Article par Nicolas SANTOLARIA
Publié le 06.12.2017
Crédit : Ina. Illustration Guillaume Long

Interfaces vocales : attention, dangers !

Les assistants vocaux se multiplient, tels Siri ou Google Now. Demain, ils occuperont une place considérable dans l'économie réelle et émotionnelle. La voix, comme appendice des outils, jouera un rôle prépondérant dans la fusion homme-machine qui se profile.

Sommaire

Diffusé en septembre 2017 aux États-Unis, le premier épisode de la 21e saison de South Park intitulé "White people Renovating Houses" a réussi un véritable tour de force : pirater les assistants vocaux de milliers de foyers américains. En prononçant simplement à l'écran les formules d'activation "Ok Google" et "Alexa", les personnages de la série animée de Trey Parker et Mat Stone sont parvenus à mettre en marche les enceintes connectées Google Home et Amazon Echo, puis à leur faire énoncer les pires grivoiseries. En effet, ces enceintes, avec lesquelles on interagit oralement, ne faisaient jusqu'alors pas de distinction entre les différents interlocuteurs et pouvaient être mises en route par n'importe quelle source sonore susceptible d'être captée par les micros ultra-sensibles qu'elles renferment.

Une fois ces enceintes activées, ces totems domestiques — qui ont pris place dans les intérieurs comme une nouvelle évidence — sont programmés pour répondre docilement aux demandes, énoncées le plus souvent sur un mode impératif : lancer la musique, faire des recherches sur Internet, effectuer une opération d'arithmétique, raconter une blague, localiser le fleuriste le plus

proche... Par le truchement de ces assistants multifonctions, le personnage Cartman de South Park, hacker à la fois fictionnel et réel, s'est donc amusé à ajouter sur les listes de courses des téléspectateurs "des gros seins", "des grosses boules velues" et des "chips aux nichons".

Ce piratage n'est pas le premier du genre, puisqu'en avril 2017 l'enseigne de fast-food Burger King avait réussi, elle aussi, à activer à distance les enceintes "Google Home" des téléspectateurs américains au moyen d'un message publicitaire télédiffusé, en vue d'assurer la promotion de son célèbre sandwich le "Whooper".

Le pouvoir suggestif de la voix

Par leur ampleur, ces détournements d'un outil technologique en vogue rappellent le fameux canular d'Orson Welles. Ayant fait croire à une invasion extra-terrestre en diffusant une adaptation sonore de La Guerre des mondes sur les ondes de CBS, le réalisateur avait mis en exergue, en 1938, l'emprise du média radiophonique sur les esprits, tout comme il avait magistralement souligné le pouvoir suggestif de la voix.

La farce interactive imaginée par les auteurs de South Park a eu, quant à elle, une vertu similaire : donner la mesure de la place grandissante prise aujourd'hui par l'interface vocale.

Depuis sa démocratisation au tout début de la décennie, cette interface a été implémentée dans les assistants personnels des smartphones (Siri, Cortana, Google Now, S-Voice, Bixby), ainsi que dans les enceintes intelligentes à commandes vocales (Amazon Echo, Google Home, HomePod), et sert de levier d'interaction plus trivial, permettant d'allumer une ampoule, de changer les chaînes de la télé, ou de régler un thermostat.

Activant des fonctionnalités, donnant accès à des services tiers, la voix permet déjà de transmettre à la machine les instructions de l'utilisateur de la manière la plus naturelle qui soit.

L'avènement des chatbots, ces agents conversationnels qui remplacent aujourd'hui l'usage des applications mobiles, s'inscrit également dans cette dynamique, puisque les réponses — encore majoritairement sous forme écrite — de ces robots bavards devenus les nouveaux vecteurs de la relation clients ont vocation dans le futur à être vocalisés.

Activant des fonctionnalités, donnant accès à des services tiers, la voix permet déjà de transmettre à la machine les instructions de l'utilisateur de la manière la plus naturelle qui soit.

Là où les technologies vocales ont longtemps rimé avec dysfonction, les progrès accomplis en termes de reconnaissance et traitement automatisé du langage, rendent désormais possible le fait de décrypter avec une précision suffisante la requête d'un utilisateur, nous faisant entrer dans ce que l'on pourrait appeler la "société de conversation".

