 3: The labels associated with the $\widehat{OV}$ stimuli compared to the labels associated with the OV stimuli. The diagonal entries are the clips for which warping did not change the perceived emotion. Note that label 'xxx' refers to a lack of majority vote agreement.

			$\widehat{OV}$					
			ang	hap	neu	sad	oth	xxx
OV	Agree	ang	0.85	0.00	0.08	0.01	0.01	0.05
		hap	0.00	0.95	0.02	0.00	0.00	0.03
		neu	0.02	0.02	0.88	0.02	0.01	0.07
		sad	0.04	0.01	0.12	0.71	0.02	0.09
	Disagree	ang	0.59	0.00	0.18	0.06	0.00	0.18
		hap	0.03	0.66	0.16	0.00	0.00	0.16
		neu	0.09	0.03	0.71	0.03	0.07	0.07
		sad	0.23	0.00	0.11	0.43	0.00	0.23

TABLE 4: Agreement between perceived and target for the $\widehat{OV}$ stimuli grouped by OV stimuli with evaluator labels that "agree" or "disagree" with the actor targets.

5.4 Reconstructed Audio-Visual (RAV)

RAV categorical perception allows us to understand how emotional assessments are made given multimodal cues that contain conflicting information. These results will be further extended in Section 6 to understand how the dimensional assessments of the RAV stimuli change from the original audio-only, video-only, and audio-visual clips. The evaluators assigned clips to the classes of anger (24.85%), happiness (19.39%), neutrality (40.61%), sadness (8.89%), and other (6.26%). This suggests that evaluators tended to assign labels of neutrality in the presence of emotional "noise." This may also relate to the large percentage of OA clips labeled neutral.

We note that the effects of the audio (OA) and visual ($\widehat{OV}$) stimuli depend on the emotional content

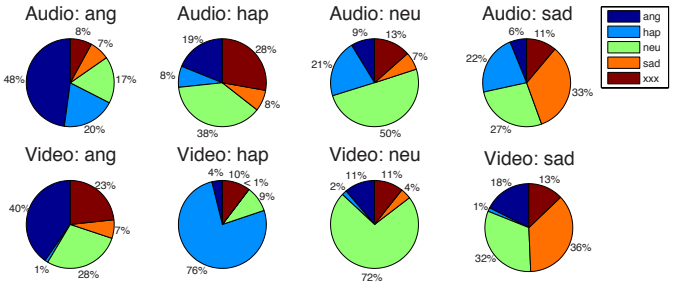

Fig. 3: The discrete evaluation of the RAV clips when grouped by *evaluated* emotion on the audio channel (upper row) and video channel (lower row). 'xxx' refers to either the class of 'other' or no agreement.

of each stimulus source. When the audio stimuli contain anger, happiness, neutrality, or sadness, the RAV clips are perceived as having the same emotion content $48\%, 8\%, 50\%$, and 33% of the time, respectively (Figure 3). However, when the video stimuli contain the same emotion content, the RAV clips are perceived as having the same content $40\%, 76\%, 72\%$, and 36% of the time, respectively (Figure 3). These results suggest that the audio and video content bias the perception of categorical emotion differently depending on emotion class. Video biases happy and neutral evaluations most strongly (in terms of categorical judgment, dimensional judgment is explored in Section 6). Sad video most weakly biases perception towards the class of sadness. The results suggest that angry video may have a weaker effect on perception compared to angry audio. Interestingly, happy audio most weakly biases RAV perception towards the class of happiness (just 8% of the time).

6 Dimensional Perception

Categorical evaluation provides important insight into how audio and video information are integrated during the emotion perception process. However, the discrete nature of the perceptual judgments makes it challenging to understand the interplay between audio and video information at a finer level. In this section, we discuss the dimensional perception associated with our stimuli set, describing how attributions of valence, activation, and dominance (VAD) change as the emotion content expressed across the face and voice change. We first describe the VAD perception of the original stimuli (OAV, OA, OV), we explore the effect of warping on VAD perception ($\widehat{OV}$), and finally investigate how changes in unimodal VAD affect the perception of RAV VAD. We focus only on the stimuli labeled as angry, happy, neutral, and sad, due to the paucity of data labeled other.

The dimensional perceptual evaluations have two important characteristics: the presentation type (e.g., OV) and the emotion class (e.g., angry). We conducted a two-way ANOVA to understand the impact of these characteristics on dimensional perception. We assert significance when $p < 0.01$ both for this analysis and all following analyses in this section. In the valence dimension, we find a significant main effect only for emotion class and a significant interaction between emotion class and presentation type. In the activation and dominance dimensions, we see that both main effects are significant and that the interaction effect is significant (Table 5). This suggests that in the activation and dominance dimensions the perceptions of the discrete classes and the presentation types are well separated. We observe significant interaction effects in all three dimensions. In the remainder of this section we explore the perpetual patterns associated with each of the presentation types.

Dimension	Presentation	Emotion Class	Interaction
Valence		**	**
Activation	*	**	**
Dominance	**	**	**

TABLE 5: The significance of the main effects (presentation and emotion class) and interaction effects on the perception of dimensional valence, activation, and dominance. The designation '*' indicates significance at $p < 0.01$, '**' indicates significance at $p \approx 0$. No designation indicates lack of significance.

6.1 Original Audio Visual

Previous research has demonstrated that anger and sadness are characterized by low valence, neutrality by medium valence, and happiness by high valence. Anger and happiness are characterized by high activation, neutrality by medium activation, and sadness by low activation [20]. The perceptual results of the current stimuli agree with previous findings (Figure 4). As seen in previous studies, the dimensions of activation and dominance are highly correlated (0.80) and specifically for the emotions of anger, neutrality, and sadness (0.80, 0.83, and 0.81, respectively). Activation and dominance are moderately correlated for happiness (0.68). All graphs will be shown in the valence-activation space.

