Cette technique a été utilisée en première mondiale pour reconstruire les voix de Marilyn Monroe (dans le film *Marilyn*, de Philippe Parreno, 2012), du maréchal Pétain (dans le documentaire *Juger Pétain*, de Philippe Saada, 2014), et de Louis de Funès (dans le film *Pourquoi j'ai pas mangé mon père*, de Jamel Debbouze, sorti en 2015). Un comédien dont la mission était d'imiter la prosodie et le jeu de Louis de Funès sur le texte du personnage dans le film a été enregistré en plusieurs séances. Parallèlement, une base de données d'une dizaine de minutes de la voix de Louis de Funès a été constituée à partir d'une collection d'enregistrements historiques. Les chercheurs ont alors pu convertir le timbre du comédien à partir de la signature vocale de Louis de Funès. L'effet est saisissant : Louis de Funès semble avoir été enregistré aujourd'hui, mais avec sa voix des films des années 1970.

Un puzzle vocal

Les chercheurs de l'Ircam repoussent encore les limites de la manipulation vocale grâce à la synthèse de la parole. Ici, il n'est plus question de transformer une voix à partir d'un enregistrement, mais bien de créer une voix de synthèse capable d'exprimer n'importe quel texte *(voir l'encadré page 59)*.

Les premières tentatives de synthèse de la parole remontent au XVIII[e] siècle, le siècle des automates et des machines parlantes. Avec l'avènement de l'électricité, au XX[e] siècle, les machines se modernisent et se perfectionnent (tel le « Voder », pour *Voice Operation Demonstrator*, fabriqué en 1939 par les laboratoires Bell), puis s'automatisent avec l'arrivée des ordinateurs. Ainsi, le synthétiseur de l'IBM 704, conçu par les laboratoires Bell en 1962, est connu pour avoir inspiré la voix de l'intelligence artificielle HAL dans le film *2001, l'Odyssée de l'espace* de Stanley Kubrick.

Aujourd'hui, les systèmes de synthèse de la parole sont totalement automatiques et offrent la possibilité de personnaliser les voix numériques. La synthèse de la parole repose essentiellement sur l'utilisation de bases de données de voix, et plusieurs heures d'enregistrement sont nécessaires pour créer une voix de synthèse. Car contrairement à la conversion de voix, la synthèse ne se limite pas à modifier les caractéristiques d'une voix existante. Il faut avoir à disposition toutes les briques élémentaires pour reconstituer intégralement le langage et prononcer n'importe quel texte. La plupart des synthétiseurs actuels fonctionnent par « sélection d'unités » dont le principe est similaire à un jeu de puzzle.

Chaque pièce est un morceau de voix (phonème, syllabe, mot, etc.) dont la prosodie et le timbre lui sont spécifiques. La base de données sert alors de réservoir de pièces de puzzle qui doit couvrir au mieux la variabilité acoustique de la voix à créer. Pour synthétiser un texte, il faut alors chercher et réassembler les pièces destinées à reconstituer le texte. La construction du puzzle relève cependant du casse-tête : des algorithmes trient une immense masse de pièces de puzzle, chacune ayant une signature sonore spécifique qui n'autorise pas à les accoler simplement les unes aux autres. La réalisation du puzzle textuel nécessite de trouver les pièces qui, mises bout à bout, se correspondent au mieux.

Afin qu'une voix de synthèse paraisse naturelle, il est indispensable de s'assurer de l'intelligibilité du texte en associant parfaitement les phonèmes. Il faut également s'assurer de la forme d'ensemble des pièces du puzzle, la prosodie. La qualité de la restitution dépend de la taille des bases de données : le grand nombre de pièces sonores disponibles pour chaque fragment de voix garantit la bonne connexion des

Un interprète dans votre poche

De voyage au Japon, vous souhaitez converser avec un habitant. Vous dégainez votre téléphone, parlez en français dans le microphone et votre voix ressort du haut-parleur quasi instantanément en japonais ! Votre interlocuteur fait de même en vous répondant dans sa propre langue traduite en français par votre téléphone.

