Dimensional perception of a ‘smiling McGurk effect’

Ilaria Torre*, Simon Holk*, Emma Carrigan†, Iolanda Leite* Rachel McDonnell† and Naomi Harte†
* KTH Royal Institute of Technology, Stockholm, Sweden
Email: {ilariat, sholk, iolanda}@kth.se
† Trinity College Dublin, Dublin, Ireland
Email: {carrige, ramcdonn, nharte}@tcd.ie

Abstract—**Multisensory integration influences emotional perception, as the McGurk effect demonstrates for the communication between humans. Human physiology implicitly links the production of visual features with other modes like the audio channel: Face muscles responsible for a smiling face also stretch the vocal cords that results in a characteristic smiling voice. For artificial agents capable of multimodal expression, this linkage is modeled explicitly. In our study, we observe the influence of visual and audio channel on the perception of the agent's emotional state. We created two virtual characters to control for anthropomorphic appearance. We record videos of these agents either with matching or mismatching emotional expression in the audio and visual channel. In an online study we measured the agent's perceived valence and arousal. Our results show that a matched smiling voice and smiling face increase both dimensions of the Circumplex model of emotions: ratings of valence and arousal grow. When the channels present conflicting information, any type of smiling results in higher arousal rating, but only the visual channel increases the perceived valence. When engineers are constrained in their design choices, we suggest they should give precedence to convey the artificial agent's emotional state through the visual channel.**

Index Terms—**Multisensory integration, Smiling, Virtual Agent, Human-Likeness**

I. Introduction

Correctly interpreting the interlocutor's emotions during a conversation contributes to successful social interactions. In an increasingly technological world, it is of paramount importance to facilitate human-machine cooperation and communication. Recent evidence shows that emotional displays in robots and virtual agents lead to the agents being perceived more positively, and improve collaboration with humans [1, 2]. Yet these emotional displays are more often than not coming from only one modality, whereas emotional expressions in humans are generally multimodal [3]. Humans are generally skilled at decoding information from visual (e.g. facial expressions) and auditory (e.g. the tone of voice) cues, and they are able to react appropriately to emotional displays [4]. When information is transmitted from more than one channel simultaneously – e.g., when we see as well as hear someone at the same time – this information is generally congruent: if we are angry and want to show it, all the available channels will transmit the information that we are angry (e.g. angry facial expression and angry tone of voice) [1]. However, when artificial agents display an emotion multimodally, the auditory and visual channels often present a certain degree of incongruency. There is no unified biological entity producing the information, but rather the information is created by two different systems – for example, a software managing the facial expressions and a Text-to-Speech system generating the voice. This can result in different degrees of expressiveness being broadcast by the two channels. Therefore, as observed by [6], when we study people's perceptions of multimodal expressive artificial agents, we cannot rule out the possibility that their perception will be affected by this incongruency.

In this paper, we study people's perceptions of an artificial agent that presents mismatched facial and vocal emotional expressions, and compare them to the baseline matched expressions. These expressions were generated from a human actor, thus ensuring ecological validity: we used motion capture to drive the agent's facial expressions, and audio recordings for its vocal expressions. We found that incongruency did not impair emotional processing, and we discuss the implication of these findings for Human-Agent Interaction.

II. Related Work

It is important to make a distinction between the terms ‘emotions’ and ‘emotional expressions’: while the former indicate biological states that can result in subjective feelings, actions, and intentions [7], the latter indicate how someone is displaying an emotion, for example by making certain facial expressions or speaking with affective prosody [8]. This implies that emotional expressions do not necessarily reflect a ‘real’ emotion, but they can be deliberate, at least to a certain extent [9]. Our experiment deals with emotional expressions, but in this brief overview of the literature we will use both terms as appropriate.

Emotions can be expressed in the voice, face and body [10, 11, 12]. These channels work together to form a unitary

The project was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713567, and by the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2016) and is co-funded by the European Regional Development Fund. The first author is funded by Wallenberg AI, Autonomous Systems and Software Program.

[1] Note however that, contrary to a common misconception, humans can naturally produce incongruent emotional stimuli as well, for example to produce a complex emotional display (e.g. sarcasm), or if one channel has been impaired (e.g. by a stroke) [5].

978-1-6654-0019-0/21/$31.00 ©2021 IEEE

emotional expression, in a so-called multisensory integration [13]. Scans of the brain have shown the existence of dedicated regions associated with the handling of multimodal input [14]. The multisensory integration phenomenon has been extensively used to study which channel is most salient in emotion perception, by presenting participants with mismatched auditory and visual stimuli – for example, overlapping the picture of a person showing an angry facial expression and the audio of a happy-sounding speaker [15]. The effect of this mismatch on a listener is related to a well-known perceptual phenomenon, the McGurk Effect [16]. From these studies on mismatched emotional expressions, it has emerged that, in human-human interaction, the face is more salient for emotion processing. Combining photographs and audio clips of people expressing different emotions resulted in higher recognition accuracy when face and voice matched, and in visual channel predominance when they did not match [15].

