FrankenFolk: Distinctiveness and Attractiveness of Voice and Motion

JAN ONDŘEJ, Disney Research, Los Angeles
CATHY ENNIS, Dublin Institute of Technology
NIAMH A. MERRIMAN, Disney Research, Los Angeles
CAROL O'SULLIVAN, Trinity College Dublin

It is common practice in movies and games to use different actors for the voice and body/face motion of a virtual character. What effect does the combination of these different modalities have on the perception of the viewer? In this article, we conduct a series of experiments to evaluate the distinctiveness and attractiveness of human motions (face and body) and voices. We also create combination characters called FrankenFolks, where we mix and match the voice, body motion, face motion, and avatar of different actors and ask which modality is most dominant when determining distinctiveness and attractiveness or whether the effects are cumulative.

CCS Concepts: • **Computing methodologies** → **Animation**; **Perception**; *Motion capture*; Rendering

Additional Key Words and Phrases: Virtual characters, multisensory perception, crowd variety

ACM Reference Format:

Jan Ondřej, Cathy Ennis, Niamh A. Merriman, and Carol O'Sullivan. 2016. FrankenFolk: Distinctiveness and attractiveness of voice and motion. ACM Trans. Appl. Percept. 13, 4, Article 20 (July 2016), 13 pages.
DOI: http://dx.doi.org/10.1145/2948066

1. INTRODUCTION

With the increased demand for realism in virtual characters in recent years, performance capture of actors has become ubiquitous. Apart from the advantages that this approach provides, such as highly realistic motions and voices, there are several potential problems. One challenge is the infeasibility of capturing all modalities simultaneously (i.e., voice, body, face, hands, avatar's appearance) from a unique actor for every character, especially for crowd creation. Other practical constraints, especially for real-time applications, are limited hardware and time budgets. Thus, it is common in games and movies to combine and reuse the voice recordings and motions (face, body, fingers) of actors and to apply them to a variety of different three-dimensional (3D) characters. However, what is the effect

This work was partially supported by a Science Foundation Ireland Principal Investigator Award (S.F.I. 10/IN.1/13003).
Authors' addresses: J. Ondrej, Disney Research LA, 1401 Flower Street Glendale, CA 91201-5020, USA; email: jan.ondrej@disneyresearch.com; C. Ennis, Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland; email: cathy.ennis@dit.ie; N. A. Merriman, Royal College of Surgeons in Ireland, Beaux Lane House, Lower Mercer Street, Dublin 2, Ireland; email: niamhmerriman@rcsi.ie; C. O'Sullivan, Trinity College Dublin, College Green, Dublin 2, Ireland; email: Carol.OSullivan@scss.tcd.ie.

DOI: http://dx.doi.org/10.1145/2948066

of combining different modalities from different actors to create new virtual characters, which we nickname *FrankenFolk*? This is the question we explore in this article.

To explore the perception of FrankenFolk, we focus on perceived Attractiveness and Distinctiveness, as in Hoyet et al. [2013]. However, they explored the perception of body motion only for different types of locomotion. In our case we focus on short speaking performances and ask the following questions:

—How distinctive and attractive is each modality (i.e., voice, face or body motion, physical appearance) when presented in isolation, that is, as a *Partial* performance?

—How does each partial performance relate to the overall, *Full* performance?

—What, if any, modality most strongly influences the attractiveness or distinctiveness of a character's performance?

To answer these questions, we conducted a series of experiments: in the Full baseline experiment we evaluate the distinctiveness and attractiveness of each actor's performance, presented both as a Real video and applied to a Virtual character. In the Partial baseline experiment, we explore each modality in isolation. Finally, we create FrankenFolk characters, where we mix and match the voice, body motion, face motion, and avatar of different actors. We found that an actor's Voice may be the most distinctive feature of his performance, but we found that only for males. Females in general were less easy to recognize. We also found that body motion and character's appearance (avatar) were most indicative of perceived attractiveness.

Our results highlight the importance of paying attention to all modalities when creating virtual performances captured from multiple different actors. If, for example, a particular actor's voice or body motion is highly distinctive or unattractive, it could adversely affect the overall performance. Furthermore, repetition of such a voice or motion in a crowd would stand out, attract undesirable attention, and detract from the overall realism of a scene.

2. RELATED WORK

Previous studies in computer graphics have shown that the appearance of a virtual character is much more distinctive than its walking motion [McDonnell et al. 2008], but that simple tricks can help to disguise similar "clones" in a crowd by simply varying their textures or accessories [McDonnell et al. 2009]. However, while walking motions tend to be relatively difficult to recognize [Prazak and O'Sullivan 2011], neither distinctiveness nor attractiveness transfers to other gaits such as jogging or dancing [Hoyet et al. 2013]. Dancing was found to be very distinctive for each actor, so how about speaking and gesturing, which can also be quite stylistic?

