Vers un modèle du « toucher vocal » pour la communication ubiquïte

Ambre Davat[1, 2] Véronique Aubergé[2] Gang Feng[1]
(1) Univ. Grenoble Alpes, Grenoble INP, GIPSA-lab, F-38040 Grenoble, France
(2) CNRS, Grenoble INP, LIG, 38000 Grenoble, France
ambre.davat@gipsa-lab.grenoble-inp.fr, veronique.auberge@univ-grenoble-alpes.fr, gang.feng@gipsa-lab.grenoble-inp.fr

RESUME

Un des enjeux de la robotique de téléprésence est d'offrir une immersion socio-relationnelle ubiquïte. Pour y parvenir, il est nécessaire de comprendre et modéliser les facteurs qui permettraient au téléopérateur de contrôler la transmission de ses productions vocales, afin de lui en donner la perception, la proprioception et l'inter-proprioception. Ce modèle de transfert de la distance vocale socialement incarnée devra tenir compte de l'ensemble des paramètres impliqués dans l'effet socio-relationnel des productions vocales, en particulier l'intensité. Il devra également intégrer des éléments de contexte pertinents à la performance intentionnelle du locuteur téléopérant. Nous présentons dans cet article une première expérience visant à mesurer, analyser et modéliser la manière dont l'humain perçoit la distance physique qui le sépare d'un interlocuteur, en fonction de variations socio-affectives dans les productions vocales de cet interlocuteur. Ces résultats seront la référence des modèles qui seront implantés sur notre robot de téléprésence : RobAir Social Touch.

ABSTRACT

Towards a model of "social touch" for ubiquitous communication

One of the challenges of telepresence robotics is to provide ubiquitous social-interpersonal immersion. In order to achieve this, there is a need to understand and model the factors that would allow the users to control the transmission of their vocal productions, and to give them perception, proprioception and inter-proprioception of this control. This model for transferring socially embodied vocal distance should take into account all parameters involved in the social-interpersonal effect of vocal productions, especially intensity. It should also integrate the background information which is relevant for the speakers to express their intentions. In this paper, we present a first experiment for measuring, analyzing and modeling how human beings perceive the distance to an interlocutor, depending on socio-affective variations in the vocal productions of this interlocutor. These results will be the reference for the models which will be implanted in our telepresence robot: Robair Social Touch.

MOTS-CLES : toucher social, robotique de téléprésence, psychoacoustique, perception, socio-affects, expérimentation écologique.

KEYWORDS: social touch, telepresence robotics, psychoacoustics, perception, socio-affects, ecological experimentation.

10.21437/JEP.2018-35

1 Introduction

Après l'invention de la téléphonie, qui permet de transporter la voix de quelqu'un d'un endroit à un autre, puis celle de la visiophonie, qui transmet également l'image, la robotique de téléprésence constitue une nouvelle étape dans l'immersion ubiquïte. En effet, il ne s'agit plus simplement de transporter la voix et l'image d'une personne vers un point de l'espace fixé par ses interlocuteurs, mais de permettre à l'utilisateur de contrôler ce point et de le déplacer dans l'espace distant dans lequel il communique à travers le « corps » physique d'un artefact robotique.

Ces robots de téléprésence reposent sous un nouvel angle la question de la fidélité avec laquelle la parole est transmise grâce aux moyens de télécommunication, question qui ne se pose plus pour le téléphone ou la visiophonie car nous avons culturellement intégré leurs artefacts dans nos usages. Ainsi par exemple, lors d'une communication téléphonique classique, les participants ont l'habitude de régler le volume sonore de leur interlocuteur au niveau qu'ils jugent acceptable. Une personne qui utilise un téléphone n'a donc pas à se préoccuper de savoir si sa voix est forte ou faible dans l'environnement distant car ce contrôle est clairement de la responsabilité de l'interlocuteur. En revanche, le pilote d'un robot de téléprésence peut être considéré par ses interlocuteurs comme maître de ce contrôle et donc responsable des artefacts liés à la télécommunication, d'autant plus que ses interlocuteurs sont réticents à l'idée de modifier les réglages du robot, car il est perçu non plus comme un simple objet de télécommunication, mais bien comme une personne physiquement présente.