Studying mismatched emotional expressions with human stimuli is interesting from a theoretical point of view. For example, from a cognitive science and signal processing perspective, it is interesting to know if one channel (audio or video) is more predominant in the transmission and decoding of a certain information. However, most of the time our emotional expressions are congruent. For example, when we smile, driving the lips upwards modifies the vocal tract, and as a result the speech uttered while smiling will present a corresponding smiling sound (more on this below). On the other hand, many emotional expressions displayed by artificial agents are incongruent. For instance, many widely used social robots, which happen to have a fixed facial expression, have their voice modulated independently of this expression. Thus, for the field of Human-Machine Interaction, studying these mismatched stimuli is also interesting from a practical point of view: in such cases where one expression channel is fixed, we need to know how people will interpret the agent's incongruent signals. Furthermore, even in cases where both channels are changeable (e.g. robots like Sophia, Furhat, or Emys, whose facial expressions are not fixed), it might be difficult to obtain the same degree of expressivity from both channels, thus creating a signal that is still not perfectly congruent. Also, expressivity for TTS voices is still a far from resolved issue [17].

To date, research on this topic within the field of Human-Agent Interaction is scarce. In a study with a female-appearing virtual agent, participants watched a virtual avatar displaying emotions in the voice and face and then rated their valence and arousal, showing that both types of expression contributed information [5]. Ratings showed bias towards the audio stimuli. However, the computer graphics technologies used in these studies are now outdated; as mentioned in [5], it is possible that it was the incongruency between expressive voices (recorded from a human actor) and not-so-expressive faces (a stylised avatar) that resulted in vocal predominance. In another study featuring a different female avatar, [18] found that matched emotional expressions were easier to recognise than mismatched ones (e.g. happy face and concerned voice). On the other hand, the mismatched expressions were not clearly perceived as either of the original expressions in the video or audio channel, resulting in a 'confused' perception of the avatar's expressions. Similarly, in a more recent study in Human-Robot Interaction, [19] found that emotion recognition plummeted with mismatched cues, and that people categorised these mismatched expressions as neither of the original, intended emotions. Results from [18, 19] suggest that mismatching audio and video emotional information might impair emotion processing, similarly to how the McGurk effect impairs phonological processing. However, as acknowledged by [19], the chosen robotic platform had a static facial expression that likely had a different degree of expressiveness from their chosen synthetic voice.

The scarcity of existing research on the topic makes it difficult to draw conclusions on how incongruent multimodal emotional expressions are perceived in Human-Agent Interaction. An added source of confusion is that the few existing studies differ in terms of how they asked participants to recognise the emotion or emotional expression, reflecting a well established divide in emotion research, namely discrete vs. dimensional theories of emotion [20]. Both major theories have found some support in neuroscience, but also faced challenges, since neither can be solely mapped into a single region in the brain [21]. Additionally, most of these studies have a fundamental methodological problem, which is, that they ask people to rate the emotions with labels ('happy', 'angry', but also 'dominant', 'aroused', etc) [22]. One method that allows to account for this issue is the Self-Assessment Manikins (SAM) [23], a series of pictorial scales to measure valence, arousal and dominance (Fig. 2). This scale has been used in Human-Machine Interaction in the past [5].

With this paper, we aim to bring new evidence into how mismatched emotional expressions are perceived in virtual agents. We chose to use the SAM scale as our evaluation tool, to remove the risk of label-bias in our data.

A. *Smiling*

In our study, we focus on smiles. Smiling is a universal [24], multimodal [25] emotional expression, which is visible in the face and audible from the voice [26]. Visually, a smile is usually displayed by activating the Zygomaticus Major muscle (lip corner puller) and, in some cases, the Orbicularis Oculi muscle (cheek raiser, or 'crow feet' around the eyes). Acoustically, smiling affects the vocal tract by shortening it and as a result, vocal frequencies tend to increase [25]. However, so far there is no consensus regarding what exactly constitutes the acoustics of a smile [27]. Previous studies using synthetic and natural speech samples, both acted and naturally-occurring, have shown that vocal fundamental frequency (F0), the first three formants (F1, F2, F3) and the spectral centroid tend to increase while smiling [27, 2, 26]. A recent summary on the findings on smiling acoustics can be found in [2]. Perceptually, both visual features (facial expressions) and auditory features (speech acoustics) contribute to the perception of 'smilingness' [26]. Smiling is not linked to only one emotion: there are sad smiles, convenience smiles, Duchenne smiles, mocking smiles, and many others [e.g. 9, 28].

