Le Test de Pieter Musk : interaction vocale en réalité virtuelle avec transformations temps réel de la voix du participant

Vincent Isnard, Trami Nguyen, Isabelle Viaud-Delmon

Le Test de Pieter Musk : interaction vocale en réalité virtuelle avec transformations temps réel de la voix du participant

Vincent Isnard[1], Trami Nguyen[2], Isabelle Viaud-Delmon[3]

[1] Institut de Recherche Biomédicale des Armées, Brétigny-sur-Orge, France
[2] Ensemble Links, Paris, France
[3] CNRS, Ircam, Sorbonne Université, Ministère de la Culture, Sciences et Technologies de la Musique et du son, STMS, F-75004 Paris, France

Résumé

Cette installation propose un nouveau type d'interaction, vocale, dédié à la réalité virtuelle (RV). Un scénario futuriste sert de prétexte au dialogue avec le personnage virtuel, Pieter Musk, incarnation anthropoïde d'une intelligence artificielle faisant passer au participant un test de Turing inversé. Celui-ci utilise alors sa propre voix pour répondre, laquelle est transformée en temps réel en timbre et en spatialisation, en correspondance avec la vidéo 360°. Dans la fiction, ce test quantifie le degré d'humanité du participant, tandis que des évaluations comportementales déterminent en parallèle la qualité de l'interaction. Globalement, le potentiel de ce nouveau vecteur narratif est confirmé par le ressenti très positif des participants, tout en favorisant l'extension de l'interactivité en RV à 360°.

Introduction

Parmi les dernières technologies de réalité virtuelle (RV), la vidéo immersive à 360° présente l'intérêt d'offrir une mise en œuvre relativement simple et accessible au grand public. Le résultat, obtenu à partir d'images et sons naturels, peut suffire à augmenter l'intensité d'une émotion spécifique et du sentiment de présence (e.g. Chirico et al., 2017). Cependant, cette technologie présente de fortes limitations d'interactivité, comme l'absence de déplacements en translation dans l'environnement virtuel.

Une interaction vocale, à la manière des « agents conversationnels animés » (e.g. Potdevin et al., 2018 ; voir aussi *Virtual Corporation*, 1996 : un jeu vidéo singulier se jouant avec la voix), pourrait permettre de compenser ces limites tout en explorant de nouvelles formes narratives sonores. Sans avoir à ajouter de sources visuelles dans l'environnement virtuel, la voix pourrait provoquer des évènements favorisant l'interactivité et, par suite, l'immersion et la qualité de l'expérience en RV (Slater & Wilbur, 1997). Enfin, des transformations vocales originales en temps réel de la propre voix du participant, incarnant directement un personnage de la fiction comme une nouvelle source narrative à appréhender au sein du monde virtuel, pourraient également amplifier le ressenti de l'expérience.

Ici, un scénario futuriste sert de prétexte à un échange dialogué entre une intelligence artificielle (IA) fictive, incarnée par un acteur, et le participant. La voix du participant est transformée en temps réel pour tester si l'interaction reste crédible même lorsque son personnage possède des qualités sonores divergeant totalement d'une voix naturelle (e.g. timbre, spatialisation). Bien plus, ces transformations semblent avoir favorisé l'interaction et l'immersion en RV.

Méthodes

Scénario et conditions expérimentales

Le scénario, basé sur un test de Turing inversé sur le modèle du test fictif de Voight-Kampff (Dick, 2014), met en scène Pieter Musk, une IA d'apparence humaine, interrogeant le participant pour déterminer son degré d'humanité et le faire réfléchir sur sa condition humaine à l'heure des augmentations technologiques (Frischmann & Selinger, 2018). Les 16 questions proposées sont volontairement dérangeantes pour susciter une réaction émotionnelle propre à un comportement considéré comme normal chez un humain (cf. Isnard & Nguyen, 2020). Les 3 réponses possibles correspondent respectivement à un comportement « humain », « post-humain » ou « machine ». Un « score d'humanité » est affiché à la fin de l'expérience en fonction des réponses du participant, pour une conclusion ludique et conforme à la trame du scénario, sans rentrer en compte dans l'analyse scientifique.