Ce scénario est d'ores et déjà possible grâce aux logiciels d'interprétation qui ont vu le jour en 2009 au sein du projet européen EMIME (www.emime.org) et développés parallèlement par IBM, Google ou Mobile Technologies avec l'application Jibbigo.

Leur fonctionnement repose sur une chaîne d'actions. Le message vocal capté par le micro est tout d'abord retranscrit sous forme de texte par un module de transcription. Le texte est alors traduit vers la langue cible par un module de traduction avant d'être prononcé par un module de synthèse et finalement émis par le haut-parleur du téléphone. Le module de traduction réalise la traduction à partir de règles, de la même manière que le ferait un interprète humain. Cependant, ces règles sont difficiles à formuler explicitement. Ainsi, plutôt que de fournir au module un ensemble limité de règles à appliquer, celles-ci sont apprises par un ordinateur à partir d'un ensemble composé d'une multitude d'exemples de textes écrits dans les deux langues. De la même façon que Champollion eut appris à décrypter les hiéroglyphes, en partie grâce à la pierre de Rosette sur laquelle était gravé un même texte écrit en trois langues, l'ordinateur va par lui-même déduire les méthodes pour passer d'une langue à l'autre. Contrairement à la pierre de Rosette, la quantité d'exemples disponibles pour l'apprentissage de ces règles, collectés entre autres sur Internet, est quasi illimitée et grandit de jour en jour.

De même, les modules de transcription et de synthèse sont en mesure de retranscrire la parole en texte et de le prononcer au moyen de méthodes apprises à partir de livres audio par exemple, mais ce n'est pas tout. Lorsque nous rencontrons une personne pour la première fois, si elle a un fort accent et si l'environnement est bruyant, la compréhension est difficile dans les premières minutes de discussion, mais s'améliore rapidement. Ainsi, à partir de quelques secondes de parole, le module de transcription s'ajuste à la voix de l'utilisateur, à sa manière de parler, mais également à l'environnement acoustique (lieu calme ou bruyant) afin d'améliorer ses performances. Cette capacité d'adaptation est également partagée par le module de synthèse qui produit de la parole à partir d'un texte avec une voix ayant des caractéristiques acoustiques semblables à celle de l'utilisateur. Nous sommes dorénavant doués pour parler le japonais (ou toute autre langue) sans l'avoir jamais appris…

– Pierre Lanchantin
Machine Intelligence Laboratory, université de Cambridge

phonèmes et la richesse prosodique de la voix. Le rendu final est obtenu par des retouches locales automatiques qui visent à rendre les connexions imperceptibles et à lisser le phrasé.

Les avancées scientifiques réalisées depuis le début des années 2000 en linguistique, en traitement du signal et en apprentissage automatique ont conduit à une amélioration sans précédent de la modélisation de la prosodie en synthèse de la parole. Nous sommes ainsi passés d'une voix de synthèse intelligible à une voix de synthèse naturelle et expressive.

La frontière entre le réel et l'artificiel se réduit, à l'exemple de la voix de l'acteur André Dussolier que nous avons recréée (dans le cadre de l'exposition « La Voix » organisée en 2014 à Paris, à la Cité des sciences et de l'industrie). La voix réelle et la voix artificielle sont devenues indifférenciables, au dire de l'acteur lui-même.

À partir de milliers de voix différentes, et grâce aux possibilités offertes par l'apprentissage automatique, il est également possible de créer une voix artificielle « moyenne », de créer des voix hybrides à partir de plusieurs voix et d'adapter le style, l'accent, ou l'émotion d'une voix.