Thus, an interesting aspect to consider is that smiles are appraised based on the context in which they are expressed. This could be situational context – e.g., smiling after beating you in a game has a different meaning than smiling after cooperatively winning a game together [4] – but also interpersonal context – for example, attractiveness can mediate whether a smile is considered 'cute' or 'creepy' [29]. For artificial agents, appearance features, such as human-likeness and rendering style, have been shown to influence how an emotion is perceived [30]. Specifically, happy, sad and surprise expressions were perceived with lower intensity as the realism of the face increased. This might be because more stylised agents can resemble cartoons, which are generally meant to exaggerate expressions [30]. In another study, [31] created 11 render styles, from very sketched to very realistic, and found that people perceived the most realistic and most cartoon styles to be the most appealing and trustworthy (among other traits), with a drop for the characters that were halfway in the realism continuum. This might be because people are not familiar with these middle styles. Thus, in our study we also manipulated agent appearance – realistic or cartoon – to see if perceived valence and arousal are mediated by agent style.

B. Research questions

With our study, we set to answer the following research questions:

- RQ1-a: Does smiling in the voice and / or face influence arousal and valence ratings?

This is our starting question, that will inform us whether smiling (in any of the two channels) influences emotion processing. From previous literature, we can expect that smiling in the face and / or voice will increase valence and arousal ratings. To answer this question, we created a set of multimodal and unimodal stimuli, showing a virtual agent that was either displaying a matching smiling expression – smiling both in the face and in the voice – or a mismatching one – smiling either only in the face, or only in the voice. As baseline, we also created a version of the agent that was showing a matching neutral expression. The unimodal stimuli featured only the agent's face (smiling or neutral) or only the agent's voice (smiling or neutral).

- RQ1-b: Is this perception mediated by virtual agent style (human-like / cartoon-like)?

Agent-related features could influence emotion perception. For this reason, we created two virtual agents, which were animated by the same motion capture and voice, differing only in human-likeness (human-like and cartoon-like). From previous literature, we would expect the cartoon-like agent to be perceived as more expressive overall, resulting in higher ratings of valence and arousal.

- RQ2: Do people perceive congruent and incongruent emotional expressions differently, in terms of valence and arousal, in virtual agents?

TABLE I
LIST OF EMOTIONAL EXPRESSION CONDITIONS

Code	Voice (V)	Face (F)
V_sF_s	smiling	smiling
V_sF_n	smiling	neutral
V_nF_n	neutral	neutral
V_nF_s	neutral	smiling
V_s	smiling	//
V_n	neutral	//
F_s	//	smiling
F_n	//	neutral

If incongruent emotion information affects emotion perception, we would expect the ratings of valence and arousal to be different in the incongruent and congruent conditions. If this was the case, we would expect the smiling voice + neutral face and neutral voice + smiling face condition to be rated differently than either the smiling voice + smiling face or the neutral voice + neutral face conditions (see Table I).

- RQ3: Is one channel (audio or video) more predominant in emotion processing in virtual agents?

If, instead, incongruent emotional expressions are not perceived differently than congruent ones, it might be that, in incongruent stimuli, the information conveyed by the predominant channel would override the information conveyed by the non-predominant channel. Previous studies using avatars found that vocal information overrode facial information, however this might have been due to higher realism in the avatar's voice [5]. We used motion capture-driven, state-of-the-art virtual agents to rule out any potential expressiveness confounds.

Our study extends previous literature in several ways: we focus on emotional expressions, and not emotions (differently from e.g. [5]); we use a more modern virtual character than in [5]; we use a graphical dimensional scale to obtain more ecologically valid data, that is free from semantic bias (differently from e.g. [15, 19]); finally, to the best of our knowledge, there are no studies comparing the effect of emotional expression mismatch in different virtual agents.

III. METHOD

The experiment was conducted on Amazon Mechanical Turk (AMT). Participants were presented with a series of clips featuring either matched (smiling face + smiling voice, or neutral face + neutral voice) or mismatched (smiling face + neutral voice, or neutral face + smiling voice) emotional expressions (see Table I). Participants were also shown clips featuring only the video (neutral or smiling face, with subtitles) or only the audio (neutral or smiling voice, played over a black box). The clips (with the exception of the audio-only clips) featured two animated characters, one in a photorealistic, human-like style, and one in a cartoon style (Fig. 1). The characters were generated using motion capture, as described below (Section III-A).

A. Motion Capture

The virtual agents were created using state-of-the-art computer graphics technology for modelling, animating, and ren-

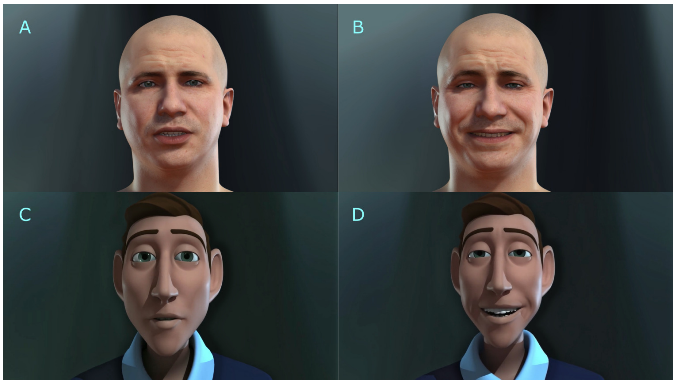

Fig. 1. Top row: photo-realistic virtual character in neutral (A) and smiling (B) facial expression. Bottom row: cartoon virtual character in neutral (C) and smiling (D) facial expression.