Les systèmes proposés à l'heure actuelle dans la littérature pour tenter de résoudre ce problème ne nous satisfont pas entièrement. Ainsi, le feedback audio proposé par (Paepcke et al., 2011) peut effectivement inciter le téléopérateur à parler plus doucement, mais ne garantit pas que sa voix soit adaptée à l'environnement distant. Le système de (Takahashi et al., 2015) adapte automatiquement le volume sonore du robot en fonction du bruit ambiant et de la distance de l'interlocuteur, mais le contrôle de l'utilisateur est très limité, puisqu'il n'a le choix qu'entre deux modes de régulation : un mode « confortable » et un mode « secret » avec lequel seule une personne proche du robot peut l'entendre. Enfin, les interfaces visuelles proposées par (Kimura et al. 2007) permettent à l'utilisateur de visualiser le volume sonore perçu par son interlocuteur, mais semblent peu adaptées à la téléconférence mobile.

La difficulté à développer ce type de systèmes vient du fait que les productions vocales dépendent de multiples éléments de contexte (Cooke, 2014). Ainsi, le signal qui serait adapté dans l'environnement du téléopérateur, ne l'est pas forcément dans celui où se trouve le robot. Par exemple, si le téléopérateur se trouve dans une pièce calme, sa voix transmise par le robot ne sera peut-être pas suffisamment forte pour être audible dans un environnement bruyant. Il est donc intéressant de pouvoir l'amplifier artificiellement, en ayant conscience toutefois que cette correction peut engendrer des artefacts sonores, des voix « schizophoniques » qui ne pourraient pas être produites naturellement (Schafer, 1977). Or, cette schizophonie a sans doute un impact sur l'effet socio-relationnel des signaux vocaux. En effet, elle est notamment utilisée par la publicité, le cinéma et la musique afin de suggérer une certaine distance sociale entre l'auditeur et le locuteur (Maasø, 2008). Par exemple, un enregistrement qui parvient à capturer les bruits de bouche et la respiration du locuteur suggère une forme d'intimité à l'auditeur, qui ne devrait pas être capable de percevoir ces sons sans qu'ils soient extrêmement proches physiquement (Collins, Dockwray, 2015).

Pour améliorer notre robot de téléprésence, Robair Social Touch, nous cherchons à développer une interface qui permette à l'utilisateur de contrôler non seulement l'intensité de sa voix dans l'environnement distant, mais surtout son impact social, ou « toucher vocal ». Pour ce faire, nous avons besoin d'un modèle qui lie ce toucher vocal à un jeu de paramètres physiques mesurables (éléments de contexte et caractéristiques de la voix). Notre hypothèse est qu'il est possible de définir ce toucher vocal en étudiant la manière dont la perception acoustique de l'espace peut être influencée socialement. En effet, (Gardner, 1969), (Brungart, Scott, 2001) ou encore (Philbeck, Mershon, 2002) ont montré que la manière dont un sujet perçoit à l'aveugle la distance qui le sépare d'une source de parole dépend de l'effort vocal utilisé au moment de la production du stimulus. Ainsi, un chuchotement est systématiquement perçu plus proche qu'il ne l'est réellement, tandis qu'un cri est perçu comme plus éloigné. La réponse des sujets est donc influencée par la distance de communication imaginée par le locuteur au moment de l'enregistrement du stimuli, et non simplement par la distance physique des sources sonores. Il y a donc bien un biais social dans la perception acoustique de l'espace, sur lequel nous allons nous appuyer pour établir notre modèle. Dans ce papier, nous présenterons le protocole d'expérimentation que nous avons conçu puis mis en œuvre, et dont la complexité est relative à la complexité du problème posé. Enfin nous présenterons nos premiers résultats.

2 Scénario expérimental en « caméra cachée »

Le cœur de notre protocole consiste à demander à un sujet d'estimer la position spatiale de son interlocuteur, pendant que celui-ci exprime différents socio-affects. Cependant, il est fondamental que le sujet ne soit pas conscient que nous mesurons sa capacité à repérer l'espace physique en fonction de variations socio-affectives des stimuli, car alors il procéderait à des méta-traitements cognitifs qui ne sont pas ceux que nous souhaitons mesurer directement. Pour assurer cette démarche d'observation en situation écologique, nous avons dû monter un scénario de type « caméra cachée » et baser notre expérience sur une tâche prétexte. De plus, comme ce sont les performances de régulation sociale humaine qui nous intéressent, nous avons décidé, sur ces deux contraintes, d'accepter les variabilités de production des stimuli inhérentes à l'écologie naturelle de production des humains interactants. Ainsi, dans cette expérience, la source sonore n'est pas un haut-parleur, mais un locuteur expert, dont les productions vocales seront analysées a posteriori pour contrôler leur régularité.