Le principe est similaire à celui de la synthèse par sélection d'unités, mais la différence est considérable. Au lieu d'utiliser les morceaux réels d'une voix pour créer une voix de synthèse, on représente

À la recherche de la parole perdue

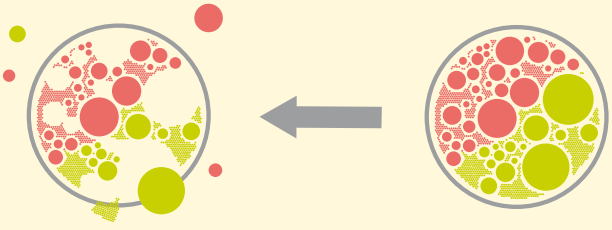
Certaines maladies neurologiques dégénératives, liées au vieillissement ou d'origine génétique, affectent la parole jusqu'à la rendre complètement inintelligible alors que les patients conservent toutes leurs facultés intellectuelles et cognitives.

Ainsi par exemple, le physicien Stephen Hawking est atteint d'une sclérose latérale amyotrophique, aussi nommée maladie de Charcot, dont les conséquences sont la dégénérescence des motoneurones du cortex cérébral et qui l'empêche d'articuler le moindre son. Il communique grâce à l'un des premiers systèmes de synthèse de la parole, conçus dans les années 1970. Le timbre robotique de la voix de synthèse utilisée par Stephen Hawking a fini par s'identifier complètement à la figure du physicien.

Cependant, de nouvelles techniques créent désormais des voix de synthèse personnalisées qui conservent les principales caractéristiques de l'identité vocale du patient telles que le timbre, l'intonation ou l'accent. Ces techniques utilisent un modèle simplifié de la production vocale. Selon ce modèle, la voix est le produit d'une excitation (souffle issu des poumons et vibration des cordes vocales) filtrée par les résonances du conduit vocal. Les différents paramètres tels que la fréquence de vibration des cordes vocales ou les résonances du conduit vocal peuvent être représentés, en fonction de chaque phonème prononcé et de son contexte linguistique, grâce à des modèles statistiques appris à partir des enregistrements de la voix. Ces modèles statistiques caractérisent en quelque sorte l'empreinte vocale du locuteur. Grâce à de puissants outils d'apprentissage automatique, les algorithmes apprennent à modéliser une voix sur la base de quelques minutes d'enregistrement seulement. Cependant, il arrive fréquemment qu'une personne ne prenne conscience de la nécessité d'enregistrer sa voix que lorsque celle-ci commence à être dégradée par la maladie. Le défi est alors d'être capable de recréer une voix de synthèse claire et intelligible à partir d'une source dégradée ou peu intelligible. L'équipe du CSTR, à l'université d'Édimbourg, est récemment parvenue à reconstruire la voix d'un patient en utilisant certaines parties de la voix d'un locuteur sain, qualifié de « donneur », et dont les caractéristiques vocales étaient suffisamment proches de celle du patient. Cette opération s'effectue non pas directement sur les enregistrements des voix, mais en « mélangeant » en partie les modèles statistiques appris à partir de ces enregistrements.

Le principe, illustré par la figure ci-dessous, repose ainsi sur le remplacement de certaines composantes « défectueuses » du modèle statistique de la voix du patient par celles du donneur. Des modules d'analyse statistique sont utilisés pour détecter les parties à corriger. Cette méthode, testée sur plus d'une cinquantaine de patients, en collaboration avec la clinique Anne Rowling à Édimbourg, utilise le plus souvent les voix des proches du patient (frère, sœur ou enfants). Pour tous ces patients, l'utilisation d'une voix de synthèse personnalisée leur a permis de conserver leur identité vocale et de jouir d'une qualité de communication améliorée avec leur entourage.

– Christophe Veaux
Centre for Speech Technology Research, université d'Édimbourg



LA RECONSTRUCTION DE LA VOIX consiste à remplacer les composantes biaisées du modèle de voix du patient *(à gauche)* par leurs homologues du modèle du locuteur sain *(à droite)*. Par exemple, si certains phonèmes sont mal articulés, on modifiera le modèle de voix du patient pour l'articulation de ces phonèmes en prenant comme référence les valeurs correspondantes du modèle du locuteur sain.