Où vont nos données ? L'exemple des assistants vocaux

Par **Martin BIERI**
Chargé d'études au Laboratoire d'innovation numérique de la Commission nationale de l'informatique et des libertés (CNIL)

La voix est l'un – si ce n'est le premier – de nos principaux outils de communication : nous l'utilisons quotidiennement, pour toutes sortes d'interactions. Si elle est restée longtemps le vecteur d'échanges entre les humains (et aussi, dans une moindre mesure, avec les animaux), elle devient une nouvelle interface dans la relation homme-machine, incarnée par les assistants vocaux. Désormais, chacun peut échanger directement avec des systèmes informatiques en utilisant le langage de tous les jours, le « langage naturel [1] ». Ce qui signifie que la machine possède des capacités allant de la compréhension à la formalisation d'une réponse, vocale également. Tout ce processus induit un traitement de notre voix et de nos paroles par l'objet communicant. Alors, comment fonctionne réellement un assistant vocal ? Et que fait-il de notre voix ?

Le rapport à la voix : une donnée à géométrie variable

Une histoire pas si récente

Avant de regarder sous le capot de ces « nouveaux majordomes », analysons la voix et la parole comme « données à interpréter ». Historiquement, la voix était une donnée volatile : selon le proverbe, les paroles s'envolent, et les écrits restent. Toutefois, depuis l'invention du phonographe par Thomas Edison et Charles Cros (chacun de son côté), à la fin des années 1870, permettant d'enregistrer les sons – et donc la voix –, cet adage semble de plus en plus obsolète. Son développement entraîne une nouvelle perception de ce moyen de communication, notamment sur le plan juridique. En effet, quelques années après, en 1890, dans un texte autour de la protection des données et de la vie privée (sur la notion du « droit d'être laissé tranquille »), les juristes américains Warren et Brandeis pointent les risques associés à ces nouvelles technologies d'enregistrement, y incluant « le possesseur de tout autre dispositif moderne de reformulation ou de reproduction de scènes ou de sons [2] ». En France, le code pénal actuel précise notamment dans son article 226-1 la protection contre la captation, l'enregistrement ou la transmission « des paroles prononcées à titre privé ou confidentiel [3]».

Des données multiples et complexes

La voix n'est pas qu'un son à interpréter, mais une donnée à géométrie variable. Tout d'abord, il faut distinguer l'information verbale, c'est-à-dire le message en lui-même que nous voulons transmettre, et les informations non verbales, dites « paralinguistiques » : l'intonation, les silences, les gestes, etc.

(1) En opposition au « langage formel », codifié pour être non ambigu et qui est utilisé pour les programmes informatiques.
(2) WARREN S. & BRANDEIS L. (1890), "The right to privacy", *Harvard Law Review*, 4(5), December 15.
(3) Version en vigueur au 1er août 2020.

Sans pour autant entrer dans la théorie entre « signifiant » (l'image acoustique associée à un mot) et « signifié » (le concept de ce mot) [4], la parole peut véhiculer un ou plusieurs sens [5]. Et selon ce qui est dit, l'information que l'on en retire sera différente : une phrase anodine n'aura pas la même notion de sensibilité qu'une phrase délivrant un secret, ou faisant état d'une opinion politique, d'une appartenance syndicale, d'une orientation sexuelle, etc.

Par ailleurs, hors du sens propre de ce qui est dit, la voix renferme également d'autres éléments, la part non verbale. En effet, nous sommes en capacité de reconnaître ou d'en déduire d'autres informations : âge, genre, état émotionnel, état de santé (par exemple, des maladies dégénératives comme Parkinson [6]), condition physique, origine géographique, accent, etc.