We investigate how the emotion classes are differentiated across each dimension. We use a one-way ANOVA, testing the effect of emotion class on valence, activation, or dominance perception. In each of the three dimensions we find a significant effect (valence: F(3, 520) = 866.58, $p \approx 0$; activation: F(3, 520) = 253.97, $p \approx 0$; dominance: F(3, 520) = 230.87, $p \approx 0$). We use a multiple comparisons test to assess the pairwise differences between emotion classes to understand if the pairs of emotion classes are differentiated in the valence, activation, or dominance dimensions (again asserting significance if $p < 0.01$). In the valence dimension, all pairs of emotions have significantly different group means excepting anger and sadness. In the activation and dominance dimensions, all emotion pairs have significantly different group means. This suggests that the dimensional perceptions associated with the discrete classes are well separated.

6.2 Unimodal

We repeat the ANOVA analysis discussed on the OAV data on the OA data and find evidence for an effect of

		Valence	Activation	Dominance
OA	ang	3.35 ± 0.66	5.58 ± 0.92	5.69 ± 1.00
	hap	6.13 ± 0.64	5.82 ± 1.05	4.12 ± 1.10
	neu	4.65 ± 0.51	3.13 ± 0.98	2.97 ± 1.00
	sad	3.13 ± 0.59	2.80 ± 1.11	2.43 ± 0.92
OV	ang	3.23 ± 0.56	4.84 ± 0.98	5.26 ± 1.09
	hap	6.67 ± 0.84	5.41 ± 1.24	3.58 ± 0.89
	neu	4.49 ± 0.47	3.12 ± 0.84	2.89 ± 0.85
	sad	3.04 ± 0.52	2.87 ± 0.89	2.65 ± 0.90
$\widehat{OV}$	ang	3.31 ± 0.63	5.11 ± 0.94	5.41 ± 1.12
	hap	6.53 ± 0.85	5.43 ± 1.15	3.65 ± 0.89
	neu	4.51 ± 0.57	3.18 ± 0.98	3.00 ± 0.95
	sad	3.29 ± 0.71	3.15 ± 1.08	2.88 ± 1.08

TABLE 6: The VAD ratings for the OA, OV, and $\widehat{OV}$ stimuli. The emotion labels of the $\widehat{OV}$ stimuli are the labels of the associated OV stimuli prior to warping, to understand the effect of warping on VAD perception.

emotion class on dimensional perception (valence: F(3, 520) = 392.42, $p \approx 0$; activation: F(3, 520) = 253.35, $p \approx 0$; dominance: F(3, 520) = 224.57, $p \approx 0$). We perform a multiple comparisons test to compare the VAD ratings of pairs of emotion classes. We find that in the valence dimension, all emotions have statistically significantly different group means excepting anger and sadness. In the activation dimension, the emotions of anger and happiness and of neutrality and sadness are not differentiated. In the dominance dimension, all pairs of emotions have statistically significantly different group means.

We continue with the OV stimuli. ANOVA analyses demonstrate an effect of emotion class for each VAD dimension (valence: F(3, 520) = 884.02, $p \approx 0$; activation: F(3, 520) = 206.06, $p \approx 0$; dominance: F(3, 520) = 166.67, $p \approx 0$). We used multiple comparisons to investigate differences in the emotion-specific VAD distributions. In the valence dimension, all pairs of emotions had statistically significantly different group means excepting anger and sadness. In the activation dimension, all pairs had significantly different group means excepting neutrality and sadness. In the dominance dimension, all pairs of emotions had statistically significantly different group means excepting neutrality and sadness.

ANOVA analyses on the $\widehat{OV}$ data also demonstrate an effect of emotion class for each VAD dimension (valence: F(3, 881) = 1332.57, $p \approx 0$; activation: F(3, 881) = 403.17, $p \approx 0$; dominance: F(3, 881) = 342.69, $p \approx 0$). In the valence dimension, all pairs of emotions have statistically significantly different group means excepting anger and sadness. In the activation dimension, the perception of all pairs of emotions differs excepting neutrality and sadness. In the dominance dimension, all differed significantly excepting neutrality and sadness. The results suggest that the warping did not strongly affect the relative distribution of emotion classes in the valence, activation, or dominance spaces.

We continue our investigation into the effect of warping on emotion perception using a two-way ANOVA, testing the main effects of emotion category and presentation type (i.e., OV vs. $\widehat{OV}$) on valence, activation, or dominance perception. In this analysis, the emotion

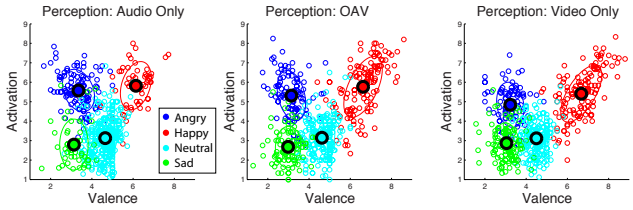

Fig. 4: The OA (left), OAV (center), and OV (right) valence-activation perception. The value of 1 represents negative (valence) and calm (activation) emotions. The value 9 represents positive (valence) and excited (activation) emotions.

labels of the $\widehat{OV}$ stimuli are the labels of the OV stimuli from which the $\widehat{OV}$ stimuli were derived (this contrasts with the previous analysis of $\widehat{OV}$ stimuli). This allows us to specifically investigate the effect of warping. We only consider $\widehat{OV}$ stimuli evaluated as angry, happy, neutral, or sad and derived from OV stimuli also evaluated as angry, happy, neutral, or sad (821 out of 990 stimuli). The results demonstrate that there is a main effect of emotion category (F(3, 1335) = 1748.71, $p \approx 0$), but no main effect of presentation type in the valence dimension (F(1, 1335) = 2.16, $p = 0.14$) or interaction effect (F(3, 1335) = 2.38, $p = 0.068$). In the activation dimension, we observe a significant main effect for emotion category (F(3, 1335) = 490.73, $p \approx 0$), but not for presentation type (F(1, 1335) = 6.5, $p = 0.0109$)) or for an interaction effect (F(3, 1335) = 1.76, $p = 0.15$). In the dominance dimension, we observe a significant effect for emotion category (F(3, 1335) = 393.09, $p \approx 0$), a borderline significant effect for presentation type (F(1, 1335) = 6.09, $p = 0.014$) and no significant interaction effect (F(3, 1335) = 0.65, $p = 0.58$). Overall, these results suggest that the categorical emotions have similar means across the presentation types.