Le contenu vocal est contrôlé avec des échanges calibrés en nombres de syllabes (174.0 ± 7.7 syllabes par question ; 16.6 ± 1.4 syllabes par réponse à lire par le participant), pour une vidéo d'une durée totale de 18 min 28 s. La qualité de l'interaction vocale est testée en fonction de distorsions sonores et visuelles, proposées à 2 niveaux (faible ou fort) et appliquées : (1) au « timbre » : voix robotiques et distorsions visuelles ; (2) à la « spatialisation » : cohérence entre la source originale et perçue (i.e. faible : pas d'effet de spatialisation surajouté ; forte : voix du participant déplacée 3 m devant lui et dissociations colorimétriques dynamiques appliquées à la vidéo ; cf. Tab. 1 et Fig. 1).

Echanges vocaux		« Timbre »	« Spatialisation »
Introduction		*Faible*	*Faible*
Questions	1 à 4	Faible	Faible
	5 à 8	Faible	Forte
	9 à 12	Forte	Faible
	13 à 16	Forte	Forte
Conclusion		*Forte*	*Forte*

Tableau 1. Transformations audiovisuelles au cours du scénario. Le participant n'intervient pas dans l'introduction et la conclusion.

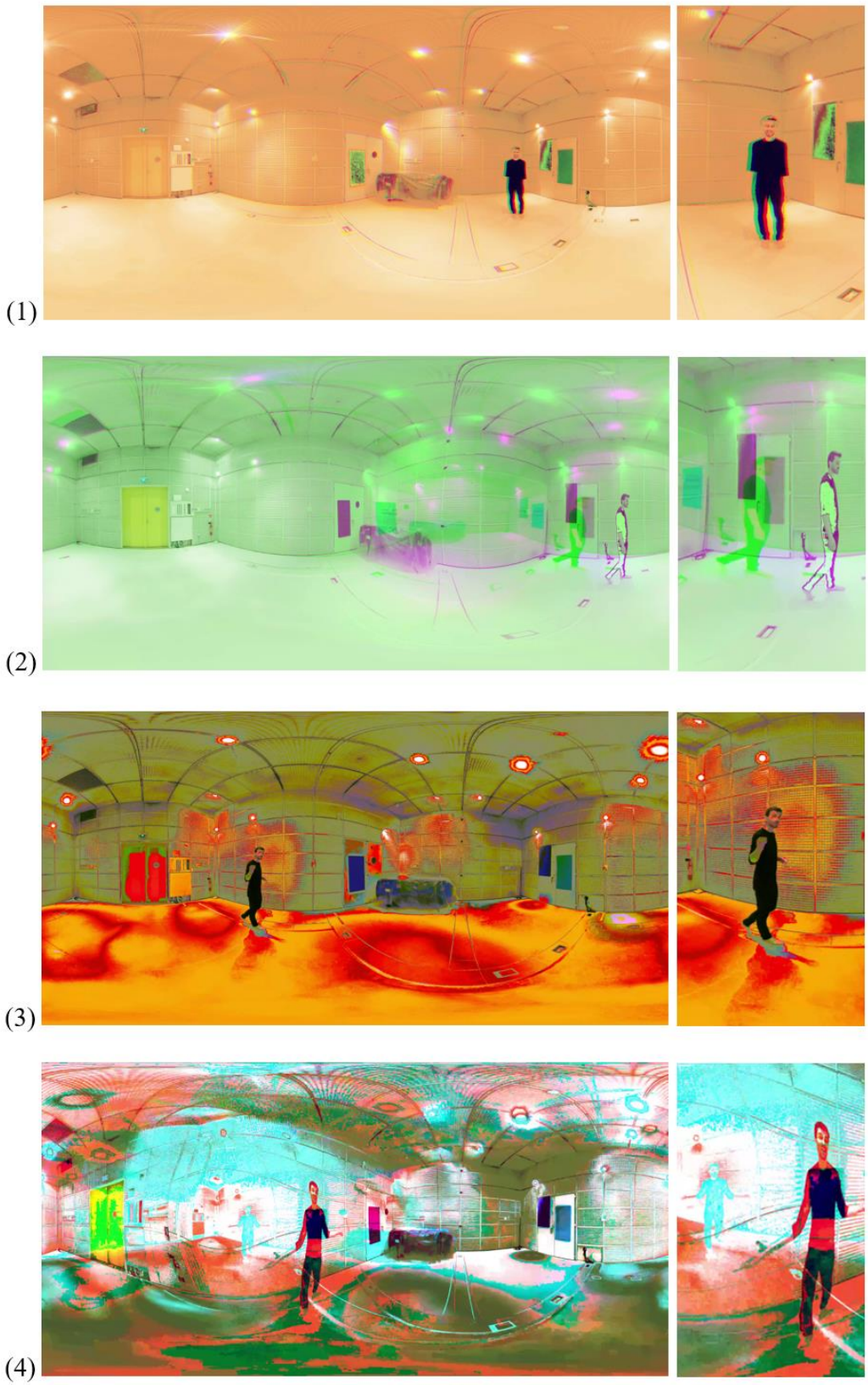

Figure 1. Rendu visuel 360° pour chaque condition de transformations audiovisuelles. Distorsions de : (1) « timbre » faible, « spatialisation » faible ; (2) « timbre » faible, « spatialisation » forte ; (3) « timbre » forte, « spatialisation » faible ; (4) « timbre » forte, « spatialisation » forte. Les figures à droites sont un agrandissement du personnage virtuel que devait suivre le participant.

Matériel

Les vidéos 360° ont été enregistrées avec un acteur dans un studio de l'IRCAM avec une caméra Insta360 Pro 2, avant d'être stitchées, montées et traitées dans Adobe Premiere Pro 2019. L'audio a été capté en ambisonique avec un microphone sphérique MH Acoustics Eigenmike (32 canaux, 24 bits, Fe = 44.1 kHz), puis monté et normalisé dans Reaper 5. Tous les bruits parasites liés à l'acteur (e.g. respirations) ont été coupés pour accentuer l'aspect numérique fictif du personnage. La restitution synchrone de l'image et du son est effectuée, sur un PC dédié à la RV, grâce à un patch Max 8 intégrant la bibliothèque « VR » pour la restitution 360° de l'image, tandis qu'un rendu sonore 3D binaural est obtenu grâce à la bibliothèque « Spat ».

