

Écosse, à l'université de Carnegie Mellon aux États-Unis, à l'Institut des sciences et technologies de Nara au Japon... Parmi ces acteurs mondiaux, l'Ircam (Institut de recherche et coordination acoustique-musique), à Paris, se distingue par le lien qu'il entretient entre l'innovation technologique et le monde de la culture, du divertissement et de la création de contenus (musique, cinéma, jeux vidéo, etc.).

Alors que les géants du numérique et d'autres laboratoires se sont longtemps contentés de voix synthétiques intelligibles et standardisées pour des usages spécialisés, les chercheurs de l'Ircam ont depuis les années 1980 focalisé leurs efforts sur ce qui rend une voix humaine – son grain, son ton, sa personnalité, son expressivité.

Au cours de la dernière décennie, grâce notamment aux travaux de l'équipe Analyse et synthèse des sons, de l'Ircam, on est passé de voix de synthèse intelligibles, mais mécaniques, à des voix de synthèse naturelles et expressives. Cette rupture a été rendue possible par un ensemble d'avancées scientifiques réalisées en traitement du signal, en apprentissage automatique et en linguistique. Aujourd'hui, nous reproduisons plus fidèlement le grain de la voix grâce à

■ LES AUTEURS



Nicolas OBIN est maître de conférences à l'université Pierre-et-Marie-Curie (Sorbonne universités) et chercheur dans l'équipe Analyse et synthèse des sons de l'Ircam, à Paris.



Axel ROEBEL est chercheur et responsable de l'équipe Analyse et synthèse des sons de l'Ircam.

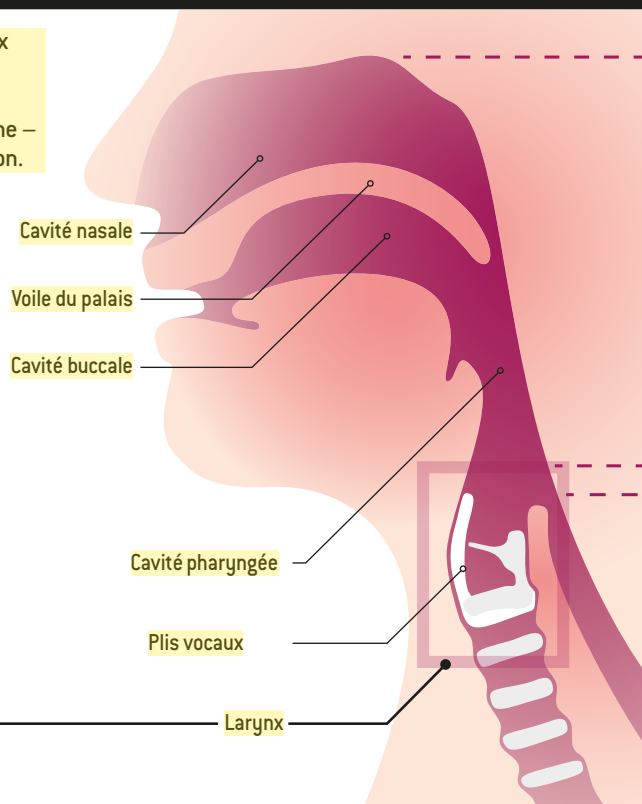
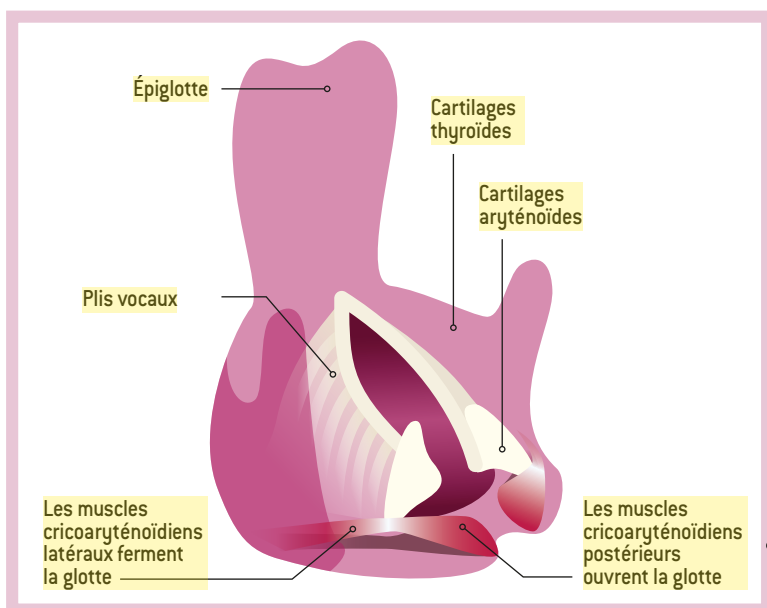
l'amélioration des algorithmes d'analyse et de synthèse de la parole. Par ailleurs, nous maîtrisons mieux la prosodie de la voix et son lien avec la syntaxe d'un texte, ce qui a considérablement amélioré le naturel des voix de synthèse. Enfin, les méthodes d'apprentissage automatique et l'augmentation de la capacité des ordinateurs ont été déterminantes pour concrétiser ces avancées et apprendre à la machine à parler (les bases de données contenant dorénavant des heures d'enregistrements pour une voix, et ce pour des milliers de voix différentes).