We use the dimensional evaluations to understand how perception changes given unimodal or multimodal stimuli. Figure 4 presents a graphical depiction for how emotion perception changes from OA (left) to OV (right) to OAV (center). The points in the figure represent the valence-activation perception associated with each utterance, where the value is calculated as the average over the set of evaluators retained after smoothing (see Section 4.2). The figure demonstrates that the video modality is associated with greater separation between happiness and the other four emotion classes. The audio modality is associated with greater activation separation.

6.3 RAV Stimuli

RAV clips provide an opportunity to understand how individuals integrate audio-visual information when there is emotional complexity introduced by audio-visual emotional mismatch. Each subplot of Figures 5 and 6 demonstrate how perception is affected when the audio emotion (Figure 5) or video emotion (Figure 6) is held constant. In each subplot of Figure 5, all utterances

		Compare to OA			Compare to $OV/\widehat{OV}$		
Audio Emo	Video Emo	ΔV_{audio}	ΔA_{audio}	ΔD_{audio}	ΔV_{video}	ΔA_{video}	ΔD_{video}
ang	ang	= 0.22	= 0.11	= 0.26	= 0.13	↑ 0.61	↑ 0.56
	hap	↑ 2.65	↑ 0.47	↓ 1.03	↓ 0.55	↑ 0.59	↑ 1.01
	neu	↑ 0.38	↓ 0.64	↓ 0.59	↓ 0.81	↑ 1.85	↑ 2.18
	sad	= 0.12	↓ 0.79	↓ 0.90	= 0.14	↑ 1.85	↑ 2.14
hap	ang	↓ 2.18	↓ 0.94	= 0.54	↑ 0.69	= 0.19	↓ 0.73
	hap	↑ 0.57	= 0.37	= 0.01	= 0.16	↑ 0.73	↑ 0.47
	neu	↓ 1.23	↓ 1.19	= 0.06	↑ 0.36	↑ 1.54	↑ 1.14
	sad	↓ 2.29	↓ 1.47	= 0.10	↑ 0.75	↑ 1.41	↑ 1.36
neu	ang	↓ 0.82	↑ 1.04	↑ 1.40	↑ 0.57	↓ 0.90	↓ 1.02
	hap	↑ 1.50	↑ 1.59	↑ 0.47	↓ 0.39	↓ 0.73	= 0.21
	neu	= 0.06	= 0.03	= 0.04	= 0.04	= 0.02	= 0.09
	sad	↓ 1.05	= 0.19	= 0.17	↑ 0.51	= 0.00	= 0.15
sad	ang	= 0.07	↑ 0.95	↑ 1.66	= 0.06	↓ 1.33	↓ 1.30
	hap	↑ 2.65	↑ 1.51	↑ 0.70	↓ 0.76	↓ 1.15	↓ 0.51
	neu	↑ 0.68	= 0.09	= 0.06	↓ 0.73	= 0.38	↓ 0.43
	sad	↓ 0.34	= 0.02	= 0.02	↓ 0.30	= 0.17	= 0.20

TABLE 7: The change in valence (ΔV), activation (ΔA), and dominance (ΔD) going from the audio-only (OA) and video-only (OV or $\widehat{OV}$ depending on if the multimodal stimuli is OAV or RAV, respectively) stimuli to the OAV and RAV stimuli. The OAV stimuli are in blue (they do not include RAV stimuli whose audio and video were evaluated with the same emotion content). ↑ represents a statistically significant increase, ↓ represents a statistically significant decrease, and = represents no statistical difference (two sample t-test, significance at $p < 0.01$).

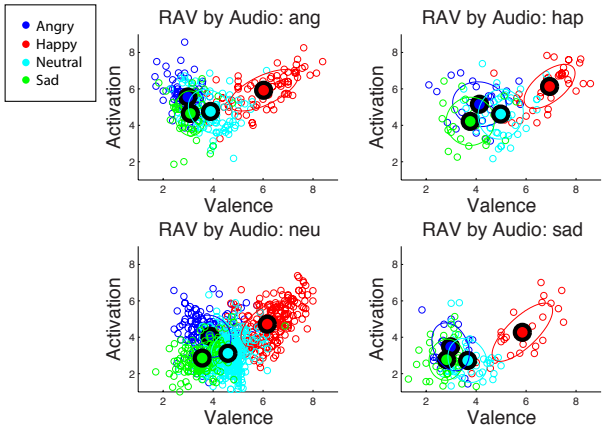

Fig. 5: The valence and activation perception of the RAV clips when grouped by emotion on the audio channel. The value of 1 represents negative (valence) and calm (activation) emotions. The value 9 represents positive (valence) and excited (activation) emotions.

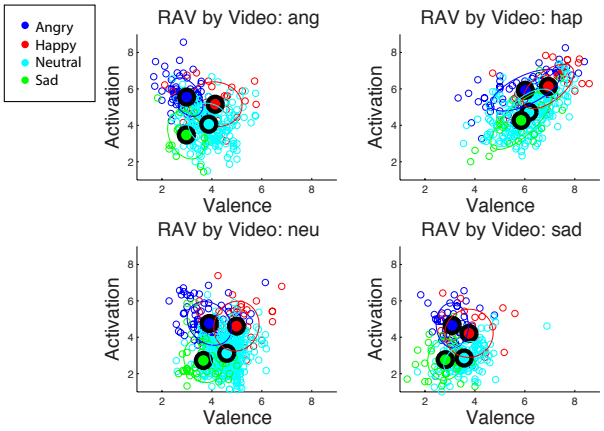

Fig. 6: The valence and activation perception of the RAV clips when grouped by emotion on the video channel. The value of 1 represents negative (valence) and calm (activation) emotions. The value 9 represents positive (valence) and excited (activation) emotions.

