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

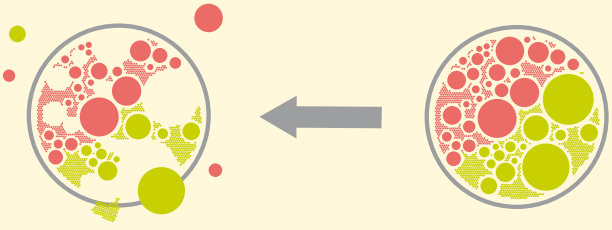
Certaines maladies neurologiques dégénératives, liées au vieillissement ou d'origine génétique, affectent la parole jusqu'à la rendre complètement inintelligible alors que les patients conservent toutes leurs facultés intellectuelles et cognitives.

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.

À qui appartiendra une voix de synthèse **créée à partir de voix humaines ?**

intégralement la voix sous la forme d'un modèle statistique, qui en réalise une abstraction mathématique capable de reproduire et de régénérer l'intégralité d'une voix de synthèse. Ainsi, la loi de distribution statistique (moyenne et variance pour une distribution gaussienne) permet de modéliser la répartition de chaque phonème dans l'espace acoustique des paramètres de la voix (hauteur, durée, intensité, timbre).

Afin de modéliser l'évolution temporelle des paramètres de la voix, des modèles statistiques séquentiels (tels que les « chaînes de Markov cachées ») sont par ailleurs utilisés : chaque phonème est alors segmenté en « états », par exemple début, milieu et fin, chaque état étant soumis à sa propre loi statistique.

Le formalisme statistique permet alors de manipuler ces « abstractions » par combinaison, interpolation et adaptation des paramètres statistiques dans l'espace des voix. On peut par exemple créer une voix hybride à partir des paramètres statistiques de deux voix réelles, de créer une voix moyenne résultant de la combinaison de milliers de voix, etc. Cette évolution technique étend considérablement les possibilités de la synthèse de la parole à partir du texte : elle ne se limite plus nécessairement à une voix réelle et s'adapte alors rapidement pour créer une nouvelle voix de synthèse à partir de quelques minutes d'enregistrement seulement. Ainsi, il est déjà possible de recréer la voix d'une personne ayant perdu l'usage de la parole à partir de quelques minutes d'enregistrement de sa voix *(voir l'encadré page 61)*, ou de faire parler une personne dans une autre langue avec sa propre voix *(voir l'encadré page 60)*.

Pour impressionnants que soient les progrès, les voix de synthèse restent encore imparfaites et l'intervention humaine est toujours nécessaire pour obtenir une bonne qualité de conversion et de synthèse. La révolution en cours de l'intelligence artificielle, de l'apprentissage profond par réseaux de neurones artificiels et des données massives devrait entraîner une nouvelle vague d'avancées.

Dans la technique des réseaux de neurones artificiels, ou réseaux neuromimétiques, le dispositif matériel ou virtuel qui « apprend » est constitué de couches d'éléments, ayant chacun deux états possibles, connectées les unes aux autres par des liens dont les caractéristiques sont modifiées au cours du processus d'apprentissage par un algorithme approprié. De tels réseaux, censés imiter les processus naturels d'apprentissage se déroulant dans le cerveau, ont fait leur entrée dans l'apprentissage automatique dans les années 1970. Mais leur développement s'est heurté à des limitations théoriques et algorithmiques, ainsi qu'aux faibles capacités des machines de l'époque.

L'apprentissage profond est en marche

Les avancées théoriques réalisées au cours de la dernière décennie et l'augmentation considérable de la puissance des ordinateurs ont remis cette technique sur le devant de la scène. Ainsi, de nouveaux algorithmes d'apprentissage ont été conçus pour des réseaux de neurones « profonds » (constitués de nombreuses couches, chacune formée de nombreux neurones), l'apprentissage portant sur une quantité massive de données. Ces techniques laissent espérer que, dans les prochaines décennies, on créera des voix numériques indiscernables des voix réelles, dans n'importe quelle langue, et personnalisables selon les besoins.

Demain, nous sculpterons notre propre voix et nous interagirons quotidiennement avec des machines intelligentes dont la voix sera indiscernable de celle d'un humain – cet ultime « miroir de l'âme ». *Deus* ou *diabolus ex machina* ? Ces avancées ne sont pas sans contrepartie et soulèvent des questions fondamentales sur la place des voix de synthèse et des machines humanisées dans nos sociétés.

Ainsi, à qui appartiendra une voix de synthèse créée à partir de voix humaines ou obtenue par conversion d'une autre voix ? À l'individu imité ? À l'individu transformé ? Aux chercheurs et aux ingénieurs qui ont rendu possible cette réalisation ? Et comment distinguer la voix synthétique de la voix réelle ? Comment authentifier l'auteur d'un message dont on peut contrefaire la voix ? L'humanisation des voix de synthèse, à l'instar de l'apparence visuelle des robots, pose également question. Avec des machines pourvues de voix trop humaines, ne faisons-nous pas un pas dans la « vallée dérangeante » dont nous avertissait le roboticien japonais Masahiro Mori ? ■

BIBLIOGRAPHIE

A. van den Oord *et al.*, **WaveNet : A generative model for raw audio**, *Proceedings of Interspeech*, San Francisco, 2016 (https://deepmind.com/blog/wavenet-generative-model-raw-audio/).

K. Tokuda *et al.*, **Speech synthesis based on hidden Markov models**, *IEEE Transactions on Audio, Speech, and Langage Processing*, vol. 101(5), p. 1234-1252, 2013.

N. Obin, **MeLos : Analysis and modeling of speech prosody and speaking style**, thèse de doctorat, Ircam-UPMC, 2011.

J. W. Mullennix et S. E. Stern (éd.), **Computer Synthesized Speech Technologies : Tools for Aiding Impairment**, IGI Global, 2010.