Revenir au sommaire

Le mythe de la métaphore

D'après des chiffres publiés en 2017 par Google, 20 % des requêtes effectuées sur son système Android aux États-Unis passeraient désormais par la voix.

"Une aide précieuse dans toute la maison, tout en gardant les mains libres" : c'est ainsi qu'est présentée l'enceinte **Google Home Mini**, nouveau vade-mecum du domicile connecté.

Destituant la main de son rapport privilégié à l'objet technique, cette nouvelle frontière opératoire est qualifiée de *"voice first"* par l'industrie, résumant son ambition non seulement supplétive, mais à bien des égards substitutive. D'après des chiffres publiés en 2017 par *Google*, 20 % des requêtes effectuées sur son système *Android* aux États-Unis passeraient désormais par la voix. Quant aux recherches sur le web, 30 % d'entre elles devraient se faire sans écran d'ici 2020, selon l'institut d'études Gartner.

Longtemps instrument privilégié de l'interaction humain-machine, l'interface graphique, qui se matérialise traditionnellement pour le grand public sous la forme d'un bureau avec des menus et des icônes, se voit donc aujourd'hui concurrencée, ou complétée dans une optique multimodale, par l'interface vocale. Dans son texte de référence *"The Myth of Metaphor"*, le développeur informatique et théoricien Alan Cooper distingue ainsi trois paradigmes d'interaction : le premier d'entre eux est le *paradigme technologique* dans lequel l'interface reflète de manière quasi transparente le fonctionnement du mécanisme, *modus operandi* qui n'est accessible qu'à un petit nombre d'initiés.

Le *paradigme de la métaphore* repose quant à lui sur une analogie entre le monde matériel (le bureau, les dossiers, la poubelle) et les actions réalisées par la machine.

"Les métaphores semblent d'abord être un gain de temps pour les utilisateurs débutants, mais elles se montrent lourdes de conséquences lorsque l'on progresse dans l'utilisation approfondie d'un logiciel. Il est donc préférable de concevoir les choses de manière idiomatique, en utilisant la métaphore de manière occasionnelle, lorsque l'une d'elle nous tombe sous la main", écrit Alan Cooper.

Simple à comprendre et à utiliser, la métaphore sur laquelle repose l'interface graphique deviendrait, dès lors, contre-productive en raison même de son formalisme contraignant.

Dans un contexte où de plus en plus d'objets deviennent communicants, il est extrêmement fastidieux d'avoir à passer systématiquement par des menus déroulants, des fenêtres, des manipulations d'icônes pour signifier à l'appareil votre volonté.

Résumé par l'acronyme *WIMP* (*Windows, icons, menus, pointers*), ce mode d'interaction dans lequel on pouvait encore percevoir la trace lointaine d'un transfert de force se révèle, de plus, mal adapté aux écrans réduits des smartphones.

Au contraire, la parole est une forme d'expression naturelle, télé-active, bien plus rapide que ne l'est le mode d'action *haptique*, mettant en jeu le toucher.

Revenir au sommaire

Destitution de la main

Cette possibilité de dialoguer avec les machines sur un mode conversationnel tel celui que l'on emploie avec un ami ou un proche est loin d'être une simple réorientation cosmétique de nos façons de faire. Si l'on suit les réflexions de l'ethnologue André Leroi-Gourhan, on peut même y voir une nouvelle étape marquante dans l'appréhension de l'objet technique :

> "À l'origine de la discrimination que nous faisons encore entre "l'intellectuel" et le "technique" se trouve la hiérarchie établie par les Anthropiens entre action technique et langage, entre l'œuvre liée au plus réel de la réalité et celle qui s'appuie sur les symboles", écrit André Leroi-Gourhan.[1]

En fait, les êtres humains utilisent la même partie du cerveau pour interagir avec une machine ou avec un humain Cette anthropomorphisation de la technique s'ancre dans les processus cognitifs particuliers, activés par l'oralité. Dans son ouvrage *Wired for speech : how voice activates and advances the Human-Computer Relationship*[2], le spécialiste de communication Clifford Nass souligne que

> "le cerveau humain fait rarement la distinction entre le fait de parler avec une machine — même les machines qui ont une compréhension très pauvre du langage et une production de mauvaise qualité de la parole — et avec une personne. En fait, les êtres humains utilisent la même partie du cerveau pour interagir avec une machine ou avec un humain".