(points) have the audio emotion specified in the title. For example, the plot in the upper left corner depicts the audio-visual perception of clips with the audio emotion fixed as angry and video emotions of angry (blue), happy (red), neutral (cyan), and sad (green). The results demonstrate that the perception of the RAV combinations become more negative and more activated than the happy, neutral, and sad OAV counterparts (Figure 4). However, the figure also hints that the impact of the audio and video information on perception depends on the emotion present in each of the two channels. For example, the effect of happy audio (Figure 5, upper right corner) is weaker than happy video (Figure 6, upper right corner). The figures demonstrate that when happiness is on the video channel, the RAV perception is generally more positively valenced and more highly activated than the OAV perceptions (for OAV angry, neutral, and sad).

The biasing effects of the audio and video information can be seen in greater detail by observing how the means of each perceptual cluster (e.g., black centers in Figures 5 and 6) change from the perceptual cluster centers in the OA and $\widehat{OV}$ presentations. Table 7 demonstrates how OAV and RAV perception differ from the perception of the unimodal perception across the dimensions of valence, activation, and dominance. Entries in blue have matched audio and video emotion content (OAV). It is important to note that this set of stimuli does not include the RAV stimuli for which evaluators disagreed with the actor targets and noted that both the audio and video components had the same emotion label. Entries in black have mismatched emotion content (RAV). A down arrow (↓) indicates that the dimensional perception is statistically significantly lower than the unimodal

perception, while the up arrow (↑) indicates that the perception is statistically significantly higher. The equal sign (=) indicates that the means are not statistically significantly different (two-sample t-test, significance is asserted at $p < 0.01$). The results demonstrate that when anger is present in the audio, the perception of the stimuli is statistically significantly more negative than any of the unimodal $\widehat{OV}$ perceptions excepting anger and sadness, where it is statistically unchanged. Additionally, the activation and dominance values are statistically significantly higher. When happiness is present on the audio channel, the resulting perception is statistically significantly more positive compared to when any video emotion is present (excepting happy video). The activation and dominance perception is significantly higher for all combinations excepting when angry is present in the video (anger is a high activation, high dominance emotion). RAV clips with sad audio are perceived significantly more negatively in all conditions excepting when angry video is present (anger is a low valence emotion). The activation and dominance perception are either significantly decreased or statistically unchanged.

When happiness is present on the video channel, we see a consistent significant (large) increase in the perceived valence of the RAV clips as compared to the valence perception of the OA clips. The perceived activation also increases compared to the OA stimuli in all cases excepting when happiness is also present on the audio channel, although the increase is in general smaller than the increase in valence (excepting in the presence of neutral audio). This may be due to the fact that happiness is primarily differentiated from the other emotions based on valence or that video contains more valence information. The results demonstrate that when anger is present on the video channel, we see a significant decrease in valence perception for RAV clips with happy or neutral audio. When sadness is present on the video channel, we see a significant decrease in the perception of valence for clips with happiness, neutrality, or sadness on the audio channel. The trends in activation perception suggest that the combination of low activation video emotions with high activation audio emotions result in a statistically significant increase in activation perception (compared to OA stimuli), and vise versa. The trends in dominance perception suggest that the combination of a low dominance video and high dominance audio result in a statistically significant increase in dominance perception.

7 Models of Audio-Visual Perception

The results presented thus far have demonstrated that both audio and video shape the perception of emotion. This section will describe the nature of the relationship between the audio and video unimodal valence, activation, and dominance perception and the corresponding perception of multimodal stimuli. In this section, we will use all audio-visual clips, independent of categorical emotion label (the labels of 'xxx' and other are merged).

	Model 1 (M_1)	Model 2 (M_2)	Model 3 (M_3)
Indep. Var.	Audio Emotion Video Emotion		Audio Emotion Video Emotion
		Audio V/A/D Video V/A/D	Audio V/A/D Video V/A/D
Dep. Var	$<-----$ Audio-Visual V/A/D $----->$		

TABLE 8: The independent variables ("Indep. Var") and dependent variable ("Dep. Var") associated with each of the models discussed in Section 7. The dimensional variables are written as "V/A/D," indicating that valence or activation or dominance values are used.

We introduce three models to test the relationship between unimodal and multimodal perception (see the overview in Table 8). *Model 1:* In our initial experiment, we posit that audio-visual dimensional perception can be estimated given knowledge of the unimodal categorical emotion content (e.g., predict audio-visual valence given anger in the audio channel and sadness in the video channel). *Model 2:* In our second experiment, we posit that audio-visual dimensional perception can be more accurately modeled given unimodal dimensional perception (e.g., predict audio-visual valence given a video valence of 5 and an audio valence of 3). *Model 3:* Finally, in our last model, we posit that models containing both sources of information will be more correlated with reported audio-visual dimensional perception, hypothesizing that the interaction between the audio and video channels depends both on the content present in each channel and the dimensional perception associated with each channel. In all cases, we model the OAV and RAV stimuli using stepwise linear regression models trained using leave-one-subject-out cross-validation (disjoint set of training and testing speakers, 11 folds). T-tests are performed on each term (e.g., valence perception of OV) and terms are entered into the model when $p < 0.05$ and removed from the model when $p > 0.1$. All models contain interaction terms if the p-values of those interaction terms suggest statistical significance. The models are implemented using $stepwiselm$ in Matlab.

In our first experiment, we test the ability of Model 1 to predict audio-visual dimensional perception. The results demonstrate that valence can be most accurately modeled in this manner (Table 9). The adjusted R^2 associated with the valence regression models range from 0.62 for the prediction of RAV to 0.75 for OAV. Interestingly, the R^2 for the combined dataset ($OAV \cup RAV = AV$) was only 0.67. This may suggest that the patterns of audio-visual integration for OAV and RAV are different. It may also be observed because the emotional content of the RAV stimuli is evaluated less consistently than the OAV stimuli and would, consequently, be harder to model (Table 1). The R^2 for the activation and dominance are lower, and exhibit the same trends (highest R^2 for $OAV > AV > RAV$).