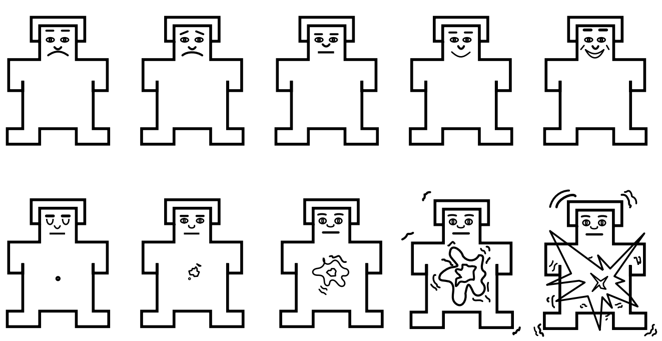

Fig. 2. The SAM (Self-Assessment Manikin) visual scale used in the study: valence (top row) and arousal (bottom row).

dering. The human-like agent, comprising over 250 scans of a real actor's facial expressions, was created by the company 3Lateral. These scans were then carefully combined into a controllable facial rig, which could then be driven by the motion capture. The cartoon-like agent was a free high-end artist rig created by Artella. We then hired a male actor to be recorded in the motion capture studio at our institution. We used a 23-camera Vicon Vantage optical motion capture system for body motion capture and a Technoprops video-based head-mounted facial capture system. The actor was asked to read a set of pre-scripted sentences in neutral and smiling expressions. These pre-scripted sentences were created as part of another experiment, where participants played the lunar and desert survival task with virtual agents [32]. The sentences were all descriptions of items to be ranked in these survival tasks. These sentences had previously been validated in terms of linguistic valence and understandability. For the current experiment, we only used sentences that were validated as neutral in valence, and as understandable by people who self-reported to be at least fluent in English [33]. The sentences can be found in [33]. Audio was recorded using a wireless microphone attached to the actor's face. The actor's facial movements were then retargeted onto the models, using Faceware Tech software for the facial movement and inverse kinematics for the movement of the head. Finally, advanced shaders (e.g., subsurface scattering for the skin) were used to create the highly realistic appearance in Autodesk Maya 2018 software. The final characters are shown in Figure 1. Since features such as key light brightness and key-to-fill-ratio were shown to have an effect on ratings of emotional intensity and appeal [34], we kept illumination parameters identical in the two characters.

B. Video file preparation

The audio recordings from the actor were processed using Audacity. First, a noise removal filter was applied to the recordings (with parameters: noise reduction 24dB, sensitivity 0dB, frequency smoothing 150 Hz, attack/decay 0.15 seconds); then the full audio file was segmented to obtain one file per utterance; finally, the individual sound files were amplitude-normalized. These files were then lip-synced to the individual, corresponding video files using Lightworks. Neutral sound files were lip-synced to neutral and smiling video files, and smiling sound files were lip-synced to smiling and neutral video files, to obtain the 4 desired multimodal conditions: $\mathrm{V}_s\mathrm{F}_s$, $\mathrm{V}_s\mathrm{F}_n$, $\mathrm{V}_n\mathrm{F}_n$, $\mathrm{V}_n\mathrm{F}_s$ (see Table I).

C. Procedure

Participants watched a total of 28 clips, in random order. Of these 28, 16 were multimodal (audio + video), 8 video only (4 with the human-like and 4 with the cartoon-like agent) and 4 audio only. Instructions to the participants specified that some of the clips would have no audio or no video, so as to prevent them from thinking that the clips were broken. Participants watched one video at a time and were then asked to evaluate them using the SAM scale, shown in Fig. 2. There was one scale for valence and one for arousal, each made of 5 points (1 = lowest valence / least aroused, 5 = highest valence / most aroused). Participants could watch each clip as many times as they wanted, but the rating scales appeared on the screen only after the clip was played in full. This evaluation study took approximately 10 minutes, and workers were compensated $3.2, which is higher than minimum wage in the USA, and higher than average Mechanical Turk pay rate [35].

D. Participants

We recruited 104 participants from AMT, however we had to exclude 6 who reported technical issues with the experiment. Of the remaining 98 participants, 30 were women, 66 men, and 2 preferred not to say; their age range was 20-67 years old (median = 35 years old). Since the experiment was conducted in English, we asked participants to self-report their English language fluency: 93 people identified as native English speakers, 1 as near-native, and 4 as fluent. The experiment was conducted in accordance with ethical guidelines from KTH Royal Institute of Technology.