When combining different body and facial motions and voices there is a need to be particularly careful as people are very sensitive to anomalies and artifacts. This is especially important in faces, where, for example, a mismatch of audio and video signals can negatively affect experience [Carter et al. 2010]. Even when very significant body animation anomalies are present, people find facial anomalies more salient [Hodgins et al. 2010]. Participants were found to be more sensitive to visual desynchronization of body motions within groups of talking characters than to mismatches between the characters own gestures and voices [Ennis et al. 2010].

It is thought that audio and visual cues are integrated in the manner of a single percept. One signal can affect the other, as demonstrated in the McGurk effect [McGurk and MacDonald 1976]. And, indeed, there are studies that show that voice and visual stimuli can rely on similar properties when it comes to perceived attractiveness, for example, averageness [Rhodes and Tremewan 1996; Feinberg et al. 2008]. There is also a stronger link between audio and visual information when it comes to ratings of attractiveness. It has been shown that female vocal and visual attractiveness are related. A study

Table I. Terminology

Actor	Performer
Video	Video recording of actor's performance
Sex	Sex of the actor
Voice	Recorded voice (with facial motion if presented)
Face	Facial motions only
Body	Full body motion, excluding facial animation
Avatar	Character's appearance, one of Rocketbox models from Figure 2
Character	Virtual human with one or more modalities combined (Voice, Body, Face, Avatar)

where male participants rated attractiveness of female photographs and recordings of vowel sounds has shown that males were in strong agreement on attractive qualities and that attractive faces had attractive voices [Collins and Missing 2003]. The authors deduce that qualities linked to attractiveness such as age and body size are inherent in both visual and audio stimuli. Other studies have found links between attractive audio and visual stimuli on a bimodal signal [Zuckerman and Driver 1989] as well as comparing the effects on static images and dynamic visual stimuli [Lander 2008].

The qualities that make a voice attractive have also been widely studied. It has been shown that women tend to find male voices of lower pitch more attractive [Hodges-Simeon et al. 2010] and that men prefer female voices to be more high pitched [Feinberg et al. 2008]. However, female voices become less attractive if they are perceived to be too high pitched ($\geq$280Hz) [Borkowska and Pawlowski 2011]. Interestingly, increasing the pitch in female voices and lowering pitch in male voices has been shown to have no effect on the perceived attractiveness, whereas the opposite modifications significantly decrease attractiveness. This implies that there are inherent attractive voice qualities that are not modified artificially by pitch changes [Fraccaro et al. 2013].

When it comes to matching voices to faces it has been shown that people can match unfamiliar voices to unfamiliar faces [Kamachi et al. 2003]. Also, familiarity in voices is more easily detected when it is accompanied with a familiar face regardless of whether it was correct voice or not [Schweinberger et al. 2007]. While changing the sentence has no effect on matching performance, modifying the intonation can disrupt it [Lander et al. 2007]. A more detailed review of research from behavioral, neuropsychological, electrophysiological, and neuroimaging studies on faces and voices processing and integration can be found in Yovel and Belin [2013].

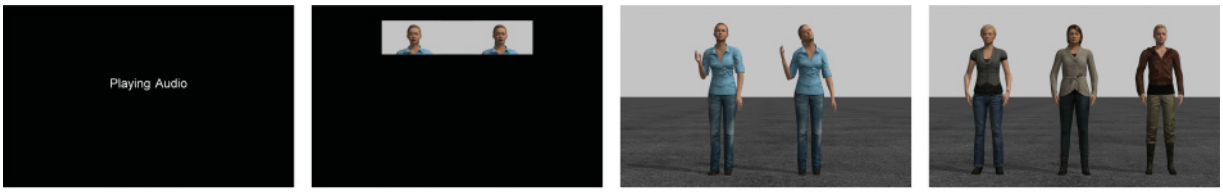
3. STIMULI

We captured the performance of six male and six female formally trained actors while they were reciting a monologue from a TV series. Actors were instructed to follow a script, which was annotated with specific gestures at required points. So, while actors were free to express their natural acting style, there was enough similarity to allow the stimuli to be compared.

3.1 Performance Recording

The voice with facial and full body movements were recorded simultaneously, and each actor performed two takes. To capture motion we used an optical motion capture system Vicon with 21 cameras. There were 52 markers placed on the body and 36 markers on the face. We did not capture finger or eye movements of our actors, which is typical for game applications. We also captured video of each actor performance.

To record voice, A RODE NT-5 studio condenser microphone was placed on a tripod in front of each actor. In advance of recording, the preamp gain was adjusted for each actor, ensuring minimal audio distortion during the capture session. The audio was synced to the motion capture using a clapboard



(a) Voice (b) Face motion (c) Body motion (d) Avatar

Fig. 1. A first screen from the distinctiveness experiment for partial baseline stimuli.

with motion capture markers attached, and each audio file was normalized post-recording to −2dB in order to ensure equal peak volume across actors and takes. All audio was recorded using a MOTU Traveler-mk3 external audio interface.