In our second experiment, we test Model 2. Models with high R^2 values suggest that there are consistent patterns that explain how evaluators integrate unimodal emotional cues (as captured by the unimodal

		Valence	Activation	Dominance
M_1	AV	0.67 ± 0.01	0.52 ± 0.01	0.46 ± 0.01
	OAV	0.75 ± 0.01	0.64 ± 0.01	0.58 ± 0.01
	RAV	0.62 ± 0.01	0.42 ± 0.01	0.36 ± 0.02
M_2	AV	0.80 ± 0.01	0.75 ± 0.01	0.66 ± 0.01
	OAV	0.85 ± 0.00	0.82 ± 0.00	0.76 ± 0.01
	RAV	0.77 ± 0.01	0.69 ± 0.01	0.59 ± 0.01
M_3	AV	0.82 ± 0.00	0.77 ± 0.01	0.68 ± 0.01
	OAV	0.86 ± 0.00	0.83 ± 0.00	0.77 ± 0.01
	RAV	0.79 ± 0.01	0.72 ± 0.01	0.61 ± 0.01

TABLE 9: The adjusted R^2 of the linear regression models (M) for the RAV, OAV, and AV stimuli for the dimensions of valence, activation, and dominance.

dimensional assessments). The results demonstrate that the models based on dimensional perception are more strongly correlated with RAV, OAV, and AV perception (Table 9) than the models using knowledge of categorical emotion content. This finding should be expected as the dimensional perception associated with the OA and $OV/\widehat{OV}$ clips (i.e., OV clips for the OAV stimuli or $\widehat{OV}$ for the RAV stimuli) provide a greater granularity of description than do the discrete labels of the first model.

In our final experiment, we test Model 3. The results demonstrate that the R^2 associated with the valence, activation, and dominance models all improve compared to either the VAD-only or emotion content-only models (Table 9). This suggests that the emotion class, in addition to the dimensional perception of that class, may influence how evaluators integrate audio-visual information.

The results of the three modeling experiments demonstrate that there are consistent patterns underlying the emotion perception of our evaluator population. We highlight Model 2, which predicts multimodal perception from unimodal perception (e.g., how RAV activation perception can be predicted from unimodal OA and $OV/\widehat{OV}$ activation perception). We investigate the learned model to understand the implications to audio-visual emotion perception. In our investigation, the absence of an interaction term would suggest that multimodal perception could be explained as merely additive, while the presence of an interaction term would provide evidence for cross-modal perceptual integration. We hypothesize that the OAV stimuli, which contain redundant information across the modalities, can be explained using a simple additive model absent interaction terms, while the RAV stimuli, which contain supplementary information due to the emotional mismatch, will be explained by models with interaction terms. We found that over the 11 cross-validation folds, there were no interaction terms selected in the OAV models for any of the dimensions. This suggests that our stimuli with emotionally matched content do not highlight the underlying interaction between the audio and video modalities. However, in all 11 folds, the RAV activation and dominance models contain interaction terms, suggesting that the RAV stimuli highlight cross-modal interaction patterns. It is important to note that the interaction term, while statistically significant at $p << 0.001$ in all cases, is very small (average value of -0.064 ± 0.0045 for activation and -0.050 ± 0.0086 for dominance). The interaction term was selected in only one of the 11 cross-validation folds in the valence model.

Training	Testing	Valence	Activation	Dominance
OAV	OAV	0.31 ± 0.09	0.50 ± 0.09	0.54 ± 0.13
RAV	OAV	0.31 ± 0.09	0.56 ± 0.15	0.58 ± 0.11
OAV	RAV	0.34 ± 0.06	0.53 ± 0.10	0.64 ± 0.19
RAV	RAV	0.34 ± 0.06	0.51 ± 0.09	0.61 ± 0.18

TABLE 10: The mean square error of OAV and RAV valence, activation, and dominance perception prediction given either OAV or RAV training data. The model is built using unimodal dimensional perception.

The presence of only a small interaction term suggests that models of OAV VAD perception may be used to estimate RAV VAD perception and vise versa. If this can be accomplished without incurring large errors it would suggest similar integration processes. This in turn would support the use of RAV stimuli in perception experiments. We assess this question using two different types of training: training and testing on the same data (e.g., train and test on OAV, "agree") or training and testing on different data (e.g., train on OAV and test on RAV, "disagree") using Model M_2 (Table 8). The results demonstrate that the mean square error associated with the agree and disagree training/testing paradigms are similar (Table 10). We run an ANOVA analysis to understand the main effects of dimension and training/testing paradigm and the associated interaction effect. ANOVA demonstrates a significant effect of dimension (F(2, 120) = 61.14, $p \approx 0$) and a non-significant main effect for training paradigm and a non-significant interaction term (F(3, 120) = 1.12, $p = 0.3431$ and F(6, 120) = 0.48, $p = 0.8257$, respectively). This suggests that not only do there exist patterns describing how unimodal perception is linked to multimodal perception (low error for training and testing on the same data, Table 10), but that the patterns through which unimodal and multimodal perception are integrated are similar for RAV and OAV stimuli. A multiple comparisons test demonstrates that the error associated with valence prediction is statistically significantly lower than that of either activation prediction or dominance prediction. This suggests that valence rating can be more accurately estimated given unimodal information than either activation or dominance ratings.

8 Discussion

Human emotion perception is inherently multimodal; when making assessments of emotion, human evaluators attune to emotional cues presented over multiple channels. These cues contain different information and evaluators must integrate this information to arrive at a cohesive narrative of emotional content even when the information on the two channels would not occur naturally in the environment. In this work, we investigated the interplay between audio and video emotion

cues to understand how such cues are integrated during the emotion perception process. We found that there are patterns underlying the integration of the audio and video cues and that these patterns can even be used to explain perception patterns for stimuli that are not naturally occurring.

Conventionally, OAV-style stimuli are used for emotion perception experiments (e.g., [70], [71]). These data are incredibly important for emotion classification experiments as they provide clear audio-visual emotion content, critical for basic research in emotion expression. However, the challenge with such stimuli is that the cross-modal redundancy may obfuscate the complex interaction patterns that are a part of the emotion perception process. Our results demonstrate that complex interaction between modalities (beyond simple additive interaction) did not occur in OAV stimuli and occurred only in the activation and dominance perception associated with RAV stimuli. This result suggests that emotionally "noisy" stimuli offer an opportunity to gain insight into emotion perception that is not available when considering emotionally clear stimuli.