Les participants sont équipés d'un visiocasque Oculus Rift CV1 et d'un casque audio Beyerdynamic DT 770, connecté à une carte son RME Fireface UC, ainsi que des 2 manettes Oculus Touch pour le déroulement de l'expérience. Leur voix est captée à l'aide d'un microphone DPA 4066, connecté à la carte son, pour être traitée en temps réel dans le même patch (spatialisation, réverbération et filtrage simulant l'écoute de notre propre voix sans microphone ; Pörschmann, 2000).

Procédure

Après chaque question du personnage, les participants doivent lire à voix haute la réponse choisie avant d'effectuer l'évaluation perceptive suivante : « Les traitements sur votre voix favorisent-ils votre interaction avec le personnage virtuel ? ». Ils déplacent alors un curseur entre « Pas du tout » et « Tout à fait ». Il leur a été précisé à l'oral que cette évaluation concerne l'immersion en RV et la cohérence de leur incarnation virtuelle dans le contexte de la fiction. L'ensemble de l'expérience en RV dure environ 30 min. Ensuite, les participants complètent un formulaire pour donner leur avis sur l'expérience.

Résultats

Les données pilotes de 8 participants (voir Fig. 2) suggèrent que ceux-ci ont globalement apprécié l'expérience, malgré les contraintes du dispositif d'évaluation expérimentale surajouté au scénario de fiction. L'évaluation moyenne obtenue sur l'ensemble des conditions est de 62/100 (échelle arbitraire). Cette interprétation est corroborée par leurs commentaires : e.g. « beaucoup apprécié », « surprenant », « intéressant ». Par ailleurs, l'appréciation de l'expérience semble être d'autant plus forte pour les participants déjà familiers avec la RV ou les nouvelles technologies.

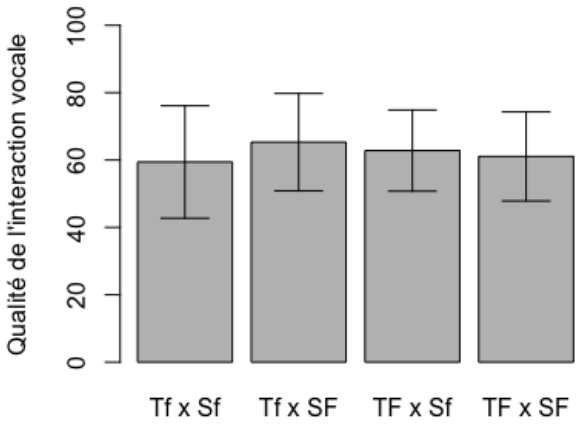

Figure 2. Evaluation perceptive de la qualité de l'interaction vocale. Les types et niveaux de distorsions sont notés : Tf = « timbre » faible ; TF = « timbre » fort ; Sf = « spatialisation » faible ; SF = « spatialisation » forte.