3.2 Motion Processing

A skeleton-based approach that used dual quaternion skinning was used to drive the body and facial geometry in line with previous work [Hoyet et al. 2013; Ennis et al. 2013]. The motion-captured body markers were used to compute the joint angle animations and mapped onto the virtual character using the Biped retargeting system in Autodesk 3DS Max 2014. The facial motion was directly exported as 3D marker motion and stabilized by removing the movement of the head. The markers were first aligned to the head and then automatically adjusted to the positions of the bones on the face of the character. The character's facial bones were then constrained to their corresponding optical markers to produce the animation. Once the facial and body motions were combined, the resulting animations were imported to NaturalMotion's Morpheme to generate animation finite state machines (FSMs). To render the stimuli, an in-house game engine based on Ogre3D was used.

3.3 Stimuli Creation

In the "Full Baseline" experiment, we studied the full performance of the actor in a video and transferred onto an avatar (i.e., voice and both body and face motion). The "Partial Baseline" experiment explored each different modality of the actors' performances (i.e., voice, facial motion, body motion) along with the appearance of the characters that were later used to visualize them. Finally, in the "FrankenFolk" experiment, we combined modalities from different actors to create a new performance. Examples of all stimuli can be found in the supplemental video.

Voice: The output sound from the scene was rendered in a stereo mode through noise-canceling headphones. The audio file (WAV) was played as an omnidirectional mono source from the character's position and was synchronized with character's lip motion if a character was used. When only the monologue was played (without a character), a black screen was shown with the text "Playing Audio" (see Figure 1(a)).

Face and Body Motion: To animate the character, we used NaturalMotion's Morpheme animation library. We created one FSM per actor containing both clips with a parameter to independently turn on/off the facial and full body animations. When the modality was turned off a default pose was used; a neutral facial expression and/or a standing pose. When only the body motion was playing, we chose to simply freeze the facial motion. If only a facial animation was played, then we occluded the avatar's body (see Figure 1(b)). We chose to display the neutral facial expression, as we found that this looked more natural and less distracting than occluding or blurring the face entirely. It allowed the participants to focus solely on the body motion without distraction. On the other hand, we occluded the body when the face was playing, as the body was much larger in scale and thus it was easier to focus on the facial motion when nothing else was present on the screen.

Fig. 2. Rocketbox avatars used in the partial baseline experiment with two versions of clothing for male and female.

Avatar: To test the effect of character's appearance we used six male and six female models from Rocketbox (see Figure 2). Each model had two materials with different clothing textures. For stimuli where model was not a factor, we used the same two models as in Hoyet et al. [2013].

Video: Each performance was recorded on a Pal video camera placed in front of the actor. To use videos as stimuli in our second experiment, we replaced the original audio from the camera with the one from the performance recording to preserve the same quality between blocks. Every video was cropped to show only actor (if possible) and scaled to match approximately the size of the virtual

character on the screen from other blocks. To play the video in our rendering engine, we used a dynamic texture and applied it on a rectangular object with a size proportion matching the video. Figure 3 shows a screen shot of all the actors.

FrankenFolk: To combine modalities together for the FrankenFolk stimuli, we synchronized the Body animation (gestures) from one actor with the Face animation and Voice from another. We did not combine voice with face animation from different actors, as modifying the intonation of a voice can change it significantly [Lander et al. 2007], while morphing the facial animation resulted in obvious artifacts in some cases. The morphed body motions, on the other hand, were of good quality and varied only minimally from the original. We also applied the resulting combinations to different avatars.

To combine facial and body animation, we created one animation FSM for the males and one for the females containing both clips from all the actors. To synchronize a body motion (gestures) with a facial animation we annotated every animation with event marks, which were placed at the time of important gestures from the monologue script. We then used a built-in event synchronization system in Morpheme to blend facial and body animations from different actors or/and clips together.

4. EXPERIMENTS

All our experiments consist of two tasks: a recognition task to evaluate *Distinctiveness* and a rating task to evaluate *Attractiveness*. Participants do either one or both of these tasks, depending on the condition. In all cases, the presentation of stimuli is fully randomized.

In the Distinctiveness condition, a group of either two or three character stimuli side by side is first displayed or, in the case of audio, presented simultaneously either via headphones or speakers. (In the FULL baseline and the Avatar partial baseline, a group of three stimuli is shown, as otherwise the task is too easy and individual differences in distinctiveness would not become evident. In all other cases, two stimuli are shown. This choice was guided by pilot experiments.) After seeing this first group, a single character stimulus is then shown, and the participant's task is to indicate whether the single stimulus was *Present* or *Absent* from the previous group. In 50% of cases the stimulus will have been present. The other distractor stimuli in the group are picked at random from the remaining ones. All clips in both conditions are 5s long.