En pratique, les sujets sont recrutés pour passer une expérience sur la perception du goût et de l'odorat. Cette expérience est censée faire partie d'un projet franco-japonais qui s'intéresse à la dimension culturelle et sociale des saveurs et des odeurs. Elle se déroule en binôme : un des sujets doit goûter des mini-pilules gustatives, l'autre respirer des boîtes à odeurs. Avant l'expérience, ils doivent remplir un questionnaire concernant leur pratique gustative et olfactive, et qui permet de recueillir des informations sur leur accent. Au moment où le sujet (S) se présente pour passer l'expérience, il rencontre le locuteur expert (L), présenté comme le second sujet de l'expérience. L se fait passer pour un nez professionnel, ce qui justifie qu'il puisse produire des énoncés à la fois très autoritaires et très hésitants. Un expérimentateur (E) présente l'expérience, en s'appuyant sur les questions posées par L afin de convaincre S qu'il passe bien une expérience sur le goût et l'odorat.

S comprend ainsi que chaque mini-pilule est assortie à une boîte à odeur. A chaque étape, L goûtera une mini-pilule disposée dans des gobelets répartis sur deux rangées de tables et annoncera ce qu'il a reconnu. Ensuite, E donnera une boîte à odeur à respirer à S, qui devra alors discuter avec L jusqu'à ce que chacun donne un avis définitif. Les deux participants seront placés dans différentes

situations d'interaction, soit disant pour observer des variations de leurs capacités perceptives. Afin d'éviter qu'ils puissent lire sur le visage de l'autre des informations de plaisir ou de déplaisir qui pourraient influencer leur perception, S devra porter un masque qui l'aveugle et dissimule le bas de son visage. Il passera donc l'expérience assis sur une chaise au centre de la pièce, à côté de E. Cependant, cette situation serait très inconfortable pour L, qui aurait l'impression de parler à quelqu'un qui l'ignore. Ainsi, pour s'assurer que S reste concentré sur sa tâche et rassurer L, on demande à S d'indiquer avant chaque boîte à odeur :

- la direction dans laquelle se trouve L (avant, gauche, derrière ou droite)
- sa distance (devant la première table, entre les deux tables ou derrière la deuxième table)
- son orientation (face au sujet, ou dos au sujet)

Par ailleurs, L reçoit un ordre de passage indiquant les positions des gobelets et leur numéro. Au dos de cette feuille se trouve les indications pour la vraie expérience, à savoir :

- la direction (avant, gauche, derrière ou droite)
- la distance (proche, milieu ou loin)
- l'orientation (face au sujet, ou dos au sujet)
- l'intensité à produire (faible ou forte)
- le socio-affect à communiquer (confiance autoritaire ou doute poli)
- le mot-clé à prononcer (ex : pomme, orange...)

Entre chaque étape, de la musique est diffusée par des hauts parleurs placés derrière les oreilles du sujet pour le faire patienter et dissimuler les pas du locuteur expert. Le sujet et le locuteur expert portent un micro serre-tête Sennheiser HSP 4 relié à un émetteur radio afin que leur échange soit enregistré.

FIGURE 1 : Photo du dispositif expérimental

A la fin de l'expérience, un débriefing est effectué. Il permet d'abord de vérifier que le sujet n'a pas deviné le but réel de l'expérience. Ensuite, la supercherie est dévoilée, et on s'assure que le sujet a bien compris les buts de l'expérience pour qu'il puisse donner son consentement éclairé.

Il s'agit d'une expérience lourde à monter, à mettre au point, puis à reproduire pour chaque sujet, puisqu'elle dure environ 1h30, introduction et débriefing compris. Il est donc important de noter que les 6 sujets déjà enregistrés, ainsi que plusieurs sujets préalables, non conservés tant que la mise au point n'était pas stabilisée, n'ont pas montré de signe d'ennui, de désintérêt ou de charge cognitive trop forte par rapport à la tâche prétexte ; en outre, nous n'avons observé a priori ni d'effet d'apprentissage, ni de dégradations des réponses des sujets au cours de l'expérience (nous le vérifierons statistiquement quand nous aurons plus de sujets). Nous allons à présent détailler la manière dont ce dispositif expérimental a été choisi.

