FUTURA

# Cette IA imite votre voix à partir de seulement 3 secondes d'enregistrement

## Podcast écrit par et lu par Emma Hollen

[*Générique d'intro, une musique énergique et vitaminée.*]

Une IA qui imite votre voix à la perfection à partir de seulement 3 secondes d'enregistrement, c'est l'actu de la semaine, dans Vitamine Tech.

[*Fin du générique.*]

On connaît toutes et tous WALL-E, le petit robot imaginé par Pixar, destiné à compacter *ad nauseam* les déchets laissés sur Terre par l'humanité. C'est de son nom qu'a été inspiré celui de DALL·E, un générateur d'images qui, aux côtés de plusieurs autres, fait depuis quelques temps grincer les dents des artistes, tant de par ses capacités créatives impressionnantes que de par sa tendance à recourir au plagiat d'œuvres pré-existantes. Mais celui-ci qui nous intéresse aujourd'hui, c'est VALL-E, la dernière création de l'équipe de recherche de Microsoft. Son super-pouvoir : imiter une voix humaine à partir de seulement 3 secondes d'échantillon audio, le tout avec la bonne émotion.

[*Une musique électronique calme.*]

Oui, 3 secondes d'audio seulement, vous avez bien entendu ! On n'arrête pas le progrès, pour le meilleur comme pour le pire – plutôt pour le pire d'ailleurs, mais ça on y viendra dans un instant. VALL-E appartient à la catégorie de ce que l'on appelle des « synthétiseurs vocaux ». Une technologie qui ne date pas d'hier, puisque c'est le mathématicien Leonhard Euler qui en pose les bases au XVIIIe siècle. La première mise en pratique, pour sa part, remonte à 1784, avec une sorte d'instrument de verre capable de reproduire la voix humaine. Alors évidemment, c'était assez rudimentaire à l'époque, et c'est au XXe siècle que la synthèse vocale prend véritablement son envol. Dans les années 30, les laboratoires Bell créent le Voder, un système permettant de recréer des mots humains en modulant des fréquences via un clavier manuel. [*Un homme demande «* Who saw you? *» ; la machine répond «* She saw me *». Puis l'homme demande «* Who did she see? *», ce à quoi la machine répond «* She saw me *», illustrant la capacité du Voder à modifier son intonation pour mettre l'emphase sur un mot ou un autre.*] Puis en 1961, l'un des chercheurs de Bell, John Larry Kelly, parvient à faire chanter « Daisy Bell » à son ordinateur. [*La chanson est chantée par une voix robotique et accompagnée d'une musique numérique composée par un humain.*] Si la chanson vous dit quelque chose, c'est probablement que vous êtes un fan de *2001, l'Odyssée de l'espace*, puisque, impressionné par la démonstration de Kelly, Arthur C. Clarke avait décidé d'inclure la mélodie dans la scène iconique du film où l'astronaute

Dave Bowman déconnecte les blocs mémoire du redoutable HAL 9000. Si l'intelligence d'HAL s'éteint sur les notes de « Daisy Bell », la synthèse vocale, elle n'en est qu'à ses débuts et continue de se perfectionner en s'appuyant sur des méthodologies et des technologies toujours plus poussées. Aujourd'hui, son alliée principale est bien évidemment l'IA et plus particulièrement les systèmes dits d'apprentissage profond ou *deep learning*. Ces derniers s'entraînent à produire du langage en s'appuyant sur un gigantesque répertoire d'enregistrements audio, accompagnés de leurs textes écrits. Grâce à cette ingestion massive d'information, les synthèses vocales dites « neuronales » s'approchent toujours plus de notre façon de parler, au point qu'il devient de plus en plus difficile de distinguer l'artificiel du naturel. Et il semblerait que les chercheurs aient réussi à accomplir un nouveau bond de géant avec VALL-E. Celui-ci fonctionne un peu différemment de ses prédécesseurs. Sans trop entrer dans le détail, lorsque vous présentez un texte à une synthèse vocale, sa première mission consiste à aller chercher la prononciation de chaque mot pour produire un transcript phonétique de ce que vous voulez lui faire lire. Il en résulte une série de phonèmes, /*a*/, /*s*/, /*d*/, qui possèdent chacun une signature sonore bien spécifique. Cette signature est non seulement distincte à l'oreille mais elle l'est également visuellement. On la représente sur un spectrogramme, une sorte de heatmap où le son est converti en image. En ordonnées, les fréquences que mobilise le son produit, en abscisses, le temps, puisque même le phonème le plus simple emploie plusieurs combinaisons de fréquences à la suite les unes des autres, et en couleur, l'amplitude, c'est-à-dire le volume de chaque fréquence. Parce qu'une image vaut mille mots, je vous invite à chercher le spectrogramme de l'alphabet sur votre moteur de recherche de prédilection ; vous verrez que chaque lettre possède sa propre empreinte digitale, composée d'un assemblage de fréquences émises à des amplitudes variées, sur une durée donnée. À noter que cette empreinte sera également amenée à varier en fonction de la voix du locuteur, de son émotion, de l'intonation, de l'emplacement du phonème dans la phrase, ou dans un mot. Quoi qu'il en soit, ce spectrogramme est ensuite converti en onde sonore et c'est ainsi que nous obtenons un texte lu par une voix de synthèse.