To test for effects of the independent variables on the participant responses, we perform Analysis of Variance (ANOVA) to test for main and interaction effects. We report the F and p values for significant effects (i.e., $p < .05$) along with effect sizes partial eta-squared (η_p^2) that can be interpreted as small (>0.01), medium (>0.06), or large (>0.14) as described by Cohen [1988].

For Attractiveness, the participants are asked to rate how attractive a single character stimulus is on a scale of 1–7, where 1 is least attractive and 7 is most attractive. They are told to evaluate this as objectively as possible, as if they were looking at a person on the street or the television. Where appropriate, they are told to consider all modalities when making their judgments.

4.1 Full Baseline

In this experiment, we studied the full performance of the actor in a Real video and also applied to a Virtual character (i.e., voice and both body and face motion). The male and female models used are shown in Figure 3. Our aim was to see whether distinctiveness and attractiveness varied between the Real and Virtual conditions. We also wished to provide a baseline against which to compare the results of each modality separately. In the distinctiveness condition, we displayed three stimuli side by side, in order to raise the difficulty level of the experiment.

Thirteen volunteers (7M/6F,20-40y) participated in this study, of whom eight completed both conditions (in counter-balanced order), two attractiveness only and three distinctiveness only. There were

Fig. 3. A first frame from actors' performance used in baseline experiments: a video recording and applied on one of the avatars (bellow). From the left: female actors A1-A6, male actors A1-A6.

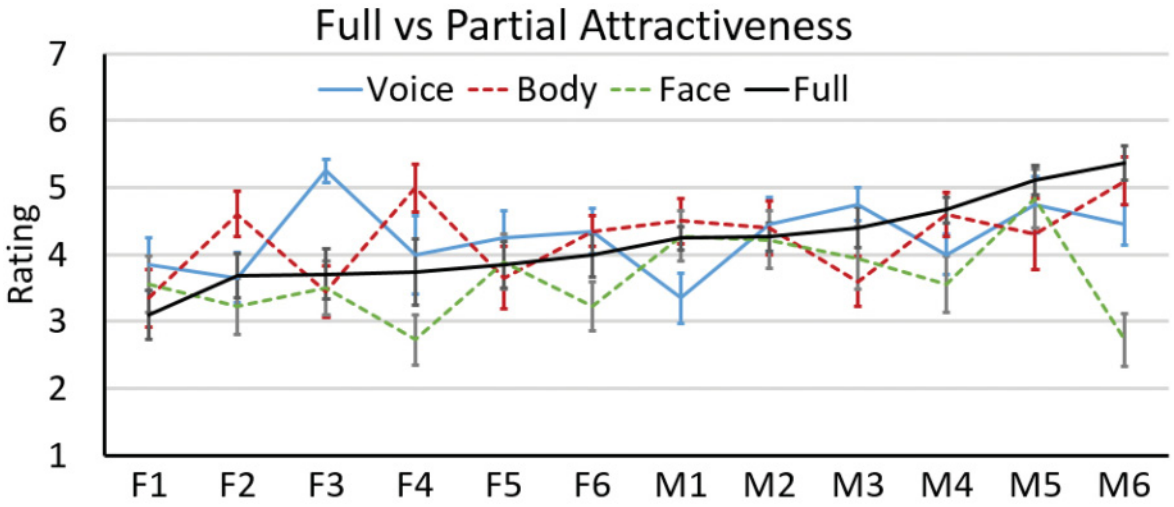


Fig. 4. Comparison of attractiveness ratings between the Full and Partial baselines for all actors (Females F1–F6, Males M1–M6).

two clips for every mode, and each was repeated 3 times. In the distinctiveness condition, there were repetitions for both present and absent.

We performed a repeated-measures ANOVA with within-subjects factors Mode (Real or Virtual) and Actor (6M, 6F) on the results of both conditions. We found a large main effect of Mode for the distinctiveness results, where people accurately recognized the Real stimuli 90% of the time, which was significantly greater than the 80% rate for Virtual ($F(1, 10) = 8.7$, $p < .05$, $\eta_p^2 = .47$). We found no effect of actor or any interaction effect, which indicates that the task may still have been too easy for participants. Of interest is the fact that we found no significant difference in Attractiveness ratings between the Real and the Virtual stimuli. However, we did find a large main effect of Actor ($F(11, 99) = 4.2$, $p < .00005$, $\eta_p^2 = .32$) and these results are shown for comparison purposes in Figure 4.

4.2 Partial Baseline

In this experiment, we explored the attractiveness and distinctiveness of each modality independently in four blocks: Voice, Face, Body, and Avatar (see Figures 1(a)–(d) for examples). Our aim was to determine whether any one modality was more significant when trying to recognize or rate the attractiveness of a performance. We ran the experiment over a 2-day science fair and recruited 67 participants for the distinctiveness condition (16, 16, 17, 18 for the Voice, Body, Face, and Avatar conditions, respectively) and 38 participants for Attractiveness (10, 10, 9, 9, respectively). The number of men and women was approximately equal with a wide range of ages (20–65y).

