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.