Discussion

Ces données doivent être complétée par de nouvelles évaluations pour confirmer les résultats. De plus, même pour cette expérience en RV comportant a priori peu d'évènements susceptibles de provoquer de la cybercinétose, le temps passé sous visiocasque reste un facteur primordial de son acceptabilité, bien qu'elle faisait déjà ici l'objet d'un compromis avec une forte limitation des conditions testées (e.g. 1 seule évaluation perceptive sous visiocasque).

Malgré les contraintes expérimentales, l'intégration de l'expérience dans un cadre de fiction est très appréciée et tolérée par les participants, et les résultats obtenus permettent de donner une première estimation des possibilités d'interaction vocale en RV. Cette expérience doit évoluer avec des interactions vocales plus libres, en contrôlant a posteriori la quantité de parole pour donner aussi une indication sur l'implication du participant ; ainsi qu'une interactivité accentuée en modifiant le cours du scénario en fonction des réponses du participant. D'autres facteurs seront approfondis : e.g. familiarité, étrangeté, jugement d'appartenance de sa propre voix (Kimura & Yotsumoto, 2018) ; acceptation de l'interaction avec un personnage virtuel (Eyssel et al., 2012).

Plus étonnamment, il n'a pas été observé ici de différence en fonction des effets appliqués à la voix et de leur intensité : l'intervalle d'intensité des effets sera donc accru pour tenter d'y remédier. Un contrôle sera notamment effectué avec leur propre voix sans transformation, et inversement, avec de fortes transformations pour examiner jusqu'à quel point celles-ci sont tolérées sans abolir le jugement d'appartenance. De plus, des mouvements dans la spatialisation sonore seront ajoutés afin de favoriser l'externalisation de leur propre voix, pour examiner la contribution de l'internalisation à l'immersion ou au sentiment de présence (Slater & Wilbur, 1997 ; Nichols et al., 2000).

Enfin, d'après les profils des participants, ceux avec une connaissance de la RV ou des nouvelles technologies, même minimale, semblent avoir davantage apprécié l'expérience que les participants plus naïfs, tandis que leur expertise aurait pu au contraire les amener à des critiques des contraintes du dispositif expérimental. Cela confirme donc son potentiel. Le profil des participants en fonction de leur familiarité avec la RV sera plus systématiquement étudié avec une étape préalable de

familiarisation avec la RV, pour vérifier si l'appréciation augmente en étant rendu plus disponible pour l'expérience immersive et les nouvelles fonctionnalités de l'interaction vocale.

Cette installation en RV ouvre la voie à d'autres expériences impliquant la voix du participant pour des créations immersives artistiques et/ou musicales, en environnements sonores spatialisés et transformés en temps réel.

Remerciements

Cette installation résulte d'un projet art/science soutenu par l'IRCAM et le GRAME dans le cadre de résidences complémentaires en 2019-2020. Une partie du protocole et des résultats a été présentée lors du Forum IRCAM 2020.

Références

Chirico, A., Cipresso, P., Yaden, D. B., Biassoni, F., Riva, G., & Gaggioli, A. (2017). Effectiveness of immersive videos in inducing awe: an experimental study. *Scientific Reports*, *7*(1), 1-11.

Eyssel, F., De Ruiter, L., Kuchenbrandt, D., Bobinger, S., & Hegel, F. (2012, March). 'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 125-126). IEEE.

Dick, P. K. (2014). *Blade Runner. Les Androïdes rêvent-ils de moutons électriques ?*, J'ai Lu.

Frischmann, B., & Selinger, E. (2018). *Re-engineering humanity*. Cambridge University Press.

Kimura, M., & Yotsumoto, Y. (2018). Auditory traits of "own voice". *PloS one*, *13*(6), e0199443.

Isnard, V., & Nguyen, T. (2020). *L'étrangeté perceptive en réalité virtuelle.* Forum IRCAM.

Nichols, S., Haldane, C., & Wilson, J. R. (2000). Measurement of presence and its consequences in virtual environments. *International Journal of Human-Computer Studies*, *52*(3), 471-491.

Potdevin, D., Clavel, C., & Sabouret, N. (2018, November). Virtual Intimacy, this little something between us: A study about Human perception of intimate behaviors in Embodied Conversational Agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (pp. 165-172).

Pörschmann, C. (2000). Influences of bone conduction and air conduction on the sound of one's own voice. *Acta Acustica united with Acustica*, 86(6), 1038-1045.

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments, 6*(6), 603-616.