When showing the first group in which the target stimulus was present or absent, the group size for Avatar was 3, as otherwise the task was too simple. All other groups were of size 2. Furthermore, it was important to present the voices together, as the application scenario in which we wish to present the results is in the simulation of crowds and groups of characters. Hence, knowing when an individual voice is distinctive in such scenarios is very important, as otherwise repetition of voices on different characters within the crowd would become very noticeable. From pilot experiments we determined that presenting the Voice stimuli in pairs did allow for individual differences in distinctiveness to become evident—presenting only one voice was too easy to recognize, whereas presenting three voices became overwhelming.

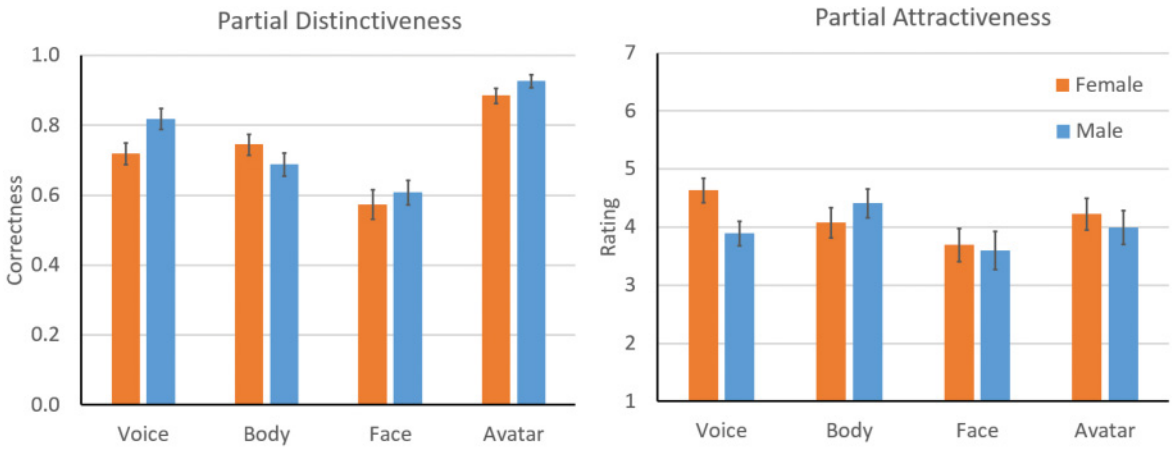


Fig. 5. Interaction effects of Sex*Mode from the Partial baseline experiment for Distinctiveness (l) and Attractiveness (r).

As all participants only evaluated one modality, we first performed a repeated-measures mixed ANOVA on the Distinctiveness data, with categorical predictor Mode (Voice, Face, Body, Avatar) and within-subjects factor Sex (F,M) of actor, averaged over Actor and repetitions. The dependent variable was % of correct present or absent responses. We found a large main effect of Mode ($F(3, 63) = 29.6$, $p < .000005$, $\eta_p^2 = .58$). A post hoc analysis (using Fisher's least significant difference (LSD) comparison between means) shows that all four modes overall significantly differ in terms of distinctiveness, with Avatar > Voice > Body > Face. A medium interaction effect Sex*Mode ($F(3, 63) = 2.7$, $p < 0.05$, $\eta_p^2 = .11$), as can be seen in Figure 5 (left), indicates that only the male Voice is significantly more distinctive than the male Body, whereas we did not find a significant difference between Body and Voice for the female actors.

We performed the same analysis on the Attractiveness data, with dependent variable average attractiveness rating. We found a large interaction effect of Sex*Mode ($F(3, 34) = 4.7$, $p < 0.05$, $\eta_p^2 = .29$), as shown in Figure 5 (right). We can see that the female voices rated significantly more attractive than the males, while both male and female face motions were rated as being quite unattractive overall.

We also performed a repeated-measures mixed ANOVA on both sets of data, with categorical predictor Mode (Voice, Face, Body) and within subjects factor Actor (1–12). We removed the virtual Avatar mode from this analysis, as each (Voice, Body, Face) group belonged to one of our real actors but not the Avatars. For Distinctiveness, we found a medium main effect of Actor ($F(11, 506) = 2.9$, $p < 0.005$, $\eta_p^2 = .06$) and a medium interaction effect of Actor*Mode ($F(22, 506) = 1.9$, $p < 0.05$, $\eta_p^2 = .07$). For Attractiveness, we also found a medium main effect of Actor ($F(11, 286) = 1.9$, $p < .05$, $\eta_p^2 = .07$) and a large interaction effect of Actor*Mode ($F(22, 286) = 3.8$, $p < 0.000005$, $\eta_p^2 = .22$). There were many significant differences between the actors in all the modes, for both conditions, and no significant correlation between any of those modes, that is, distinctiveness or attractiveness of one modality did not predict that of another.