3 Dispositif expérimental

L'expérience se déroule sur la plateforme d'expérimentation Domus, du LIG. Il s'agit d'une salle de 7.1 x 8.6 m, que nous avons aménagé avec des paravents pour former un espace carré de 7.1 x 7.1 m, soit 10 m en diagonale. Nous avons mesuré un temps de réverbération de 0.8 s[1]. C'est donc une salle particulièrement réverbérante, dans laquelle une simple mesure de l'intensité acoustique ne permet pas de deviner la distance d'une source sonore. Une photo du dispositif expérimental apparaît en Figure 1.

3.1 Choix des distances

Pour définir les distances mises en jeu dans cette expérience, nous nous sommes intéressés à l'incertitude avec laquelle un sujet estime la distance d'une source sonore. En effet, la psychoacoustique a montré que lorsqu'on demande à un sujet d'estimer à plusieurs reprises la distance d'une même source sonore, ses réponses varient. En choisissant des distances suffisamment proches les unes des autres, il est donc possible d'induire le sujet en erreur, et donc peut-être d'observer un biais socio-affectif dans ses réponses. Au contraire, si les distances mises en jeu sont trop différentes, le locuteur n'a aucune difficulté à deviner si son interlocuteur est devant, derrière, ou entre les deux tables. Or, cette incertitude dépend probablement des caractéristiques acoustiques de la salle où se déroule l'expérience, et varie d'un sujet à l'autre. Ainsi, (Zahorik et al., 2005) évoquent un flou perceptif évalué entre 5 et 20 % de la distance effective d'après un article de Haustein de 1969, et entre 20 et 60 % lorsque les distances sont représentées en échelle logarithmique d'après réanalyse de leurs propres travaux. Dans un autre article, (Calcagno, Abregú, 2012) indiquent l'écart type des distances estimées par leurs sujets. Celui-ci augmente linéairement dans le cas où les sujets doivent estimer la distance des sources sonores à l'aveugle. A titre indicatif, l'écart type est d'environ 40 cm pour une source située à 2 m, lorsque les sujets ont eu l'occasion de voir la salle avant le début du test. Par ailleurs, (Anderson, Zahorik, 2014) montrent que l'erreur de jugement des sujets suit une distribution normale lorsque la distance est représentée en échelle logarithmique.

Pour rester dans cette zone de flou perceptif, nous avons dû fabriquer de petites tables de 20 x 60 cm à partir de panneaux de bois posés sur des trépieds. Les 8 tables sont disposées sur 2 rangées, respectivement à 2 et 3 m de l'emplacement du sujet. Le locuteur se place donc à 1 m 70, 2 m 50 ou 3 m 30 du sujet. Non seulement cet aménagement convient parfaitement aux dimensions de la pièce, mais il est également intéressant en termes perceptifs. D'une part, il permet d'étudier le mode proche et le mode éloigné de la sphère sociale définie selon la théorie proxémique (Hall, 1966). D'autre part, les études psychoacoustiques ont montré que nous avons tendance à surestimer la distance des sources sonores proches et à sous-estimer celles des sources éloignées. La distance à laquelle s'inverse la tendance dépend des propriétés acoustiques de la pièce, mais semble être dans l'ordre de grandeur des distances que nous utilisons. Ainsi, (Anderson, Zahorik, 2014) ont observé un point d'inflexion à 1 m 90, puis à 3 m 22 dans une salle plus réverbérante, dans laquelle les distances sont perçues plus éloignées.

[1] Il s'agit du temps nécessaire pour que l'intensité du son diminue de 60 dB.

