Le laboratoire parisien d'IA Kyutai lance un modèle de traitement de la voix

Financé par des mécènes, Moshi a vocation à offrir une solution française et open source aux assistants vocaux des géants de l'intelligence artificielle OpenAI ou Google.

Par Alexandre Piquard

Publié le 03/07/2024 à 20h40, modifié à 08h48

Lors de la présentation de Moshi, à Paris, le 3 juillet 2024. KUYTAI

Kyutai lance Moshi. Ces deux noms mignons mais cryptiques empruntent aux mots japonais « sphère » et « allô ». L'un désigne un laboratoire parisien d'intelligence artificielle (IA) fondé en novembre, et l'autre, son premier outil rendu public, un modèle de traitement de la voix.

Comme ChatGPT ou Gemini pour le texte, comme Dall-E ou Midjourney pour les images, celui-ci se place sur le terrain des assistants virtuels, mais vocaux. Capable de décrypter une instruction orale et de générer une réponse dans un style conversationnel, Moshi se veut une alternative aux outils équivalents d'OpenAI (le créateur de ChatGPT), Google ou Apple (Siri). Mais elle se revendique fabriquée en France et open source, c'est-à-dire utilisable et modifiable librement.

Moshi est la première publication de Kyutai depuis son lancement en grande pompe, le 17 novembre. Ce laboratoire doté de 300 millions d'euros est atypique sur la scène parisienne, car il est financé par des mécènes : les Français Xavier Niel (fondateur de l'opérateur télécoms Iliad et actionnaire à titre personnel du Groupe Le Monde) et Rodolphe Saadé, PDG de l'armateur CMA-CGM, ainsi que l'Américain Eric Schmidt, ex-PDG de Google devenu investisseur.

Un modèle « hybride » financé par des mécènes

Il a été créé en débauchant six chercheurs en IA issus des géants américains comme Meta ou Google DeepMind. Son projet est de *« fabriquer des modèles de fondation en IA innovants et*

de les publier, résume son directeur, Patrick Pérez. *L'idée à l'origine de Kyutai est de créer un hybride bénéficiant du meilleur des deux mondes, la recherche académique pour sa liberté et le milieu de l'entreprise pour ses moyens. »*

Moshi se veut donc innovant, même par rapport à la concurrence mondiale. Kyutai a choisi le domaine du son, moins occupé que celui des modèles de génération de texte (où opèrent déjà OpenAI, Google ou Anthropic, mais aussi les français Mistral ou LightOn). *« D'ordinaire, les IA vocales utilisent plusieurs modèles successifs : l'un pour détecter la présence d'une instruction sonore, un autre pour la transcrire en texte, un autre pour comprendre la requête, un autre pour produire la réponse et un dernier pour la transformer en voix. Mais cela produit une latence de trois à cinq secondes, désagréable dans une conversation »*, explique le chercheur Neil Zeghidour, qui a travaillé chez Google sur le modèle d'IA musicale AudioLM.

Pour obtenir des réponses *« en temps réel »* (en quelques centaines de millisecondes), Moshi s'appuie sur un modèle d'IA unique, entraîné directement sur des extraits sonores. Cela permet de mieux décoder et imiter les émotions ou les accents, assurent les chercheurs. Moshi pourrait adopter *« soixante-dix styles et tons »* : chuchoter, prendre une *« voix de pirate »*, parler anglais tel un Français… Autant de nuances inspirées de la voix d'une actrice enregistrée.

« Parler tout en écoutant »

Pour pallier le manque de données sonores disponibles et améliorer la fiabilité, Kyutai a adjoint au modèle sonore un modèle de traitement du texte maison (Helium), qui travaille en parallèle. Pour permettre les interruptions dans une conversation, Moshi utilise aussi deux flux, ce qui lui permet de *« parler tout en écoutant »*. Effet démo oblige, mercredi 3 juillet, l'assistant continuait d'ailleurs parfois à répondre, bien qu'on lui pose une autre question…

Ces fonctionnalités ressemblent à celles montrées en démo par OpenAI le 13 mai dans sa présentation du modèle GPT4o ou à celles promises par Apple en juin. Celles-ci ne sont toutefois pas encore disponibles pour le grand public, ce qui fait dire à Kyutai que sa démonstration accessible en ligne est *« une première »*.

Le laboratoire se félicite aussi d'avoir réduit les besoins en calcul informatique de Moshi, ce qui permet de l'utiliser hors ligne, sur un ordinateur de type MacBook Pro, et *« bientôt »* sur un smartphone. Les sons produits peuvent aussi être identifiés comme *« créés par une IA »*, grâce à un filigrane (*watermark*) inséré dans les fichiers, afin de lutter contre la désinformation ou les usurpations d'identité.

Et maintenant ? Kyutai espère que des entreprises et institutions vont déployer Moshi. Le laboratoire a encore du budget : il n'a, selon nos informations, dépensé qu'environ 10 millions d'euros, en salaires et en calcul informatique (via 1 000 processeurs Nvidia du supercalculateur de Scaleway, filiale d'Iliad). Ses futures pistes de recherche touchent à l'IA *« multimodale »* mélangeant texte, image et son, ainsi qu'à l'amélioration des modèles en langue française. Kyutai espère accéder à des contenus francophones d'institutions publiques, mais cela pose des questions de droits d'auteur.