Cet état de fait est encore accentué par la tendance d'une partie de l'industrie à jouer jusqu'au bout la carte de l'humanisation. Alexa, l'assistante virtuelle des enceintes connectées Echo d'Amazon, est ainsi susceptible de s'exprimer en utilisant les nuances telles que le chuchotement, les variations de débit ou d'intonation, les pauses. Soit autant d'éléments prosodiques qui facilitent l'assimilation de la voix de synthèse à une voix humaine.

Quant aux assistants vocaux des smartphones, ils ont été conçus en fonction des codes de civilité en vigueur dans la communication interpersonnelle, maniant à la perfection les formules de politesses et, à l'instar des hommes politiques, l'adresse directe à l'interlocuteur. Si vous dîtes par exemple "Merci" à Siri, l'assistant vocal d'Apple pourra vous répondre alors :

> "Tout le plaisir est pour moi".

Revenir au sommaire

Personnalité artificielle

Outre cette similarité dans les "façons de parler"[3], le phénomène d'empathie est également suscité par le fait que ces programmes ont été dotés, pour nombre d'entre eux, d'une "personnalité", laquelle se traduit au travers de leurs réponses. Ces dernières font parfois référence à une corporalité elliptique,

---

1 André LEROI-GOURHAN, Le Geste et la Parole. Technique et langage, Volume 1, Albin Michel, 1964, 2013.

2 Clifford NASS, Scott BRAVE, Cambridge (MA), USA, MIT Press, 2005.

3 Erving GOFFMAN, La mise en scène de la vie quotidienne, Tome 1, La présentation de soi, Les éditions de Minuit, 1973

en évoquant des sensations et utilisent le plus souvent le registre de l'humour pour donner à penser qu'une subjectivité agissante opère au cœur du dispositif.

Un personnage enfermé dans le smartphone comme le génie dans la bouteille, c'est bien à cela que l'on pense lorsque l'on converse avec un assistant personnel intelligent

En réalité, ce sont en général des réponses écrites en amont par des pools d'auteurs qui, diffusées sur un mode random, permettent par cet aléatoire de synthèse de donner l'illusion de l'intelligence artificielle en action, plaçant ces dispositifs dans un registre qui oscillerait entre l'ingénierie de haut vol et un art forain cinématographique de la suggestion. Professeur de théorie des nouveaux médias à l'université de Californie de San Diego, Lev Manovich écrit dans son ouvrage de référence *Le Langage des nouveaux médias* (Paris, Dijon, Les Presses du réel, 2015) :

> *"La tradition de l'imprimé qui dominait à l'origine le langage des interfaces culturelles a perdu de son importance, alors que le rôle joué par les éléments cinématographiques devient de plus en plus prégnant."*

Un personnage enfermé dans le smartphone comme le génie dans la bouteille, c'est bien à cela que l'on pense lorsque l'on converse avec un assistant personnel intelligent, vision mise en scène dans le film *Her* de Spike Jonze où le héros, incarné par Joaquin Phoenix, tombe amoureux de la voix de synthèse féminine d'un *software*.

De par sa dimension ontologiquement mimétique, l'interface vocale, qui reproduit nos processus les plus intimes, n'est donc plus un simple levier, mais devient un objet social à part entière, une sorte de tiers de qui l'on attend, non seulement un accès facilité aux fonctions d'un système opérationnel, la curation (sélection) d'informations dans l'océan des données ou bien encore l'accès à des services tiers, mais également une forme de compagnonnage émotionnel. De cela découle le fait que les réactions de cet objet doivent s'insérer dans le cadre moral qui régit les sociétés humaines.

Revenir au sommaire

Honte prométhéenne

En s'hominisant au travers de son expression dialogique, l'interface vient dès lors bouleverser l'humanisme classique, déniant à l'être humain son exceptionnalité dans un domaine qui lui était jusqu'alors exclusif : celui du maniement du langage.