3.2 Choix des directions et orientations

Dans cette expérience, il ne s'agit pas de vérifier si nos sujets sont capables d'évaluer la direction d'arrivée des sons, mais de pouvoir éventuellement observer des biais perceptifs différents selon la manière dont le sujet et son interlocuteur sont positionnés dans l'espace. Parler en étant rigoureusement face à face avec quelqu'un, ce n'est pas la même chose que de lui parler de profil, ou même de dos. En effet, l'orientation du locuteur a d'abord une conséquence acoustique, car la voix humaine a une directivité : au lieu de se propager uniformément dans toutes les directions de l'espace, la puissance acoustique se répartit sous une forme cardioïde (Chu, 2002). En particulier, les hautes fréquences de la voix sont particulièrement atténuées derrière la tête du locuteur. En conséquence, un auditeur est capable de deviner à l'aveugle l'orientation de la tête d'un locuteur (Edlund et al., 2012). Par ailleurs, l'orientation du corps a un sens social ; elle est donc généralement prise en compte dans les études sur la proxémie, soit directement, par exemple comme dans (Remland et al.), soit indirectement, lorsque c'est le regard des interlocuteurs qui est analysé. Pour limiter la durée de l'expérience, l'orientation de L ne varie que lorsqu'il est entre les deux tables. Dans les autres cas, L est toujours tourné vers le sujet.

3.3 Choix des distances sociales des productions vocales

L doit être capable d'exprimer deux socio-affects différents : une confiance autoritaire par laquelle il marque une distance sociale grande et de dominance avec son interlocuteur, et un doute poli, par lequel il se rapproche socialement de son interlocuteur et inverse la dominance. Intrinsèquement à la nature prosodique de ces énoncés, L a tendance à parler plus fort pour exprimer la confiance autoritaire, et plus doucement pour exprimer le doute. Nous avons donc ajouté une consigne de contrôle de son intensité de production, faible ou forte, sur chacune des deux variables socio-affectives, afin d'observer si celle-ci a un impact sur la perception de l'auditeur.

4 Analyse des premiers résultats

Notre objectif est de faire passer 20 sujets sur cette expérience. Nous présentons ici les résultats obtenus avec nos 6 premiers sujets.

4.1 Vérification de la régularité de production du locuteur expert

Il est important de vérifier que L arrive à produire les différents socio-affects et intensités demandés, tant au niveau des contenus attitudinaux que des réalisations acoustiques.

Pour pouvoir faire une mesure étalonnée en décibels de l'intensité acoustique produite par L, nous aurions besoin d'un sonomètre placé à une distance fixe de sa bouche, ce qui est incompatible avec notre dispositif expérimental. Nous nous contentons donc de mesurer l'intensité des signaux numériques enregistrés durant l'expérience, ce qui est suffisant pour pouvoir les comparer entre eux. Pour ce faire, nous avons choisi la procédure suivante, réalisée par un script Praat :

1. extraction du pitch et de l'intensité du mot-clé (il s'agit bien sûre de l'intensité de l'enveloppe, et pas de l'intensité instantanée)
2. échantillonnage toutes les 10 ms du pitch et de l'intensité
3. moyennage des intensités pour lesquelles le pitch est défini

Les intensités mesurées à partir des enregistrements du micro porté par L sont présentées dans le Tableau 1. En moyenne, il y a bien un écart de plus de 7 dB entre les mots qui devaient être prononcés avec une intensité faible, et ceux devant être prononcés avec une intensité forte. Même si l'écart type ainsi que les valeurs min et max indiquent que l'intensité de certains mots-clés ne convient pas à leur catégorie, L arrive donc le plus souvent à suivre les consignes concernant les variations d'intensité. Cette méthode de mesure est néanmoins sensible à la position du micro, qui varie d'une expérience à l'autre, voire au cours d'une même expérience. En effet, ce micro n'est situé qu'à 2-3 cm de la bouche du locuteur, donc la moindre variation de son écartement fait varier l'intensité mesurée. Pour avoir accès à une seconde mesure d'intensité, nous avons ajouté un micro supplémentaire, placé à la verticale du sujet.

Attitude	Doute poli		Confiance autoritaire	
Intensité	Faible	Forte	Faible	Forte
Min	49,6	57,8	51,1	58,9
Max	60,2	68,0	62,5	67,3
Moyenne	54,8	62,9	57,0	63,5
Ecart type	2,9	2,9	3,3	2,5

TABLEAU 1 : Mesure de l'intensité (dB) des mots-clés prononcés par le locuteur expert

Un test perceptif devra également être effectué pour confirmer que les attitudes sont bien reconnues, quelle que soit l'intensité utilisée. Nous n'avons pas encore fait d'analyse rigoureuse des différents mots-clés en terme de qualité de voix et de prosodie, mais nous avons déjà pu remarquer qu'il est difficile pour L d'exprimer une confiance autoritaire d'une voix faible et un doute poli d'une voix forte. A l'oreille, la stratégie qui semble efficace consiste dans un cas à parler très vite pour avoir une voix sèche, et dans l'autre à faire trainer les mots-clés et leur donner une tournure interrogative.