Et ce jusqu'à l'identité du locuteur [7]. La voix possède des caractéristiques qui sont propres à chaque individu : des attributs biométriques « comportementaux [8] » qui permettent d'en authentifier l'identité, à travers la création d'un modèle de voix ou gabarit. La voix est d'ailleurs comprise dans le droit à l'image, en tant qu'« attribut de la personnalité, une sorte d'image sonore ». Puisqu'elle permet l'identification d'une personne physique, de manière directe ou indirecte, la voix est de fait une donnée personnelle au sens du Règlement général sur la protection des données (RGPD). Les données biométriques, à l'instar des données de santé, sont donc considérées comme « sensibles » dans le RGPD, c'est-à-dire des données dont l'utilisation est particulièrement encadrée : l'article 9 prévoit ainsi une interdiction à leur traitement, en permettant des exceptions. Le traitement de la voix n'est donc pas un processus anodin…

La voie des données

Fonctionnement d'un assistant vocal et circulation des données

Il faut d'abord préciser ce qu'est un assistant vocal, à commencer par ce qu'il n'est pas : une enceinte connectée n'est pas forcément un assistant vocal, mais peut en être équipée. Ce que l'on considère être un assistant, c'est la part logicielle offrant des capacités de dialogue oral en langage naturel avec l'utilisateur de l'objet l'embarquant. De manière plus pratique, nous pouvons définir trois grandes entités entrant dans la composition d'un assistant vocal [9] :

- **l'instance physique** : tout élément matériel dans lequel va prendre forme l'assistant, et qui se concrétise aussi par la présence de microphones, haut-parleurs et capacités de calcul ;
- **l'instance logicielle** : c'est la partie qui met en œuvre l'interaction homme-machine (à travers les modules de transcription automatique de la parole, de compréhension et génération du langage naturel, de dialogue et de synthèse vocale) ; elle peut être opérée au sein de l'objet, mais, de façon générale, est réalisée de manière distante ;
- **les ressources** : toutes les données externes, bases de connaissances et applications qui fournissent la réponse à la question posée ou permettent de déclencher l'action demandée par l'utilisateur.

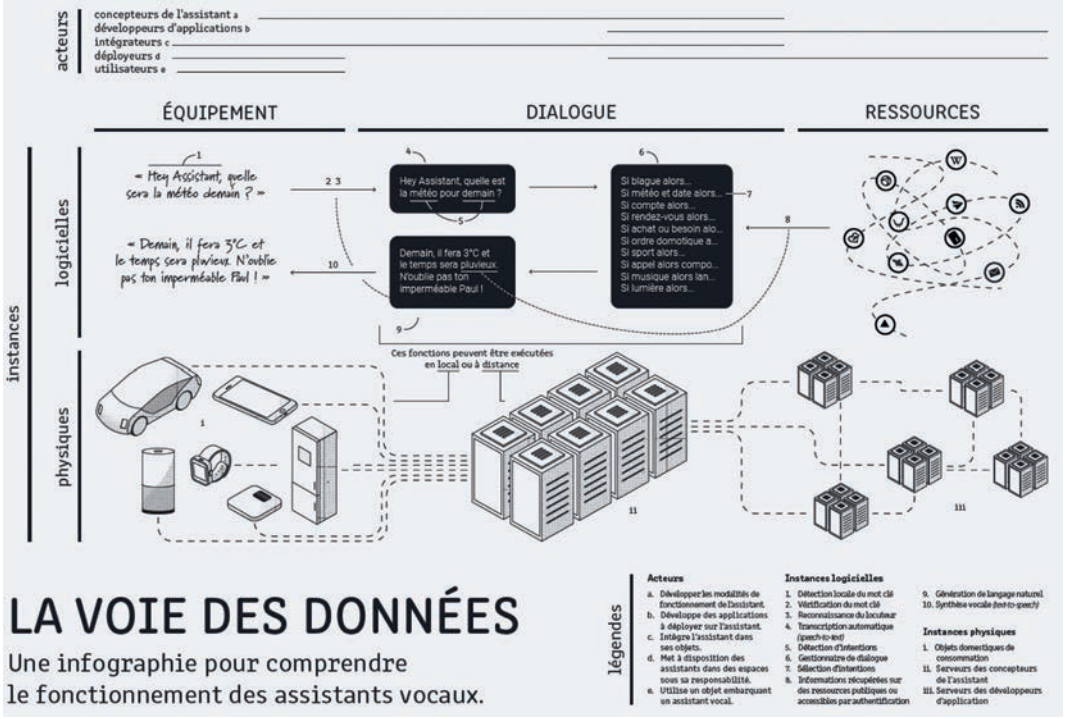
Un assistant vocal n'enregistre pas en continu tout ce qui est à portée de microphone : il est nécessaire d'utiliser un mot-clé pour « réveiller » l'assistant, qui sinon reste en veille. Il n'utilise