IV. Results

First of all, we looked at outliers in the data as potential indicators that participants were not paying attention to the task, so we examined the time it took participants to evaluate each video; a rule of thumb to discover outliers is to filter out data points that are 3 standard deviations away from the mean. We excluded individual trials for which participants took longer than 3 standard deviations away from the log of the mean of all participants. This resulted in the exclusion of 35 individual trials, for a total of 1547 analysable trials. As our response variables are ordinal categorical in nature, we used non-parametric statistical inference techniques to evaluate the effect of our independent variables. The analyses were conducted in R version 4.0.1, using packages `ordinal` [36] and `RVAideMemoire` [37].

A. Does smiling in the voice and / or face influence valence and arousal ratings?

To answer RQ1-a and RQ1-b, we fitted cumulative link mixed models to predict arousal and valence ratings, with smiling in the face and voice and agent style as predictors, and participant id as random effect. We fitted these models separately for the audio-visual, audio only, and video only conditions. The full results are shown in Table II.

For valence, in the audio-video condition, only face smiling was a significant predictor, with higher ratings for smiling faces (Fig. 3). There was also a main effect of agent style, with higher ratings for the human-like agent. In the audio-only condition, voice smiling significantly increased ratings. In the video-only condition, face smiling significantly increased ratings; there was also a main effect of agent style, with the human-like agent eliciting higher ratings.

For arousal, in the audio-video condition, we found both face smiling and audio smiling to be significant predictors: as can be seen from Fig. 3, the presence of smiling in the face and / or voice increased the arousal ratings. There was no main effect of virtual agent style. In the audio-only condition, voice smiling significantly increased ratings. In the video-only condition, face smiling significantly increased ratings. There was also a main effect of virtual agent style, with higher ratings for the human-like agent.

B. Do people perceive incongruent emotional expressions differently from congruent ones, or is there a channel predominance?

To answer RQ2 and RQ3, we ran a series of planned comparisons between the ratings of the incongruent and congruent conditions. We ran 4 Wilcoxon rank sum tests with continuity correction: V_sF_n vs. V_sF_s and V_nF_n, and V_nF_s vs. V_sF_s and V_nF_n (see Table I). To account for multiple comparisons, we set the significance level α to $0.05/4 = 0.0125$. Details of the comparisons can be found in Table III. For valence, the V_sF_n condition received lower ratings than the V_sF_s condition, but received the same ratings as the V_nF_n condition. On the other hand, the V_nF_s condition received higher ratings than the V_nF_n condition and the same ratings as the V_sF_s condition. The arousal ratings followed the same pattern: the V_sF_n condition received lower ratings than the V_sF_s condition and the same ratings as the V_nF_n condition; the V_nF_s condition received higher ratings than the V_nF_n condition and the same ratings as the V_sF_s condition. Thus, overall, the incongruent condition with smiling face and neutral voice was rated similarly to the congruent smiling condition, and the incongruent condition with smiling voice and neutral face was rated similarly to the congruent neutral condition.

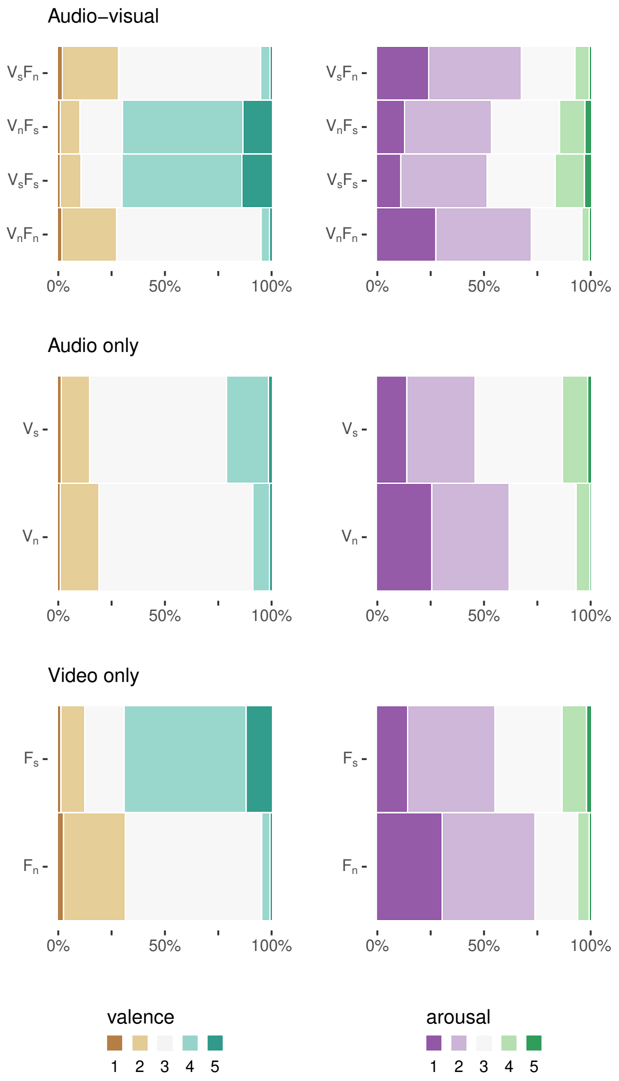

Fig. 3. Percentage of each level of valence (left column) and arousal (right column) ratings (1 = lowest, 5 = highest). The ratings are broken down by emotional expression (Table I) and experimental condition: audio-visual (top row), audio only (middle row) and video only (bottom row).