The RAV stimuli creation is motivated by the McGurk effect phenomenon, in which the presentation of conflicting audio-visual phonetic information results in the perception of a third distinct phoneme. The results in this paper have not found support for the literal existence of such a phenomenon for emotion. It is not clear that the presentation of two distinct categorical emotions over the facial and vocal channel result in a third distinct emotion. However, the results do show that the dimensional perception of emotionally mismatched audio-visual emotions is different from either the unimodal or emotionally matched stimuli. This suggests that the emotional McGurk effect phenomenon may be better described in terms of its effects on dimensional, rather than categorical, perception.

The dimensional evaluations of the RAV stimuli provide evidence for the complexity of the stimuli. The standard deviations of the evaluations associated with the RAV stimuli (Table 1) are higher than those of the OAV stimuli across the three dimensions. This may suggest that the RAV stimuli were more challenging to evaluate than the OAV stimuli. This may occur because no modality dominated the perception of the evaluators and, as a result, the evaluations were more varied. The varied impact of the modalities can be observed in terms of how evaluators assigned discrete labels to the RAV stimuli. For example, the presentation of anger across either the video or the audio modality only biased the perception of the evaluators to the class of anger in approximately 48% and 40% of the stimuli, respectively (Figure 3), suggesting that neither auditory nor visual displays of anger consistently biased perception. We hypothesize that in clips where neither of the modalities dominates the emotion perception of the evaluators, the resulting audio-visual perception may correspond to a different emotion, with associated differences in the individual activation, valence and dominance values (Table 7), which would highlight the potential for an emotional McGurk-type effect.

We modeled the hypothesized dimensional McGurk effect using regression models that fit and predicted the link between unimodal (OA and OV) perception and multimodal dimensional perception. The results demonstrated that we could make this prediction with low error. This suggests that there are consistent patterns that describe how individuals are integrating emotional information presented on either the audio or the video channels. Further, we demonstrated that we could estimate RAV perception using models trained on OAV perception and vise versa without incurring large error penalties. This result suggests that we can use RAV stimuli, in addition to the more common OAV stimuli, to study human emotion perception because the link between unimodal and multimodal perception follows similar trends. This is important because it suggests that we can use RAV stimuli to better understand audio-visual feature reliance, investigations that are more challenging in OAV stimuli given the inherent cross-modal correlations. In fact, the RAV stimuli highlighted a small interaction between unimodal perception (OA and $\widehat{OV}$) and multimodal perception that was not seen in the OAV stimuli, again highlighting the potential of a dimensional McGurk-type effect. Our future research will further explore the nature of this emotional McGurk-type effect observed on dimensional descriptors.

The similarity of the results of the matched and mismatched modeling scenarios suggests that individuals integrate the audio-visual affective messages in a similar manner, independent of emotional congruence. However, this does not necessarily suggest that individuals are also using the multimodal cues in the same manner in the two conditions. Prior work has suggested that listeners focus their attention on salient cues [72], [73]. The results suggest that an emotional attribution of happiness is strongly affected by the presence of happy video information. This suggests that cues such as smiles may be salient and may provide an explanation for why happy video information overpowers affective cues on other modalities. Additional research is needed to understand the causal link between features hypothesized as salient and resulting emotion perception, particularly research in the prediction of perception from multimodal cues and the analysis by synthesis methodologies.

One of the potential critiques associated with the stimuli creation used in this paper relates to our method for audio-visual time synchronization, which we achieve via video warping. The challenge is that timing affects emotion perception. For example, if we consider neutrality as having moderate vocalic durations, anger can be characterized by relatively rapid vocalic duration and sadness by relatively slow vocalic duration. Consequently, our method of warping the video by adding and dropping video frames has the potential to alter the perception of the emotional content associated with the

video because it changes the durational information. We observe that for emotionally clear OV information, the warping process does not have a strong effect on categorical perception; after warping, 91.61% of the emotionally clear OV stimuli originally perceived as anger, happiness, neutrality, or sadness, are perceived as having the same emotion after warping (for $\widehat{OV}$ stimuli also rated with a primary label from this subset of emotions). This contrasts with 83.08% of emotionally subtle OV stimuli, suggesting that the effect of video warping depends on the level of subtly present in the original display. Our ANOVA analysis demonstrates that warping does not significantly affect dimensional VAD perception associated with the categorical emotions. It is important to note that even given affective change introduced by warping, it is still possible to investigate multimodal perceptual integration by leveraging the known categorical and dimensional evaluations associated with the OV and $\widehat{OV}$ stimuli. It is this focus on ratings that allows for a detailed understanding of perception change.

9 Conclusions

This paper presented the UMEME dataset, a resource for studying emotion perception. The results demonstrate that there are consistent patterns underlying audio-visual emotion perception, even given emotionally mismatched stimuli. This suggests that this style of stimuli can provide insight into emotion perception and provide tools to better understand how audio and video information are integrated during perception.

The patterns that underlie emotion perception are described at a high-level in this paper as we investigated the interaction patterns between unimodal and multimodal perception. However, there remain many open questions relating how the features embedded within the audio and visual information interact cross-modally and over time. Our future work will investigate this audio-visual modeling to uncover the audio-visual feature cues that affect multimodal emotion perception. Finally, these stimuli represent a new opportunity to understand perception patterns not only in those who have "healthy" emotion perception, but also in those whose emotion perception processes are disordered. We anticipate that the stimuli will be of interest to researchers studying systematic deviations in emotional perception by individuals with disordered perception. These studies can leverage the comprehensive perceptual evaluation of the congruent and conflicting audio-visual stimuli. Furthermore, the corpus can be used as an instrumental tool to assess salient audio-visual features, and their role in perception. We expect that the corpus will be of interest to researchers seeking to understand how emotion modulates speech and facial expressions at the phoneme level, since the database provides spontaneous renditions of the same sentences under different emotional contexts. In our future work, we will broaden the evaluator pool to explore how emotionally mismatched stimuli can lead to new insight into the differences in emotion perception.