La perfection affichée de ce type d'objet renvoyant l'utilisateur à son incomplétude, le penseur allemand Günther Anders y voit les germes d'un sentiment grandissant, qu'il nomme la *"honte prométhéenne"*.

Là où l'interface graphique matérialisait une forme de distance prothétique — la souris pouvant s'apparenter à une sorte de bras télescopique virtuel —, l'interface vocale s'inscrit, au contraire, dans une dynamique d'incorporation de l'opération technique, devenue congruente avec nos propres processus. Soit un mouvement de fusion humain-machine.

"Parce que l'objet automatisé "marche tout seul", il impose une ressemblance avec l'individu humain autonome, et cette fascination l'emporte. Nous sommes devant un nouvel anthropomorphisme", écrit Jean Baudrillard dans le Système des objets (Gallimard, 1978).

Le mode dialogique incorporé dans les machines accentue encore la force de ce constat, s'accompagnant d'un ensemble de significations, d'enjeux, qui excèdent la simple dimension opératoire, pour dériver vers le terrain narcissique.

Si elle se calque à bien des égards sur la relation entre humains, l'interaction avec l'interface vocale possède ainsi ses rites singuliers et déroutants

Si elle se calque à bien des égards sur la relation entre humains, l'interaction avec l'interface vocale possède ainsi ses rites singuliers et déroutants, de nombreux utilisateurs insultant de façon très violente la personnalité mimétique, comme pour se défendre de cette duplication miroitante.

"La machine énergétique avait dévalorisé l'homme physique, comme travailleur manuel, mais cela avait plutôt été perçu comme un progrès, comme la fin d'un esclavage. La dévalorisation par la machine informationnelle de l'homme pensant posait évidemment bien d'autres problèmes", écrivait l'historien Philipe Breton, dans Une histoire de l'informatique (Seuil, 1990).

D'autant que cette dimension concurrentielle n'est pas uniquement cosmétique, mais concerne également la façon dont nous percevons le monde, dont nous y accédons.

Revenir au sommaire

L'avènement des "do engine"

Accompagnant une automatisation de plus en plus poussée des fonctions cognitives, l'interface vocale marque le passage d'une ère dominée par les "search engine" à une période où les "do engine" s'imposent comme vecteurs alternatifs d'accès aux masses exponentielles de données, traitées au moyen de l'intelligence artificielle. Le "search engine", dont le moteur de recherche de Google est la manifestation la plus emblématique, se caractérise par le fait que l'algorithme va produire une série de liens classés selon leur pertinence supposée, en réponse à une requête donnée. Même si cela est fastidieux, vous avez alors tout loisir de compulser l'ensemble des résultats pour vous forger une opinion propre, en une forme de gymnastique cognitive basée sur la pondération.

Selon les bases de données où ils puisent, les assistants vocaux produisent des réponses extrêmement fluctuantes, assénées comme des vérités immanentes

Lorsqu'ils répondent oralement à une requête, les "do engine", après avoir analysé votre demande et effectué les recherches nécessaires, condensent les résultats en une réponse unique et vocalisée. Il s'agit là d'un "moteur actif", pour reprendre l'expression de Norman Winarsky, un des créateurs de

Siri. Si le gain de temps est indéniable, le "do engine" opère en fait un court-circuit de nos processus de décision, désormais arraisonnés par la machine. Le pouvoir normatif du dispositif s'en trouve logiquement accru, posant désormais la question de la source. D'où provient l'information ? Quelles raisons ont présidé à ce choix final ? Questionnements d'autant plus nécessaires que, selon les bases de données où ils puisent, les assistants vocaux produisent des réponses extrêmement fluctuantes, assénées comme des vérités immanentes.

> "La propriété des dispositifs d'apprentissage profond non supervisé tient à ce qu'ils court-circuitent le passage jadis nécessaire par des subjectivités humaines, non seulement dans le moment de la reconnaissance d'objet, mais plus fondamentalement dans la formulation d'hypothèses de pertinences", écrit le chercheur Yves Citton, dans la revue Multitudes n°68 ("Le court-circuitage néo-libéral des volontés et des attentions", automne 2017).