V. Discussion

We studied how people perceive incongruent emotional information from the face and voice of a virtual agent, in terms of valence and arousal (the two main dimensions in the Circumplex model of emotions [38]). We created an artificial

TABLE II

OUTPUT OF THE CUMULATIVE LINK MIXED MODELS FOR VALENCE AND AROUSAL, IN THE THREE EXPERIMENTAL CONDITIONS (AUDIO-VISUAL, VIDEO ONLY, AUDIO ONLY). ODDS RATIOS, CONFIDENCE INTERVALS, AND P-VALUES ARE REPORTED. NEUTRAL VOICE, NEUTRAL FACE, AND CARTOON-LIKE AGENT WERE THE REFERENCE LEVELS.

		Valence			Arousal		
		OR	95% CI	p-value	OR	95% CI	p-value
Audio-visual	Voice (smiling)	0.99	[0.81, 1.21]	n.s.	1.26	[1.03, 1.53]	.022
	Face (smiling)	25.98	[1.98, 34.14]	<.001	3.57	[2.91, 4.39]	<.001
	Agent (human-like)	0.67	[0.55, 0.82]	<.001	1.02	[0.84, 1.24]	n.s.
Video only	Face (smiling)	25.91	[17.63, 38.08]	<.001	3.76	[2.80, 5.04]	<.001
	Agent (human-like)	0.63	[0.47, 0.84]	. 001	1.38	[1.05, 1.82]	.022
Audio only	Voice (smiling)	2.29	[1.43, 3.67]	<.001	3.26	[2.12, 5.01]	<.001

TABLE III

NON-PARAMETRIC COMPARISONS BETWEEN THE CONGRUENT AND INCONGRUENT STIMULI IN VALENCE AND AROUSAL RATINGS. 'N' = NUMBER OF OBSERVATIONS, 'W' = WILCOXON TEST STATISTIC.

		V_sF_s			V_nF_n		
		N	W	p-value	N	W	p-value
Valence	V_sF_n	772	24844	<.001	773	74083	n.s.
	V_nF_s	771	74019	n.s.	772	124206	<.001
Arousal	V_sF_n	772	56519	<.001	773	79733	n.s.
	V_nF_s	771	71675	n.s.	772	95054	<.001

agent – in two different styles – whose facial expressions were driven by motion capture, and whose voice was obtained from the same actor that provided the facial expressions. We then mixed and matched these two channels to create 4 multimodal and 4 unimodal stimuli (Table I) that were evaluated in an online study.

Confirming our expectations (RQ1-a), we found that smiling in the face and / or voice increased arousal and valence ratings. This is consistent with our previous studies, where these same stimuli were also rated as 'happier' [32, 33]. For arousal, this happened for all our stimuli, whether unimodal (only the visual or audio channel was present) or multimodal (both visual and audio channels were present). This suggests that both face and voice smiling convey arousal information, even when there is a channel mismatch. Interestingly, for valence face seemed to be the predominant channel for transmitting this information. While in unimodal stimuli smiling in the voice or face increased valence ratings, for multimodal stimuli we only found a main effect of face smiling. This means that, when presented with conflicting information (smiling voice + neutral face), information from the visual channel prevailed. This is consistent with previous studies that hypothesised that visual and audio channels transmit different aspects of emotion: specifically, according to an 'emotional McGurk effect' theory, the audio channel is used to infer arousal, and the visual channel to infer valence [39]. In our study, smiling in the voice increased ratings of arousal even when the overlapped visual channel showed that the agent had a neutral expression, suggesting that voice can convey arousal even when there is conflicting information coming from the face. On the other hand, the visual channel prevailed over conflicting valence information. This phenomenon can be framed as an information allocation problem [5]: the emotional information to be transmitted is determined by the bandwidth of the available channels. In the case of audio-visual modalities, people have more available information to perceive a certain emotion, similarly to having more contextual information. When there is only one modality, this information is reduced, and the message needs to be perceived from incomplete sources.

Furthermore, this effect is at least partially dependent on agent-related factors, namely human-likeness. We had expected cartoon-like agents to be perceived as more expressive (RQ1-b). However, our results suggest that the human-like style was perceived as more expressive, as we found a main effect of agent style on valence in the audio-visual stimuli and on arousal in the video-only stimuli. This is unexpected, but it might be due to the fact that both agents were animated by the same motion capture, so the cartoon-like agent was not 'exaggerated' as they usually would be. For this reason, maybe people were expecting a more exaggerated expression from the cartoon-like avatar and gave it overall lower ratings as a result.