This dataset will be released to the community due to its potential to impact the field of affective computing.

References

[1] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, April 2003.

[2] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, September 2006.

[3] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, November 2009.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, January 2001.

[5] D. Keltner and J. Haidt, "Social functions of emotions," in *Emotions: Currrent issues and future directions. Emotions and social behavior*, T. J. Mayne and G. A. Bonanno, Eds. Guilford Press, New York, NY, 2001.

[6] J. A. Deonna, "Emotion, perception and perspective," *Dialectica*, vol. 60, no. 1, pp. 29–46, 2006.

[7] C. G. Kohler, J. B. Walker, E. A. Martin, K. M. Healey, and P. J. Moberg, "Facial emotion perception in schizophrenia: a meta-analytic review," *Schizophrenia Bulletin*, vol. 36, no. 5, pp. 1009–1019, 2010.

[8] C. G. Kohler, L. J. Hoffman, L. B. Eastman, K. Healey, and P. J. Moberg, "Facial emotion perception in depression and bipolar disorder: a quantitative review," *Psychiatry research*, vol. 188, no. 3, pp. 303–309, 2011.

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

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

[11] 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 attributions," *Social Cognitive and Affective Neuroscience*, vol. 1, no. 2, pp. 95–106, August 2006.

[12] H. K. M. Meeren, C.C.R.J.V. 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. 16518–16523, 2005.

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

[14] 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.

[15] B. de Gelder, K. B. E. Bocker, 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, January 1999.

[16] 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.

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

[18] 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.

[19] 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, Apr. 2008, pp. 2201–2204.

[20] Emily Mower Provost, Irene Zhu, and Shrikanth Narayanan, "Using emotional noise to uncloud audio-visual emotion perceptual evaluation," in *International Conference on Multimedia and Expo (ICME)*, San Jose, CA, July 2013.

[21] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. To appear, 2015.

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

[23] J. R. Davitz, *The language of emotion*, Academic Press, 1969.

[24] T. Engen, N. Levy, and H. Schlosberg, "The dimensional analysis of a new series of facial expressions," *Journal of Experimental Psychology*, vol. 55, no. 5, pp. 454–458, May 1958.

[25] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[26] R. S. Lazarus, J. R. Averill, and E. M. Opton Jr, "Towards a cognitive theory of emotion," in *Feeling and emotion: The Loyola Symposium*, 1970, pp. 207–232.

[27] G. Mandler, *Mind and Emotion*, Wiley, 1975.

[28] S. Schacter and J. Singer, "Cognitive, social, and physiological determinants of emotional state," *Psychological Review*, vol. 69, no. 5, pp. 379–399, September 1962.

[29] P. Ekman, "Basic emotions," *Handbook of cognition and emotion*, pp. 45–60, 1999.

[30] J. M. Zelenski and R. J. Larsen, "The distribution of basic emotions in everyday life: A state and trait perspective from experience sampling data," *Journal of Research in Personality*, vol. 34, no. 2, pp. 178–197, 2000.

[31] G. A Calvert, M. J Brammer, and S. D Iversen, "Crossmodal identification," *Trends in cognitive sciences*, vol. 2, no. 7, pp. 247–253, 1998.

[32] R. Campbell, "The processing of audio-visual speech: empirical and neural bases," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1493, pp. 1001–1010, 2008.

[33] L. M Miller and M. D'esposito, "Perceptual fusion and stimulus coincidence in the cross-modal integration of speech," *The Journal of neuroscience*, vol. 25, no. 25, pp. 5884–5893, 2005.

[34] D. W. Massaro and P. B. Egan, "Perceiving affect from the voice and the face," *Psychonomic Bulletin & Review*, vol. 3, no. 2, pp. 215–221, 1996.

[35] R. J. Dolan, J. S. Morris, and B. de Gelder, "Crossmodal binding of fear in voice and face," *Proceedings of the National Academy of Sciences*, vol. 98, no. 17, pp. 10006–10010, 2001.

[36] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, 2014.

[37] M. Swerts, K. Leuverink, M. Munnik, and V. Nijveld, "Audiovisual correlates of basic emotions in blind and sighted people.," in *Interspeech*, Portland, OR, September 2012.

[38] J. H. G. Williams, D. W. Massaro, N. J. Peel, A. Bosseler, and T. Suddendorf, "Visual–auditory integration during speech imitation in autism," *Research in developmental disabilities*, vol. 25, no. 6, pp. 559–575, 2004.

[39] E. G. Smith and L. Bennetto, "Audiovisual speech integration and lipreading in autism," *Journal of Child Psychology and Psychiatry*, vol. 48, no. 8, pp. 813–821, 2007.

[40] O. Collignon, S. Girard, F. Gosselin, S. Roy, D. Saint-Amour, M. Lassonde, and F. Lepore, "Audio-visual integration of emotion expression," *Brain research*, vol. 1242, pp. 126–135, 2008.

[41] P. Barkhuysen, E. Krahmer, and M. Swerts, "Crossmodal and incremental perception of audiovisual cues to emotional speech," *Language and speech*, vol. 53, no. 1, pp. 3–30, 2010.

[42] O. Doehrmann and M. J. Naumer, "Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration," *Brain research*, vol. 1242, pp. 136–150, 2008.

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

[44] 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.

[45] 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.

[46] 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. 16518–16523, 2005.

[47] J. Cassell, D. McNeill, and K.-E. McCullough, "Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information," *Pragmatics & cognition*, vol. 7, no. 1, pp. 1–34, 1999.

[48] R. J Compton, M. T. Banich, A. Mohanty, M. P. Milham, J. Herrington, G. A. Miller, P. E. Scalf, A. Webb, and W. Heller, "Paying attention to emotion," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 3, no. 2, pp. 81–96, 2003.

[49] A. Katsamanis, M. Blackand P. Georgiouand 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.