In Figure 4, we compare the Attractiveness results from the Full baseline with those from the partial baselines for Voice, Body, and Face. The largest correlation was between Full and Body (0.54), although it was not yet significant ($p < 0.1$), so more participants would be needed to explore this effect further.

Table II. Each of the Eight Voice/Body/Avatar Combinations for Creating FrankenFolks in Our Attractiveness and Distinctiveness Experiments. MD, LD, MA, LA Refer to Most and Least Distinctive and Attractive Resp

Distinctive			Attractive		
Voice	Body	Avatar	Voice	Body	Avatar
LDV	LDB	Fixed	MAV	MAB	MAA
MDV	MDB	Fixed	LAV	LAB	LAA
LDV	MDB	Fixed	LAV	MAB	MAA
MDV	LDB	Fixed	MAV	LAB	MAA
Distractor 1			MAV	MAB	LAA
Distractor 2			MAV	LAB	LAA
Distractor 3			LAV	MAB	LAA
Distractor 4			LAV	LAB	MAA

Note: The clips were selected so that MA = LD and LA = MD for each modality.

4.3 FrankenFolk

In this experiment, we combined modalities from different actors to create a new performance. We used the results from the partial baseline experiment to select for each mode a clip that had both a high attractiveness and low distinctiveness result and vice versa. In the case of model, we only selected the most and least attractive models, as they were all equally distinctive. We then generated our "FrankenFolks" by combining these modalities, as described in Section 3.3 and in Table II.

Twelve volunteers (5M/7F,20–65y) participated in this study, 11 of whom completed both conditions in counterbalanced order and one extra female for the Distinctiveness condition only. There were two repetitions of each stimulus combination. In the case of the Distinctiveness condition, we did not vary the model used, as in our partial baseline we determined that there was little or no variation in their distinctiveness. Instead, we again used the models shown in Figure 3.

This resulted in four FrankenFolk stimuli, as shown in Table II. To prevent a learning effect by repeated exposure to only four stimuli, we used an additional four distractor performances from the other full virtual performances of actors from whom no modalities were chosen. The eight Attractiveness combinations are also shown in Table II.

We performed a repeated-measures ANOVA for both conditions, each with within-subjects factors Sex (M,F) and Frank (8). For Attractiveness, we found a large main effect of Sex ($F1, 10) = 14.1$, $p < 0.005$, $\eta_p^2 = .58$), as Females were rated more attractive on average (4.6) than Males (3.95). We also found a large main effect of Frank ($F(7, 70) = 11.1$, $p \approx 0$, $\eta_p^2 = .53$). See Figure 6 (right) with all eight FrankenFolk. We found no interaction effect between Sex and Character, which means that the perceived attractiveness of each type of combination was consistent across sex.

For Distinctiveness, we found no main effects, but we did find a large interaction of Sex*Frank for distinctiveness ($F(7, 77) = 2.8$, $p < 0.05$, $\eta_p^2 = .20$). The results are summarized in Figure 6 (left) for the four FrankenFolk (we omit the four distractors). From a post hoc analysis using Fishers LSD tests as above, the male FrankenFolk with the most distinctive Voice were found to be more distinctive than the other two. However, whether it was paired with the most distinctive or least distinctive Body did not appear to have had a significant effect. We found no significant effects for the Female FrankenFolk characters. This is consistent with feedback from participants who said that the men were easier to recognize than the women and that they mainly relied on voices to make their decision. It is also