(4) DE SAUSSURE F. (1916), *Cours de linguistique générale*, Payot.
(5) REVIS J. (2018), « Notre voix porte en elle toutes les intentions qui sont les nôtres », *Linc.cnil.fr*, entretien, mars.
(6) JEANCOLAS L. *et al.* (2016), « L'analyse de la voix comme outil de diagnostic précoce de la maladie de Parkinson : état de l'art », *Compressions et Représentation des Signaux audiovisuel*, mai.
(7) BONASTRE J.-F. (2017), « La voix n'est pas une biométrie classique », *Linc.cnil.fr*, entretien, février, https://linc.cnil.fr/jean-francois-bonastre-la-voix-nest-pas-une-biometrie-classique
(8) BONASTRE J.-F., *Ibid.*
(9) CNIL (2020), « À votre écoute – Exploration des enjeux éthiques, techniques et juridiques des assistants vocaux ».

alors qu'une mémoire tampon qui n'a pour but que de vérifier si ce mot de réveil a bien été prononcé. Cette vérification est faite localement au sein du dispositif, mais, pour les cas où les traitements sont faits à distance, il peut y avoir une deuxième passe effectuée sur des serveurs, afin de déceler un faux positif (un mot qui aurait été pris pour le mot-clé par erreur). Il peut y avoir à ce moment une vérification de l'identité du locuteur, si ce dernier s'est enrôlé, c'est-à-dire s'il a préalablement partagé ses caractéristiques vocales à l'assistant. Une fois réveillé, l'assistant va recevoir la requête de l'utilisateur qui sera traitée localement ou à distance (sur le modèle « mot de réveil, quel temps fera-t-il demain ? »). L'audio est alors transcrit en texte (*speech to text*), et le texte va être interprété par les algorithmes de traitement automatique du langage : seront alors identifiées les intentions du message ainsi que les variables d'informations. Concrètement, dans l'exemple de l'interrogation de l'assistant sur la météo pour demain, l'intention est la demande du temps, et la variable est donc « demain », la temporalité. À partir de cette identification, un gestionnaire de dialogue va pouvoir construire le scénario de la réponse à apporter. Si nécessaire, les bases d'informations distantes (ici, la base de données météorologiques) seront interrogées pour préciser la réponse selon les variables identifiées. La phrase de réponse est ainsi générée – et/ou l'action, si c'était l'objet de la requête (comme « monter le chauffage »). Elle est alors synthétisée (*text to speech*) et est donc ensuite mise en œuvre par l'objet embarquant l'assistant, et l'assistant repasse en veille !

Les acteurs en présence

Pour faire fonctionner un assistant, plusieurs acteurs sont nécessaires. Ils interviennent à différents moments dans la chaîne de valeur, à commencer par le **concepteur** de l'assistant. C'est celui qui va mettre au point l'assistant, en choisir le fonctionnement et les paramètres (techniques, matériels, etc.). Interviennent également les **développeurs d'application** : ce sont eux qui vont élaborer les applications utilisables *via* l'assistant, à l'image de ce qui est fait pour les *smartphones*. Bien qu'ils



doivent respecter le cadre instauré par le concepteur, ils ont une responsabilité également puisque, généralement, ce sont eux qui vont définir les finalités de l'application et le traitement des données (voir le cas d'usage n°2 dans le livre blanc « À votre écoute »). Il y a ensuite les **intégrateurs**, ceux qui vont implémenter dans leurs objets et équipements l'assistant vocal avant que les **déployeurs** (même si le terme n'est pas très élégant) n'installent l'assistant dans des endroits partagés, des endroits de passage, des environnements de travail, etc. Enfin, il ne reste plus que **l'utilisateur** lui-même, qui va interagir avec l'assistant vocal (et éventuellement être lui-même « déployeur » s'il installe chez lui des équipements embarquant l'assistant vocal).