[50] J. Fiscus, "Sclite scoring package version 1.5," *US National Institute of Standard Technology (NIST), URL http://www. itl. nist. gov/iaui/894.01/tools*, 1998.

[51] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks," in *Conference on empirical methods in natural language processing (EMNLP 2008)*, Honolulu, HI, USA, October 2008, pp. 254–263.

[52] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014.

[53] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle," in *Association for Computational Linguistics (ACL 2012) System Demonstrations*, Jeju Island, Republic of Korea, July 2012, pp. 26–34.

[54] A. Gravano, R. Levitan, L. Willson, S. Beòuš, J. B. Hirschberg, and A. Nenkova, "Acoustic and prosodic correlates of social behavior," in *Interspeech*, Florence, Italy, 2011.

[55] 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*, March 2010, pp. 5270–5273.

[56] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2010, HLT '10, pp. 207–215, Association for Computational Linguistics.

[57] J. Heer and M. Bostock, "Crowdsourcing graphical perception: using mechanical turk to assess visualization design," in *Proceedings of the 28th international conference on Human factors in computing systems*, New York, NY, USA, April 2010, CHI '10, pp. 203–212, ACM.

[58] S. Mohammad and P. Turney, "Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, CA, June 2010, pp. 26–34, Association for Computational Linguistics.

[59] J. P. Bigham, C. Jayant, A. Miller, B. White, and T. Yeh, "Vizwiz::locateit - enabling blind people to locate objects in their environment," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, June 2010, pp. 65–72.

[60] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the crowdworkers?: shifting demographics in mechanical turk," in *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, New York, NY, USA, April 2010, CHI EA '10, pp. 2863–2872, ACM.

[61] 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.

[62] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *IEEE International Conference on Multimedia and Expo (ICME 2008)*, Hannover, Germany, June 2008, pp. 865–868.

[63] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Lang. Resources and Evaluation*, pp. 335–359, Nov. 5 2008.
[64] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," pp. 381–385, December 2005.
[65] C. Busso and S. S. Narayanan, "The expression and perception of emotions: Comparing assessments of self versus others," in *Proceedings of InterSpeech*, Brisbane, Australia, Sept. 2008, pp. 257–260.
[66] 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.
[67] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 1179–1182.
[68] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1, pp. 227–256, 2003.
[69] M. Bulut and S. Narayanan, "On the robustness of overall F0-only modification effects to the perception of emotions in speech," *Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4547–4558, June 2008.
[70] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech.," in *Interspeech*, 2005, vol. 5, pp. 1517–1520.
[71] J. Kim, S. Lee, and S. Narayanan, "An exploratory study of the relations between perceived emotion strength and articulatory kinematics.," in *INTERSPEECH*, 2011, pp. 2961–2964.
[72] Patrik Vuilleumier, "How brains beware: neural mechanisms of emotional attention," *Trends in cognitive sciences*, vol. 9, no. 12, pp. 585–594, 2005.
[73] Mark J Fenske and Jane E Raymond, "Affective influences of selective attention," *Current Directions in Psychological Science*, vol. 15, no. 6, pp. 312–316, 2006.

Emily Mower Provost (S'07-M'11) is an Assistant Professor in Computer Science and Engineering at the University of Michigan. She received her B.S. in Electrical Engineering (summa cum laude and with thesis honors) from Tufts University, Boston, MA in 2004 and her M.S. and Ph.D. in Electrical Engineering from the University of Southern California (USC), Los Angeles, CA in 2007 and 2010, respectively. She is a member of Tau-Beta-Pi, Eta-Kappa-Nu, and a member of IEEE and ISCA. She has been awarded the National Science Foundation Graduate Research Fellowship (2004-2007), the Herbert Kunzel Engineering Fellowship from USC (2007-2008, 2010-2011), the Intel Research Fellowship (2008-2010), and the Achievement Rewards For College Scientists (ARCS) Award (2009 2010). She is a co-author on the paper, "Say Cheese vs. Smile: Reducing Speech-Related Variability for Facial Emotion Recognition," winner of Best Student Paper at ACM Multimedia, 2014 (with Y. Kim). She is also a co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge (with C. Busso). Her research interests are in human-centered speech and video processing, multimodal interfaces design, and speech-based assistive technology. The goals of her research are motivated by the complexities of human emotion generation and perception.

Yuan Shangguan (S'13) is a PhD student in Computer Science and Engineering at University of Michigan. She received her B.A.S. in Engineering Sciences from Dartmouth College and B.E. in Computer Engineering from Thayer School of Engineering at Dartmouth (magna cum laude) in 2013 . She is a member of Tau-Beta-Pi, Gamma Sigma Alpha, and Eta-Kappa-Nu. She has been awarded Philip E. Lippincott 1957 Scholarship (2009-2013) for her B.A.S in Dartmouth College, James O. Freedman Presidential Scholar research grant and Paul K. Richter And Evalyn E. Cook Richter Memorial Fund for her project on computational opinion change (2011-2012). She was selected for Singapore Agency for Science, Technology and Research (A*STAR) industrial attachment with Institute for Infocomm Research (I2R) to develop more natural human-robotic dialogues in 2012. Her research interests are in analyzing and extracting dynamic patterns from human-centered visual-audio interactions, audio-visual integration in human-computer interactions, and the extension of multimodal affective perception into gestures and artistic expressions.

Carlos Busso (S'02-M'09-SM'13) is an Assistant Professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He received his B.S (2000) and M.S (2003) degrees with high honors in electrical engineering from University of Chile, Santiago, Chile, and his Ph.D. (2008) in electrical engineering from University of Southern California (USC), Los Angeles, USA. He was selected by the School of Engineering of Chile as the best Electrical Engineer graduated in 2003 across Chilean universities. At USC, he received a Provost Doctoral Fellowship from 2003 to 2005 and a Fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. In 2014, he received the ICMI Ten-Year Technical Impact Award. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interests are in digital signal processing, speech and video processing, and multimodal interfaces. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, modeling and synthesis of verbal and nonverbal behaviors, sensing human interaction, in-vehicle active safety system, and machine learning methods for multimodal processing.