We also found a clear clustering of our mismatched multimodal stimuli: clips featuring a smiling face and neutral voice were rated similarly to the congruent smiling clips, and clips featuring a neutral face and smiling voice were rated similarly to the congruent neutral clips (Fig. 3). This result suggests the presence of channel predominance, and that information coming from the face overrode information from the voice (RQ3). This confirms results previously obtained with human stimuli [15] and virtual agents whose facial and bodily expressions were mismatched [40]. However, our findings contrast with results previously obtained with virtual characters [5]. Technology advances since these studies were conducted meant that we could generate a higher quality virtual character, and our results suggest that the emotional expressions of such high quality agents undergo similar processes as human ones [15]. However, while explicit perceptual evaluations – such as those performed in the current study – suggest a similar processing of emotional expressions in humans and virtual characters, we cannot conclude that these similarities will occur at the behavioural or physiological level as well [cf. 41].

Our results have practical implications for Human-Machine Interaction. In cases where we might not have complete control over the audio channel – for example, where technical limitations prevent us from creating an expressive synthetic voice – the current results suggest that expressivity in the face of a virtual agent alone might be enough to convey emotional information. With the current study, we have provided new insights into how mismatches in highly expressive channels are perceived; this investigation however is still ongoing, and more studies are needed on other virtual agents that might have inherent expressiveness mismatches, such as robots.

There are some limitations that should be mentioned. To the best of our knowledge, this is the first study looking at perceptual evaluations of a mismatched emotional expression – smiling – rather than of an 'emotion' (e.g. 'happiness' and 'sadness', as done in previous studies [5]). We purposely did not mention any emotion labels to our participants, to try and elicit unbiased ratings. We believe this to be an important distinction to make, since the same emotional expression can be a sign of different emotions. However, for this same reason, emotional expressions are made sense of in context [28, 42]. Thus, our participants might have found it difficult or meaningless to evaluate these smiling expressions without a context. We intend to continue this line of research and add contextual information to the stimuli; for now, the current work provides initial results that facial expressivity overrides vocal expressivity in emotion processing in virtual agents, when contextual information is not available.

Also, it should be noted that the two virtual agents differed along more than the human-likeness dimension (for example, the human-like agent was bald, the cartoon-like agent had a longer face, etc). While we kept some other features, such as lighting, constant [34], it is still possible that the partial effects that virtual agent style had on the ratings were due to these other differences, and not to the intended manipulation. However, manipulating all these possible differences quickly results in an exponential growth of experimental conditions, which would have been impossible to adequately study in the current experiment, especially given that agent human-likeness was a secondary variable of interest. Future studies will need to disentangle this potential confound by evaluating a wider range of agent styles (including genders); other modalities (such as gestures and body posture) should be investigated as well.

VI. Conclusion

Forty-five years have passed since the accidental discovery of the McGurk effect [16]. Since then, technological advances have allowed us to study the fascinating phenomenon of multisensory integration from a wide variety of angles. Here, we created a 'smiling McGurk effect' in a virtual character and found evidence that information from the visual channel overrode information from the audio channel when the two channels were mismatched. Many artificial agents currently in use allow to independently design and manipulate their visual and audio expressions. Our results suggest that, in the absence of equal expressivity capabilities in both channels, visual expressivity should be prioritised.

References

[1] D. L. Johanson, H. S. Ahn, C. J. Sutherland, B. Brown, B. A. MacDonald, J. Y. Lim, B. K. Ahn, and E. Broadbent, "Smiling and use of first-name by a healthcare receptionist robot: Effects on user perceptions, attitudes, and behaviours," *Paladyn, Journal of Behavioral Robotics*, vol. 11, no. 1, pp. 40–51, 2020.

[2] I. Torre, J. Goslin, and L. White, "If your device could smile: People trust happy-sounding artificial agents more," *Computers in Human Behavior*, vol. 105, p. 106215, 2020.

[3] 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.

[4] C. M. de Melo, J. Gratch, and P. J. Carnevale, "Humans versus computers: Impact of emotion expressions on people's decision making," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 127–136, 2015.

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

[6] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2011.

[7] A. R. Damasio, "Emotion in the perspective of an integrated nervous system," *Brain Research Reviews*, vol. 26, no. 2-3, pp. 83–86, 1998.

[8] A. S. Cowen, P. Laukka, H. A. Elfenbein, R. Liu, and D. Keltner, "The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures," *Nature Human Behaviour*, p. 1, 2019.

[9] P. Ekman and W. V. Friesen, "Felt, false, and miserable smiles," *Journal of Nonverbal Behavior*, vol. 6, no. 4, pp. 238–252, 1982.

[10] M. G. Calvo and L. Nummenmaa, "Perceptual and affective mechanisms in facial expression recognition: An integrative review," *Cognition and Emotion*, vol. 30, no. 6, pp. 1081–1106, 2016.

[11] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-cultural Psychology*, vol. 32, no. 1, pp. 76–92, 2001.

[12] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2012.

[13] B. E. Stein and T. R. Stanford, "Multisensory integration: Current issues from the perspective of the single neuron," *Nature Reviews Neuroscience*, vol. 9, no. 4, pp. 255–266, 2008.