4.2 Premières observations

Une première manière d'analyser les réponses données par les sujets est d'étudier leur répartition en fonction des cinq variables étudiées à l'aide de matrices de confusion (Figure 2). Par manque d'espace, nous ne représenterons pas les résultats liés à la direction, car outre les confusions avant/arrière régulièrement observées en psychoacoustique, les sujets se trompent rarement pour estimer la direction. Pour simplifier la mise en page, les deux facteurs proxémiques et les quatre variables retenus sont présentés au même niveau, bien qu'a priori ils ne contribuent pas de façon équivalente au toucher vocal. Des tests du χ^2 avec un seuil de tolérance de 5% ont été utilisés pour déterminer parmi ces huit matrices de confusion celles qui présentent des résultats significatifs, non explicables par le hasard.

Concernant la perception de la distance, on constate que les sujets perçoivent L comme plus éloigné lorsqu'il leur tourne le dos. En revanche, il n'y a pas d'effet notable du socio-affect ou de l'intensité. Par ailleurs, on note que les sujets sont indécis lorsqu'ils doivent estimer l'orientation de L, mais se trompent rarement lorsque L leur tourne le dos. Ils ont tendance à répondre que L leur tourne le dos lorsque le socio-affect indiqué à L est « doute poli », et l'intensité faible.

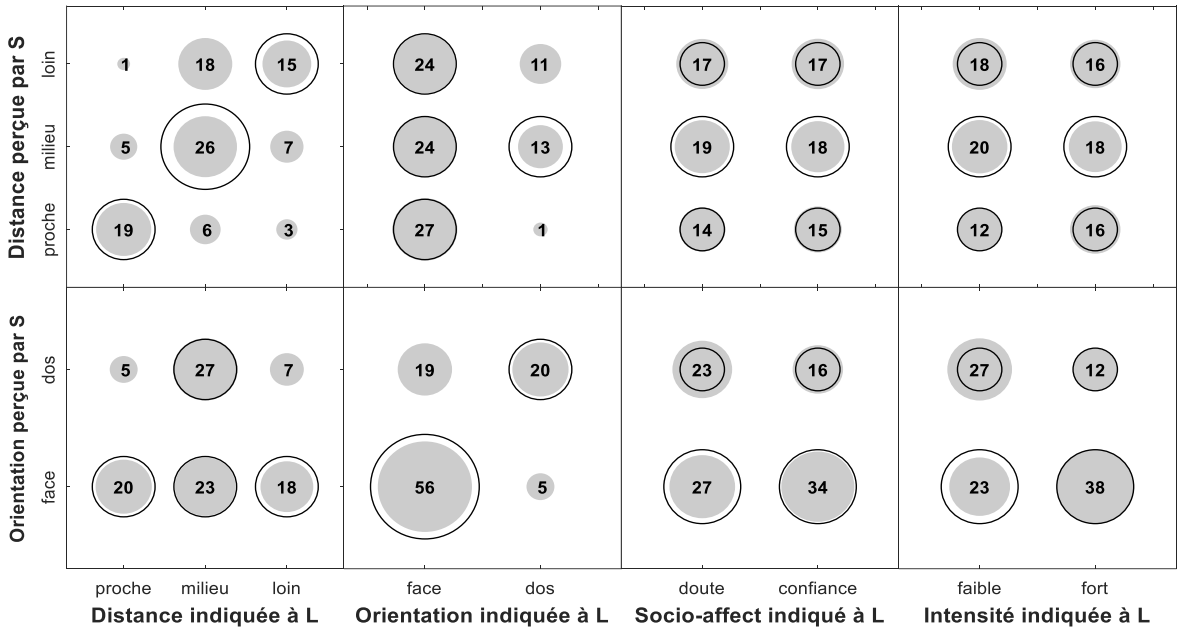

FIGURE 2 : Répartition des réponses données par les sujets (en %). *Le diamètre des disques gris est normalisé par rapport au pourcentage de réponses données pour une combinaison particulière d'indication donnée à L (indiquée en abscisse) et de réponse de S (indiquée en ordonnée). Les cercles noirs correspondent à la répartition qui aurait été obtenue si les sujets avaient donné la position réelle de L.*