[*Virgule sonore, une cassette que l'on accélère puis rembobine.*]
[*Une musique de hip-hop expérimental calme.*]

VALL-E, pour sa part, emploie une autre méthode, car son objectif n'est pas seulement de produire de l'audio à partir d'un texte, mais de le faire en imitant votre voix. Pour ça, le système d'apprentissage profond a été abreuvé de 60 000 heures d'enregistrement en langue anglaise, produites par plus de 7 000 voix différentes, et accompagnées de leur texte. En digérant cette immense base de données, il identifie et extrait des composants discrets, des espèces de blocs élémentaires, que l'on appelle *tokens*. C'est un peu comme un enfant qui apprend à isoler de nouveaux mots en écoutant les adultes parler, si ce n'est que dans le cas de VALL-E, les tokens qu'il identifie sont découpés sur une échelle bien plus petite et plus précise. Ils sont enregistrés sous la forme d'un code qui pourra, à son tour, être restitué en son. Bien. Dans l'étude, les chercheurs fournissent un nouvel échantillon de voix à VALL-E, ainsi qu'un texte qu'il devra lire en imitant cette voix. Le système va analyser le nouvel enregistrement et le découper en tokens comme précédemment, et il va en faire de même pour le texte écrit. Mais cette fois-ci, il va tenter de les rattacher aux tokens dont il dispose déjà dans sa bibliothèque. En identifiant les motifs qui rapprochent le nouvel enregistrement de ceux qu'il a déjà étudiés, il va pouvoir extrapoler la manière dont le locuteur s'exprimerait dans toutes sortes de contextes. Ainsi, s'il respecte la première étape

de la synthèse vocale en convertissant bel et bien le texte qu'on lui soumet en phonèmes à lire, il ne génère pas de spectrogramme à partir du néant, mais va plutôt piocher les pièces dont il a besoin dans la base de tokens qu'il a assimilée. Pour résumer et pour faire simple, il assemble un puzzle de micro-enregistrements humains au lieu de fabriquer de toutes pièces un son de synthèse. Et ça s'entend dans les résultats. Si certains n'arrivent pas à dissimuler des sonorités artificielles, d'autres sont confondants de réalisme et l'imitation se révélera d'autant plus parfaite que la voix se rapprochera de celles qui ont déjà été étudiées par le système. Voici par exemple un échantillon de voix qui lui a été fourni : [*une voix masculine humaine : «* We live by the rule of law *»*], et la phrase qu'il a produite en imitant cette voix et son intonation : [*une voix identique à la précédente : «* Because we do not need it *»*]. Maintenant on les met bout à bout : [*«* We live by the rule of law because we do not need it *» ; la phrase donne l'impression d'être énoncée par une seule et même personne*.] Comme vous pouvez le voir, c'est plutôt impressionnant. Alors, forcément, certains s'inquiètent, à commencer par les créateurs de VALL-E eux-mêmes. En conclusion de leur étude, les chercheurs écrivent : « *Étant donné que VALL-E est capable de synthétiser une parole qui conserve l'identité du locuteur, le modèle comporte des risques potentiels d'utilisation abusive, comme tromper une reconnaissance vocale ou usurper l'identité d'un locuteur spécifique.* » Pour limiter ces risques, l'équipe a donc choisi de ne pas partager publiquement le code de VALL-E. Ils évoquent également la nécessité de créer un modèle de détection qui pourra signaler tout audio généré artificiellement par leur synthèse vocale. Car si les deepfakes vidéo ont déjà démontré les dangers qu'ils comportent, les deepfakes audio, croyez-moi, possèdent quant à eux des ramifications encore plus inquiétantes. L'année dernière, des comédiens russes ont réussi à tenir une discussion de 7 minutes avec le président polonais en se faisant passer pour Emmanuel Macron, alors qu'un missile venait d'exploser à la frontière ukrainienne. Leur piètre imitation de l'accent français leur a valu de se faire finalement démasquer par Andrzej Duda, mais un outil comme VALL-E pourrait devenir une arme dangereuse entre les mains de ces imposteurs, dans un contexte politique pour le moins vacillant. Saluons donc la posture éthique des chercheurs qui ont choisi de garder leur innovation derrière un garde-fou en attendant que ses risques potentiels puissent être mitigés.

[*Virgule sonore, un grésillement électronique*.]

C'est tout pour cet épisode de Vitamine Tech. Pour ne pas manquer nos futurs épisodes, pensez à vous abonner dès à présent à ce podcast et si vous le pouvez, laissez-nous une note et un commentaire. C'est toujours un plaisir de vous lire et de pouvoir connecter avec vous. Cette semaine, je vous invite à faire un tour de nos productions : famille, actus, histoire, astronomie, environnement, il y en a pour tous les goûts et pour tous les âges, donc n'hésitez pas à chercher Futura sur vos apps de prédilection ou Futura Podcasts sur les moteurs de recherche. Pour le reste, je vous souhaite une excellente journée ou une très bonne soirée, et je vous dis à la semaine prochaine, dans Vitamine Tech.

[*Un glitch électronique ferme l'épisode*.]