Les grands enjeux

Des stratégies économiques, notamment fondées sur la récupération des données

La plupart des assistants vocaux font peu de traitement des données embarqué directement dans les dispositifs, mais plutôt sur des serveurs distants. Ce qui implique que le transfert des données vocales ainsi que le traitement se passent directement chez le concepteur de l'assistant. Ceci est parfois justifié par le fait qu'il faudrait embarquer directement dans l'objet connecté des capacités de calcul plus importantes que ce qui est fait majoritairement aujourd'hui, et qu'il est plus facile et moins coûteux de tout centraliser.

Cependant, il y a aussi – et surtout – des enjeux économiques dans ce mode de fonctionnement : les modèles d'affaires des grands constructeurs occidentaux (à commencer par les deux firmes qui dominent ce marché, Google et Amazon) reposent sur une stratégie classique de récupération des données à des fins publicitaires. Si la technologie vocale est « nouvelle » dans nos usages, son modèle économique reste celui de la publicité classique sur Internet : l'assistant vocal n'est qu'un avatar, un point d'entrée de plus pour récupérer des données d'utilisation, afin de pouvoir enrichir un profil qui sera ensuite utilisé pour du ciblage publicitaire. Par ailleurs, pour pouvoir utiliser l'assistant à 100 %, il est souvent nécessaire de créer un compte ou de le synchroniser avec un compte déjà existant (par exemple, un compte Gmail pour Google ou Amazon pour Alexa).

Le fait de n'embarquer que peu de puissance de calcul directement dans l'objet est aussi un moyen pour baisser son coût : le but est ainsi la pollinisation du marché par une forte production d'équipements à moindre frais, permettant une assise certaine sur le marché. Les grands constructeurs ont aussi fait en sorte qu'il soit aisé de mettre au point une application (pour la déployer dans un assistant) comme de l'intégrer dans un équipement tiers (montre, frigo, aspirateur, etc.).

Quels enjeux concernant les données pour les assistants vocaux ?

Les enjeux posés par les assistants vocaux et les assistants tout court par ailleurs, comme en témoignent les travaux du Comité national pilote d'éthique du numérique (CNPEN) sur les *chatbots* [10], sont nombreux, et n'ont pas tous trait à l'utilisation des données (et notamment les données personnelles [11]) : relation homme-machine, genre et identité de l'assistant, anthropomorphisation, l'implémentation d'IA et la question des biais qui lui est souvent associée, en passant par des problématiques liées au fait que les noms utilisés pour les assistants soient déjà existants [12], etc.

(10) CNPEN (2020), « Les enjeux éthiques des agents conversationnels », https://www.ccne-ethique.fr/fr/actualites/cnpen-les-enjeux-ethiques-des-agents-conversationnels, 26 juin.
(11) CNIL (2020), « À votre écoute – Exploration des enjeux éthiques, techniques et juridiques des assistants vocaux ».
(12) PINSKER J. (2021), "Amazon killed the name Alexa", *The Atlantic*, 18 août.

La question de la circulation des données se pose : la chaîne de valeur d'un assistant vocal comprend un certain nombre d'acteurs, dont plusieurs sont directement destinataires des données ou les voient transiter sur leurs serveurs. Qui est alors responsable du traitement de données ? La question du stockage de certains types de données a soulevé des interrogations, notamment le fait de proposer un enrôlement à l'utilisateur. Ce processus implique de garder quelque part un gabarit biométrique (modèle de voix) propre à l'utilisateur – la plupart des assistants donnent désormais la possibilité de stocker cette partie directement en local, au sein de l'objet connecté embarquant l'assistant.