5 Conclusion

L'interaction face à face in situ est un système complexe et fin qui permet aux interactants de se situer autant dans l'espace physique que dans l'espace social, la proxémie de leurs déplacements relatifs utilisant, différemment selon leurs cultures, cet espace physique pour signifier des informations sur leur espace social. Par cette étude, nous voulons montrer qu'il en est de même dans l'espace acoustique et que nous intégrons les distances physiques à nos productions vocales et notre audition pour exprimer et percevoir les distances socio-relationnelles. Ainsi, pour pouvoir assurer une immersion ubiquïte aussi bien physique que sociale, il faudra assister le téléopérateur dans la gestion de ces contrôles fins sous peine de générer des malentendus socio-relationnels importants. Nos premiers résultats montrent que le socio-affect exprimé par un locuteur et l'intensité de sa voix influencent la manière dont son orientation est perçue par un interlocuteur. Des analyses plus poussées devront être menées pour pouvoir établir un modèle du toucher vocal, qui sera testé et raffiné grâce à l'implémentation d'une interface pour robot de téléprésence.

Références

ANDERSON P. W. & ZAHORIK P. (2014). Auditory/visual distance estimation: accuracy and variability. Frontiers in psychology, 5, 1097.

BRUNGART D. S. & SCOTT K. R. (2001). The effects of production and presentation level on the auditory distance perception of speech. The Journal of the Acoustical Society of America, 110(1), 425–440.

CALCAGNO E. R., ABREGÚ E. L., EGUÍA M. C. & VERGARA R. (2012). The role of vision in auditory distance perception. Perception, 41(2), 175–192.

CHU, W. T. ; SEARCH FOR : WARNOCK A. C. C. (2002). Detailed Directivity of Sound Fields Around Human Talkers. Rapport interne, Institute for Research in Construction (National Research Council of Canada, Ottawa ON, Canada). pp. 1–47.

COLLINS K. & DOCKWRAY R. (2015). Sonic proxemics and the art of persuasion: An analytical framework. Leonardo Music Journal, 25, 53–56.

COOKE M., KING S., GARNIER M. & AUBANEL V. (2014). The listening talker: A review of human and algorithmic context-induced modifications of speech. Computer Speech & Language, 28(2), 543–571.

EDLUND J., HELDNER M. & GUSTAFSON J. (2012). Who am i speaking at? Perceiving the head orientation of speakers from acoustic cues alone. In Proceedings of LREC Workshop on Multimodal Corpora for Machine Learning : LREC.

GARDNER M. B. (1969). Distance estimation of 0 or apparent 0-oriented speech signals in anechoic space. The Journal of the Acoustical Society of America, 45(1), 47–53.

HALL E. T. (1966). The hidden dimension.

KIMURA A., IHARA M., KOBAYASHI M., MANABE Y. & CHIHARA K. (2007). Visual feedback: its effect on teleconferencing. Human-Computer Interaction. HCI Applications and Services, p. 591–600.

MAASØ A. (2008). The proxemics of the mediated voice. Lowering the boom: critical studies in film sound, p. 36–50.

PAEPCKE A., SOTO B., TAKAYAMA L., KOENIG F. & GASSEND B. (2011). Yelling in the hall: using sidetone to address a problem with mobile remote presence systems. In Proceedings of the 24th annual ACM symposium on User interface software and technology, p. 107–116: ACM.

PHILBECK J. W. & MERSHON D. H. (2002). Knowledge about typical source output influences perceived auditory distance. The Journal of the Acoustical Society of America, 111(5), 1980–1983.

REMLAND M. S., JONES T. S. & BRINKMAN H. (1991). Proxemic and haptic behavior in three european countries. Journal of nonverbal behavior, 15(4), 215–232.

SCHAFER R. M. (1977). The tuning of the world. Alfred A. Knopf.

TAKAHASHI M., OGATA M., IMAI M., NAKAMURA K. & NAKADAI K. (2015). A case study of an automatic volume control interface for a telepresence system. In Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), p. 517–522: IEEE.

ZAHORIK P., BRUNGART D. S. & BRONKHORST A. W. (2005). Auditory distance perception in humans: A summary of past and present research. ACTA Acustica united with Acustica, 91(3), 409–420.