[14] S. Campanella and P. Belin, "Integrating face and voice in person perception," *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 535–543, 2007.

[15] 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, 2000.

[16] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[17] P. Arias, L. Rachman, M. Liuni, and J.-J. Aucouturier, "Beyond correlation: acoustic transformation methods for the experimental study of emotional voice and speech," *Emotion Review*, vol. 13, no. 1, pp. 12–24, 2020.

[18] C. Creed and R. Beale, "Psychological responses to simulated displays of mismatched emotional expressions," *Interacting with Computers*, vol. 20, no. 2, pp. 225–239, 2008.

[19] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze, "Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots," *International Journal of Social Robotics*, vol. 11, no. 4, pp. 555–573, 2019.

[20] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, vol. 39, no. 1, pp. 18–49, 2011.

[21] S. Hamann, "Mapping discrete and dimensional emotions onto the brain: Controversies and consensus," *Trends in Cognitive Sciences*, vol. 16, no. 9, pp. 458–466, 2012.

[22] J. A. Russell, "Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies." *Psychological Bulletin*, vol. 115, no. 1, p. 102, 1994.

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

[24] M. Mehu and R. I. M. Dunbar, "Relationship between smiling and laughter in humans (homo sapiens): Testing the power asymmetry hypothesis," *Folia Primatologica*, vol. 79, no. 5, pp. 269–280, 2008.

[25] S. Fagel, "Effects of smiling on articulation: Lips, larynx and acoustics," in *Development of multimodal interfaces: active listening and synchrony*. Springer, 2010, pp. 294–303.

[26] K. El Haddad, I. Torre, E. Gilmartin, H. Çakmak, S. Dupont, T. Dutoit, and N. Campbell, "Introducing amus: The amused speech database," in *Statistical Language and Speech Processing*, N. Camelin, Y. Estève, and C. Martín-Vide, Eds. Springer International Publishing, 2017, pp. 229–240.

[27] P. Arias, C. Soladie, O. Bouafif, A. Roebel, R. Seguier, and J.-J. Aucouturier, "Realistic transformation of facial and vocal smiles in real-time audiovisual streams," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 507–518, 2018.

[28] M. Rychlowska, R. E. Jack, O. G. B. Garrod, P. G. Schyns, J. D. Martin, and P. M. Niedenthal, "Functional smiles: Tools for love, sympathy, and war," *Psychological Science*, vol. 28, no. 9, pp. 1259–1270, 2017.

[29] K. Dion, E. Berscheid, and E. Walster, "What is beautiful is good." *Journal of Personality and Social Psychology*, vol. 24, no. 3, p. 285, 1972.

[30] E. Zell, C. Aliaga, A. Jarabo, K. Zibrek, D. Gutierrez, R. McDonnell, and M. Botsch, "To stylize or not to stylize? the effect of shape and material stylization on the perception of computer-generated faces," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 1–12, 2015.

[31] R. McDonnell, M. Breidt, and H. H. Bülthoff, "Render me real? investigating the effect of render style on the perception of animated virtual humans," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, pp. 1–11, 2012.

[32] I. Torre, E. Carrigan, R. McDonnell, K. Domijan, K. McCabe, and N. Harte, "The effect of multimodal emotional expression and agent appearance on trust in human-agent interaction," in *Motion, Interaction and Games*. New York, NY, USA: ACM, 2019.

[33] I. Torre, E. Carrigan, K. Domijan, R. McDonnell, and N. Harte, "The effect of audio-visual smiles on social influence in a cooperative human-agent interaction task," *ACM Transactions on Computer-Human Interaction*, in press.

[34] P. Wisessing, K. Zibrek, D. W. Cunningham, J. Dingliana, and R. McDonnell, "Enlighten me: Importance of brightness and shadow for character emotion and appeal," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 3, pp. 1–12, 2020.

[35] K. Hara, A. Adams, K. Milland, S. Savage, C. Callison-Burch, and J. P. Bigham, "A data-driven analysis of workers' earnings on amazon mechanical turk," in *Proceedings of the 2018 SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2018, pp. 1–14.

[36] R. H. B. Christensen, "ordinal—regression models for ordinal data," 2019, r package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

[37] M. Hervé and M. M. Hervé, "Package 'rvaidememoire'," 2020, https://CRAN. R-project. org/package= RVAide-Memoire.

[38] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

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

[40] C. Clavel, J. Plessier, J.-C. Martin, L. Ach, and B. Morel, "Combining facial and postural expressions of emotions in a virtual character," in *International Workshop on Intelligent Virtual Agents*. Springer, 2009, pp. 287–300.

[41] A. W. de Borst and B. de Gelder, "Is it the real deal? Perception of virtual characters versus humans: an affective cognitive neuroscience perspective," *Frontiers in Psychology*, vol. 6, p. 576, 2015.

[42] L. F. Barrett and E. A. Kensinger, "Context is routinely encoded during emotion perception," *Psychological Science*, vol. 21, no. 4, pp. 595–599, 2010.