En moins d'une décennie, nous sommes donc passés de la recherche d'informations potentialisée par les moteurs et les méta-moteurs à une forme évoluée et encore évolutive de conciergerie cognitive (soit une sorte de cyber-valet qui nous accompagne au quotidien), permettant de réaliser sans effort, mais non sans déperdition, des tâches documentaires et perceptuelles. Cette forme d'assistanat en vogue s'inscrit dans une vision anthropologique marquée par une supposée maximalisation de nos potentialités au moyen de la technologie, que résume l'expression d'"homme augmenté".

Revenir au sommaire

Attention captive

Demandez à Siri : "Est-ce que je dois acheter un téléphone Samsung ?" Et il vous répondra : "Eh bien, je suis peut-être de parti pris mais je préfère les produits Apple"

Outre leur fonction affichée, ces nouveaux outils ont pour objectif sous-jacent de maintenir le consommateur dans un contexte d'achat potentiel, l'invitant à demeurer dans un écosystème où son attention restera captive. Si vous avez, via votre enceinte connectée, l'univers commercial d'Amazon à portée de main, vous serez plus enclins à y réaliser aveuglément vos emplettes, avec d'autant plus de facilités que l'interaction vocale rend tout cela extrêmement léger et que l'algorithme vous bombardera de suggestions pertinentes, formulées grâces aux informations recueillies via cette même enceinte connectée. L'agent conversationnel est si proche du client que ses propositions ciblées et personnalisées pourraient à terme faire disparaître le marketing tel qu'on le connaît aujourd'hui. Orientant ouvertement nos perceptions, ces assistants intelligents sont porteurs de biais qu'ils évoquent parfois eux-mêmes sans détour. Demandez par exemple à Siri :

> "Est-ce que je dois acheter un téléphone Samsung ?" Et il vous répondra : "Eh bien, je suis peut-être de parti pris, mais je préfère les produits Apple".

Si elle peut sembler récente dans son usage quotidien, la commande vocale a en réalité plusieurs décennies d'existence, se développant à partir des années 1950 aux États-Unis, 1970 en France. Malgré quelques antécédents notables que nous ne pouvons ici recenser, c'est l'intégration, à partir d'octobre 2011, du logiciel Siri à l'iPhone 4S d'Apple qui aura incontestablement marqué un tournant,

démocratisant auprès du grand public ce type de *software*. Fruit d'un vaste programme de recherche américain lancé par la Darpa en 2003 et visant à soulager les commandements militaires de la surcharge cognitive, *Siri*, et par extension ce type d'outils, nous invite à un usage finaliste et instrumental du langage, que l'on doit envisager dans le contexte élargi de ce que le philosophe italien Maurizio Ferraris nomme la *Mobilisation totale* (titre d'un de ses récents ouvrages paru aux PUF en 2016).

Nous sommes, au travers de tous ces objets, mobilisés en permanence, dans un état d'alerte pareil à celui du soldat en guerre.

Nous sommes, au travers de tous ces objets, mobilisés en permanence, dans un état d'alerte pareil à celui du soldat en guerre. Cette porosité des logiques militaires et civiles, cette contamination de l'une par l'autre, doit nous faire envisager l'interface conversationnelle moins comme un instrument de dialogue facilité avec les machines, que comme un vecteur efficace d'enrégimentement.

Comme l'avait fort bien montré Stanley Kubrick dans *2001, l'odyssée de l'espace* au travers de la figure inquiétante de l'ordinateur de bord *HAL 9000* — lequel s'exprimait au moyen d'une voix de synthèse —, l'individu, une fois reproduit dans ses processus les plus intimes, est alors sommé d'entrer dans le rang, en réduisant la parole à sa dimension uniquement performative. Envisagé du point de vue cybernétique, on l'invite alors à penser qu'il n'est pas une voix singulière avec tout ce que cela suppose de mystère et d'unicité, mais un simple processeur traitant de l'information de manière désuète, ralentie et imparfaite, une variable superflue, un grain de sable dans la machine.

Revenir au sommaire

À lire également dans le dossier *"De la radio aux robots parlants, métamorphoses de la voix"* :