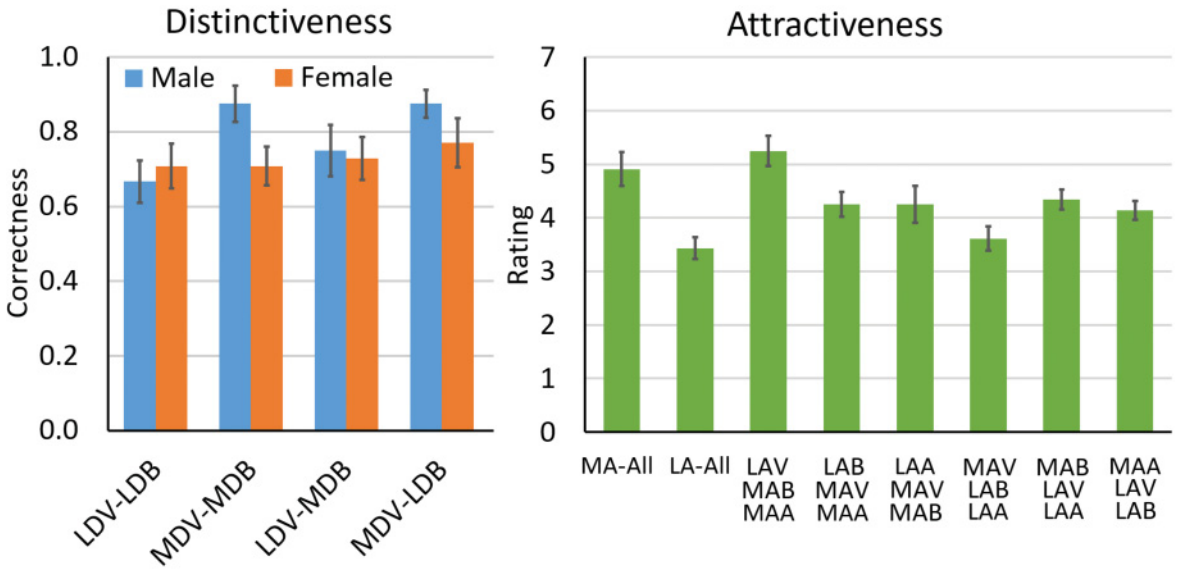


Fig. 6. FrankenFolks Results: Sex*Character interaction effect for Distinctiveness (l) and Character main effect for Attractiveness (r).

consistent with results from the Partial Baseline (Section 4.2, Figure 5), where we found that the Male audio was significantly more distinctive than the other modalities but not the female audio.

We can see that a combination of the most attractive Body motion and Avatar appears to be the best predictor of perceived attractiveness, whereas the Voice in this case is not. In fact, we found that the combination of the most attractive body and model with the least attractive voice appears to be just as attractive, if not more so, than the case where all are attractive. We should point out that there is a slight risk of a potential confound, in that the facial animation for each FrankenFolk was matched to the voice. However, we ensured that the facial clips for those actors were not among the most or least attractive. The influence of Body motion on attractiveness is also consistent with the results of the comparison between the Full and Partial baselines (Figure 4), where we found that the Body motion was most closely correlated (albeit not significantly) with the perceived attractiveness of the actor's Full performance. Model was not included in that comparison, because it is not related to any of the specific actors recorded; however, it is quite obvious that physical appearance would have a strong effect on perceived attractiveness. What is interesting, however, is that body pose/motion appears to have an equally significant effect.

5. CONCLUSIONS AND FUTURE WORK

We have performed a comprehensive evaluation of the attractiveness and distinctiveness of speaking performances. We can now answer the questions we asked in our introduction:

—***Q:*** *How distinctive and attractive is each modality (i.e., face or body motion, voice, physical appearance) when presented in isolation, that is, as a Partial performance?*
A: Physical appearance is the most distinctive feature, followed by Audio, then Body motion, then Face motion. However, results for attractiveness are less clear-cut: The female audio is found to be more attractive than the male.

—***Q:*** *How does each partial performance relate to the overall, Full performance?*

A: The full performances were all very distinctive, so a comparison of results was not meaningful with the partial performances for this condition. However, although not significant, the body motion attractiveness ratings were found to be most correlated with those for the full baseline, so further exploration of this question may be valuable. We also found that the body had the strongest effect on the perceived attractiveness of the male FrankenFolk.

—***Q:*** *What, if any, modality most strongly influences the attractiveness or distinctiveness of a character's performance?*

A: While further studies are needed to explore our observations in more detail, we found that the voice was the strongest indicator of distinctiveness but only for the males. This is consistent with the results of the partial baseline. Furthermore, the female audio was less distinctive on average than for males but was much more attractive, which is consistent with previous research. A combination of body motion and physical appearance influenced attractiveness, with the voice being less important in this case.

Our work is by no means conclusive and further studies would be valuable to explore our results further. For example, it is possible that the contribution of the facial motion was less than that of the body because of the smaller amount of screen space it occupied. We chose to display the face at the same scale as the body because this is how characters would be viewed in most practical applications. However, we did not explore the effect of the size of the character in this article, as the addition of this extra factor would have increased the scope and length of the experiments considerably, so this would be an interesting extension for future work. As discussed in Section 2, it has been shown that average faces, voices, and body motions are found to be less distinctive and more attractive than others. It would be interesting to explore whether this would be the case with the stimuli in our study.

We are wary of drawing any more general conclusions about our work with respect to human perception of conversing characters in general. As outlined in the introduction, the focus of our work is on the perception of virtual characters in animated groups and crowds, and the face and body motions have both gone through a retargeting process in order to be displayed. As we saw from the Virtual vs Real full baseline experiment, recognition of the Real stimuli was much higher than for the Virtual characters. Only the Voice has not been altered. Furthermore, we have tested only a relatively small number of actors (six females, six males), and a larger dataset would be needed to draw any more general conclusion about the interactions of the modalities in the perception of real humans talking. This would, however, be a very interesting direction to pursue in the future.

Our results will be useful for those seeking to create composite characters for movies and games. It is clear that the body motion, voice, and appearance of a virtual character all contribute to the overall perception of a performance, and care must be taken when combining them.