Aujourd'hui, des logiciels développés à l'Ircam sont utilisés dans des productions cinématographiques. Ils permettent de sculpter la voix pour en modifier son grain, son ton, son identité et sa personnalité. Nous transformons facilement une voix d'homme en celle d'une femme, une voix jeune en une voix plus âgée, et inversement. Nous sommes également capables de convertir notre voix pour imiter celle d'une autre personne. Nos ordinateurs et nos téléphones sont dotés de voix de synthèse naturelles, expressives et personnalisables, capables de prononcer n'importe quel texte.

Demain, il sera possible de rendre leur voix à des personnes ayant perdu l'usage

COMMENT LA VOIX PREND FORME

Les organes intervenant dans la parole sont nombreux. On les sépare en deux groupes. Les poumons et la trachée excitent, par leurs mouvements, les plis vocaux et les font vibrer. Cela détermine les fréquences sonores émises. Le corps de résonance – du larynx à la cavité nasale, en passant par la bouche – a pour effet de moduler l'amplitude associée à chacune des fréquences du son.



de la parole (voir l'encadré page 61), de parler dans une autre langue, mais avec sa propre voix (voir l'encadré page 60), de changer de voix comme on change de coiffure, d'interagir au quotidien avec des robots dotés d'une personnalité vocale indiscernable de celle d'un être humain. Ces progrès scientifiques seront sans doute au cœur d'une bataille technologique, commerciale et éthique. Mais pour y parvenir, les chercheurs devront bien comprendre la singularité de la voix humaine et la maîtriser.

Comment la voix est-elle produite ? La source d'émission est constituée de plis vocaux – nommés aussi, de façon trompeuse, cordes vocales (ils ne ressemblent pas à des cordes). Ces plis, situés dans la gorge au niveau de la glotte, sont mis en vibration par l'air que les poumons insufflent à travers la trachée. Ces vibrations entraînent des variations de la pression de l'air, qui se propagent et produisent un son à une certaine fréquence dite fondamentale, la hauteur de la voix, et ses multiples, les harmoniques (voir l'encadré ci-dessous).

L'irrégularité des vibrations, la tension des muscles autour des plis vocaux, la quantité d'air qui circule ainsi que les turbulences d'air engendrées modifient la « qualité » du

son émis (voix douce ou tendue, pressée, soufflée, craquée, rauque, cassée, etc.).

Le son émis au niveau de la glotte se propage à travers le conduit vocal (la trachée, puis les cavités buccales et nasales), lequel sert de corps de résonance, amplifie et colore le son. L'effet produit est celui d'un « filtre » qui module l'amplitude des composantes fréquentielles du son. La voix acquiert ainsi une coloration, un « timbre » particulier. L'anatomie propre à chaque personne rend son timbre unique, qui constitue une sorte de signature vocale.

Qui plus est, la coloration du son émis dépend de la configuration des articulateurs de la bouche (les lèvres, la langue et le voile du palais). Le contrôle de ces configurations fait de l'appareil vocal un instrument de communication. Ainsi, chaque langue est pourvue d'un ensemble de sons spécifiques nommés phonèmes, qui en constituent les unités élémentaires. Le français en compte un peu plus d'une trentaine : les voyelles et sons apparentés, comme dans « on » ou « un », les consonnes et sons apparentés (occlusives comme *b*, *d* ou *p*, fricatives comme *f*, *ch* ou *s*, etc.).

Lorsqu'une personne parle, elle est capable de contrôler et de modifier en

permanence la configuration de son appareil vocal. Ainsi, les articulateurs de la bouche se déplacent en permanence et articulent les phonèmes pour former des syllabes, des mots, des phrases. Également, la dynamique de la glotte et des plis vocaux permet de contrôler la prosodie, ou « musique de la voix ». Elle se traduit par des variations de hauteur, d'intensité, de rythme et de qualité vocale. Son évolution au cours du temps construit, organise et module le sens des mots et des phrases, indique nos origines géographiques et socioprofessionnelles, et traduit nos intentions et nos émotions.

Sculpter la voix

L'identité d'une voix est donc constituée de deux composantes propres à chaque individu : le timbre et la prosodie. Pour changer l'identité ou la personnalité d'une voix, il faut alors être capable de modéliser ces composantes afin de pouvoir les reproduire et les modifier. À cette fin, les chercheurs de l'Ircam ont développé en 2010 le logiciel nommé IrcamTools TRAX. À partir d'un signal sonore issu d'un enregistrement ou d'un microphone, des traitements numériques modifient la hauteur de la voix pour obtenir une voix plus grave ou plus aiguë, ou le filtre de la voix pour en modifier le timbre. Il est alors très facile de transformer une voix d'homme en voix de femme (et inversement), la voix d'un adulte en une voix d'adolescent ou de personne âgée...

En effet, les hommes et les femmes ont des caractéristiques physiologiques (taille, corpulence) qui imposent des différences au niveau de leur appareil phonatoire, et donc de leur voix. Les hommes ont des plis vocaux plus longs (de 17 à 25 millimètres) et un conduit vocal plus grand (17 centimètres en moyenne). La conséquence est que les voix d'homme sont caractérisées par une faible hauteur (de 70 à 160 hertz) et un timbre « grave ». Les femmes ont des plis vocaux plus courts (12,5 à 17,5 millimètres) et un conduit vocal plus court (14 centimètres en moyenne). Une voix de femme sera donc plus haute (de 130 à 300 hertz) et son timbre sera plus « aigu ». En outre, la voix d'une personne évolue au cours de sa vie, à l'instar de son corps. Les adolescents ont une voix aiguë qui devient plus grave à l'âge adulte, puis qui remonte au cours de la vie.

Ainsi, agir au moyen d'un logiciel sur la hauteur et le timbre d'une voix crée l'illusion