De la même manière, la sensibilité des données – dans le sens de ce qu'elles peuvent révéler sur l'intimité des individus – a été au cœur d'un scandale à l'été 2019. Comme pour tout système d'apprentissage automatique, il est nécessaire d'avoir une supervision humaine et un contrôle du fonctionnement. Pour les assistants vocaux, cela veut dire améliorer les performances par la réécoute d'interactions (par exemple, en vérifiant ce qui avait déclenché un faux positif dans le réveil de l'assistant, ou pourquoi l'assistant avait mal compris). L'information apportée aux personnes ne contenait pas de précision sur « cette amélioration à des fins de service », et la base légale d'un tel traitement posait question – l'autorité de protection des données d'Hambourg [13] avait alors demandé à Google de suspendre ces écoutes le temps de clarifier le fondement juridique de ce type de traitement. Par ailleurs, ces écoutes pouvaient contenir des éléments intimes (voire des scènes violentes), et ces données pouvaient être traitées par des prestataires externes aux entreprises conceptrices, jetant encore un peu plus de flou sur qui avait accès à ces données. Les grands concepteurs ont alors revu leurs pratiques, arrêtant cette manière de faire pour certains ou permettant une option de retrait (ou *opt-out*) de ce système associée à une meilleure information des utilisateurs. Cette question est d'autant plus problématique lorsque les enregistrements proviennent de personnes n'ayant pas été informées de ce traitement ou n'ayant pas donné leur consentement. En réveillant l'assistant vocal sans avoir connaissance qu'il est dans la pièce (en prononçant un mot-clé proche par exemple), la personne ignore que l'assistant va traiter l'extrait comme étant une requête.

L'assistant vocal et son incarnation physique (enceinte, véhicule, montre, etc.) sont donc un nouveau point d'entrée pour les concepteurs, leur permettant de récolter des données, de la même manière que pour une navigation sur le *web*. C'est aussi une porte d'entrée vers les comptes et les données des utilisateurs. De nombreuses publications ont montré que les assistants vocaux (notamment à travers les enceintes communicantes) étaient sensibles à des attaques plus ou moins sophistiquées. Les angles sont multiples pour accéder à cette porte d'entrée vers nos données : à travers des sons inaudibles par l'homme (*dolphin attack* [14]), à travers un laser pointé vers l'assistant (à une centaine de mètres de distance ! [15]), ou tout simplement se trouver dans son rayon d'action (environ 5 mètres). Des enjeux qui ont amené un cabinet d'avocats irlandais à bannir leur utilisation dans ses locaux – notamment pour préserver le secret professionnel et se prémunir d'éventuelles écoutes intempestives.

(13) The Hamburg Commissioner for Data Protection and Freedom of Information (2019), "Speech assistance systems put to the test - Data protection authority opens administrative proceedings against Google", août.
(14) ZHANG G. *et al.* (2017), "DolphinAttack: Inaudible voice commands", *ACM Conference on Computer and Communications Security*, novembre.
(15) SUGAWARA T. e*t al.* (2019), "Light commands: Laser-based audio injection on voice-controllable systems", novembre, https://lightcommands.com/

Conclusion

Il y a déjà eu de multiples changements depuis l'arrivée du premier assistant vocal grand public, Siri, en 2011. Encore récemment, au printemps 2021, Apple a annoncé des modifications : l'assistant deviendra « un peu plus privé [16] », faisant en sorte qu'aucun contenu ne soit enregistré par défaut, et que le maximum de traitement de la voix se fasse localement. D'autres concepteurs ont engagé des travaux pour permettre une meilleure gestion des données, ont corrigé également la sensibilité aux mots de réveil, etc. La CNIL, dans son livre blanc, avait alors listé plusieurs pistes d'amélioration, selon les problématiques et les acteurs concernés [17], à travers quatre grands principes cardinaux : entretenir les frictions désirables (profiter des moments de choix et de paramétrage pour présenter la réalité des traitements), privilégier le local au distant, assurer les moyens de contrôle par l'utilisateur et s'adapter au média vocal (présentation de l'information, recueil du consentement, etc.).

(16) HERMANN V. (2021), « Vie privée et sécurité : Apple monte d'un cran », *NextInpact.com*, 23 juillet.
(17) CNIL (2020), *Ibid.*, voir pp. 66 à 84.