ACKNOWLEDGMENTS

The authors thank Tiarnán McNulty for helping with assets creation, all the reviewers for their comments, and the participants in our experiments. This work was partially sponsored by Science Foundation Ireland (Captavatar).

REFERENCES

Barbara Borkowska and Boguslaw Pawlowski. 2011. Female voice frequency in the context of dominance and attractiveness perception. *Anim. Behav.* 82, 1 (2011), 55–59.

Elizabeth J. Carter, Lavanya Sharan, Laura Trutoiu, Iain Matthews, and Jessica K. Hodgins. 2010. Perceptually motivated guidelines for voice synchronization in film. *ACM Trans. Appl. Percept.* 7, 4, Article 23 (July 2010), 12 pages.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Sarah A. Collins and Caroline Missing. 2003. Vocal and visual attractiveness are related in women. *Anim. Behav.* 65, 5 (2003), 997–1004.

Cathy Ennis, Ludovic Hoyet, Arjan Egges, and Rachel McDonnell. 2013. Emotion capture: Emotionally expressive characters for games. In *Proceedings of Motion on Games (MIG'13)*. 31:53–31:60.

Cathy Ennis, Rachel McDonnell, and Carol O'Sullivan. 2010. Seeing is believing: Body motion dominates in multisensory conversations. In *ACM SIGGRAPH 2010 Papers*. 91:1–91:9.

David R. Feinberg, Lisa M. DeBruine, Benedict C. Jones, and David I. Perrett. 2008. The role of femininity and averageness of voice pitch in aesthetic judgments of women's voices. *Perception* 37, 4 (2008), 615–623.

Paul J. Fraccaro, Jillian J. M. O'Connor, Daniel E. Re, Benedict C. Jones, Lisa M. DeBruine, and David R. Feinberg. 2013. Faking it: Deliberately altered voice pitch and vocal attractiveness. *Anim. Behav.* 85, 1 (2013), 127–136.

Carolyn R. Hodges-Simeon, Steven J. C. Gaulin, and David A. Puts. 2010. Different vocal parameters predict perceptions of dominance and attractiveness. *Hum. Nat.* 21, 4 (2010), 406–427.

Jessica Hodgins, Sophie Jörg, Carol O'Sullivan, Sang Il Park, and Moshe Mahler. 2010. The saliency of anomalies in animated human characters. *ACM Trans. Appl. Percept.* 7, 4 (2010), 22:1–22:14.

Ludovic Hoyet, Kenneth Ryall, Katja Zibrek, Hwangpil Park, Jehee Lee, Jessica Hodgins, and Carol O'Sullivan. 2013. Evaluating the distinctiveness and attractiveness of human motions on realistic virtual bodies. *ACM Trans. Graph.* 32, 6 (2013), 204:1–204:11.

Miyuki Kamachi, Harold Hill, Karen Lander, and Eric Vatikiotis-Bateson. 2003. Putting the face to the voice': Matching identity across modality. *Curr. Biol.* 13, 19 (2003), 1709–1714.

Karen Lander. 2008. Relating visual and vocal attractiveness for moving and static faces. *Anim. Behav.* 75, 3 (2008), 817–822.

Karen Lander, Harold Hill, Miyuki Kamachi, and Eric Vatikiotis-Bateson. 2007. It's not what you say but the way you say it: Matching faces and voices. *J. Exp. Psychol.: Hum. Percept. Perform.* 33, 4 (2007), 905.

Rachel McDonnell, Michéal Larkin, Simon Dobbyn, Steven Collins, and Carol O'Sullivan. 2008. Clone attack! Perception of crowd variety. In *ACM SIGGRAPH 2008 Papers (SIGGRAPH'08)*. ACM, New York, NY, 26:1–26:8.

Rachel McDonnell, Michéal Larkin, Benjamín Hernández, Isaac Rudomin, and Carol O'Sullivan. 2009. Eye-catching crowds: Saliency based selective variation. *ACM Trans. Graph.* 28, 3 (2009), 55:1–55:10.

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264 (1976), 746–748.

Martin Prazak and Carol O'Sullivan. 2011. Perceiving human motion variety. In *Proceedings APGV*. 87–92.

Gillian Rhodes and Tanya Tremewan. 1996. Averageness, exaggeration, and facial attractiveness. *Psychol. Sci.* 7, 2 (March 1996), 105–110.

Stefan R. Schweinberger, David Robertson, and Jürgen M. Kaufmann. 2007. Hearing facial identities. *Q. J. Exp. Psychol.* 60, 10 (2007), 1446–1456.

Galit Yovel and Pascal Belin. 2013. A unified coding strategy for processing faces and voices. *Trends Cogn. Sci.* 17, 6 (2013), 263–271.

Miron Zuckerman and Robert E. Driver. 1989. What sounds beautiful is good: The vocal attractiveness stereotype. *J. Nonverb. Behav.* 13, 2 (1989), 67–82.

Received May 2016; accepted May 2016