Traitement automatique de la parole :
Quelle écoute pour nos systèmes ?

La parole est communément définie comme la faculté de communiquer oralement propre aux hommes. Intrinsèquement liée au langage, elle est une réalisation manifeste de ce dernier. Notre voix, quant à elle, assure mécaniquement la production de la parole par la vibration des cordes vocales, grâce à la pression de l'air expiré des poumons et aux résonateurs constitués par les cavités buccales et nasales. D'un point de vue conceptuel, parole et voix peuvent être appréhendées comme les deux constituants d'un signe linguistique selon la définition donnée par Ferdinand de Saussure[1], la parole composant le signifié et la voix, le signifiant. En pratique, nous sommes quotidiennement confrontés à l'analyse multiniveaux des signaux de parole. En plus de comprendre le « sens » du message qui nous est transmis, cette analyse nous révèle de nombreuses autres informations sur notre interlocuteur : âge, sexe, origines géographiques et socioculturelles, physionomie, état de santé, émotions, *etc.* Notre voix, véhicule privilégié de nos interactions sociales, révèle donc énormément sur nous.

Ayant connu une très forte expansion dans les années 1960 suite à la publication de la théorie de l'information de Claude Shannon, le traitement de la parole (*speech processing*) est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Situé au croisement de la physique (acoustique, propagation des ondes), des mathématiques appliquées (modélisation, statistique), de l'informatique (algorithmique, techniques d'apprentissage) et des sciences de l'homme (perception, raisonnement), le traitement de la parole a rapidement été décliné en de nombreux domaines d'étude : reconnaissance et vérification du locuteur, transcription automatique de la parole, synthèse vocale, détection des émotions, *etc.* Depuis une quinzaine d'années, la discipline dans son ensemble a progressé de manière remarquable et de grandes avancées ont été enregistrées. Les grands acteurs du numérique ne s'y trompent d'ailleurs pas. Pour eux, l'avenir de nos interactions avec les systèmes passe par l'analyse des signaux de parole. Nous avons coutume de dire que « les paroles s'envolent, les écrits restent ». Toutefois, les changements profonds induits par le numérique pourraient rendre caduc le proverbe. Avec l'explosion annoncée de l'Internet des Objets, nous serons de plus en plus encouragés à interagir de façon « naturelle » avec nos systèmes. Téléviseurs, agents conversationnels ou encore systèmes d'authentification seront littéralement « à notre écoute ». Toutefois quelle sera la qualité de celle-ci ? Nos nouveaux compagnons de vie auront-ils la décence de parfois entendre sans écouter ? Sauront-ils conserver les secrets que nous leur confierons ?

L'actualité met quotidiennement en exergue ces questionnements. Très récemment, la police de l'Arkansas a ainsi émis un mandat demandant à Amazon de lui fournir les données enregistrées par son dispositif Echo

[1] [Saussure] "Cours de linguistique générale", Paris, 1916.

afin de l'aider dans la résolution d'un meurtre ayant eu lieu en novembre 2015[2]. En effet, les enquêteurs ont découvert que le produit d'Amazon était allumé et diffusait de la musique non loin de la scène du crime. Convaincue du fait que le dispositif enregistre les sons de son environnement en permanence, la police de l'Arkansas a donc demandé d'accéder aux précieuses données. Ce fonctionnement est fermement démenti par la multinationale qui précise que des enregistrements ne sont réalisés, transférés et analysés que si un mot déclencheur est prononcé (par défaut « Alexa », nom de l'assistant virtuel). Quelle que soit l'issue de cette opposition, elle illustre parfaitement la sensibilité des signaux de parole. En effet, quelle confiance accorder à des systèmes qui pourraient, si mal implémentés, écouter en permanence nos conversations ou identifier les personnes présentes dans notre foyer ?

L'exemple précédent soulève une autre question : celle de l'utilisation des signaux de parole dans le cadre juridique. Dans une récente publication scientifique[3], des chercheurs d'INTERPOL ont présenté les résultats d'une enquête réalisée sur l'utilisation de la reconnaissance du locuteur par les forces de l'ordre à travers le monde. L'article met clairement en avant l'extrême prudence qui doit entourer le recours à ces méthodes. En France, depuis 1997, la Société française d'acoustique (SFA) appelle publiquement à ne pas recourir à l'expertise en matière de reconnaissance du locuteur dans le domaine judiciaire. Il s'agit là de se prémunir des promesses de ce que le philosophe Evgeny Morozov appelle le solutionnisme technologique[4] et d'indiquer clairement que si le domaine du traitement de la parole connait d'indéniables avancées, en l'état actuel des connaissances en matière d'identification vocale, il n'existe aucune méthode scientifique qui permette d'identifier une personne avec certitude.

La voix est considérée par le Droit comme une image sonore de la personne et doit à ce titre être protégée comme les autres attributs de la personne humaine. Cette vision est consacrée par l'article 9 du Code Civil. De plus, au regard de la loi Informatique et Libertés, notre voix peut tour à tour être considérée comme une donnée biométrique, une donnée de santé (car révélatrice de pathologies), une donnée sensible (car faisant apparaître des opinions politiques, philosophiques ou religieuses), *etc.* A la lueur des précédentes réflexions et des transformations induites par le numérique, il semble donc indispensable de repenser le statut juridique de la voix et des données de parole. De plus, d'un point de vue technique, les questions de loyauté des systèmes et d'appréciation de leurs capacités doivent également être posées au risque de constater une crise de confiance des utilisateurs de ces technologies.

Félicien Vallet, Ingénieur au Service de l'expertise technologique de la CNIL

---

[2] [Journal Le Monde] "Dans l'Arkansas, la police veut entendre « Alexa », l'assistant à commande vocale d'Amazon", 30 décembre 2016.
[3] [Morrisson *et al.*] "INTERPOL survey of the use of speaker identification by law enforcement agencies", Forensic Science International, Volume 263, Juin 2016.
[4] [Morozov] "Pour tout résoudre, cliquez ici ! L'aberration du solutionnisme technologique", FYP Editions, 2014.