Mozilla restructure aussi sur la reconnaissance vocale

Quel avenir pour les projets de Mozilla dans le domaine de la reconnaissance vocale ? La question s'était posée l'été dernier avec l'annonce de la restructuration des activités de la fondation. Elle a véritablement trouvé réponse hier. Aussi bien sur le moteur DeepSpeech que sur l'initiative Common Voice, destinée à constituer un jeu de données d'entraînement.

Qu'en est-il pour DeepSpeech ? Dans les grandes lignes, les équipes de Mozilla arrêteront, d'ici quelques mois, de contribuer au code. La fondation prendra alors un rôle d'accompagnateur pour le développement d'applications concrètes. Elle a œuvré dans ce sens ces dernières semaines, en réduisant les dépendances nécessaires à l'implémentation du modèle. Un guide doit par ailleurs paraître dans les prochaines semaines. S'y adjoindra un programme de subvention de projets.

Lorsque la nouvelle de la restructuration était tombée, DeepSpeech n'était plus très loin de la v1. On en est finalement resté à la 0.9.3, publiée voilà quatre mois. Elle repose sur un réseau de neurones probabiliste à cinq couches entraîné avec TensorFlow. En inférence, il peut fonctionner sur un Raspberry Pi 4.

Mozilla met ses pions sur Common Voice

Le projet avait pris son envol à la mi-2017, sur la base de travaux de recherche signés Baidu. Après avoir travaillé à partir de jeux de données libres tels que TED-LIUM et LibriSpeech, Mozilla avait enclenché la démarche Common Voice. Le principe : faire appel à la communauté – sur la base du volontariat – pour mettre davantage de matière à disposition de DeepSpeech.

En fin d'année, une première version publique du moteur avait vu le jour, assortie d'un corpus de 500 heures d'audio en anglais. À la mi-2018, le français, l'allemand et le gallois avaient rejoint la liste des langues dans lesquelles la communauté pouvait réaliser des enregistrements. Début 2019, on en comptait une vingtaine, dont le basque, le kabyle et l'espéranto.

Au dernier pointage, on a atteint les 60. L'anglais reste la plus représentée, avec environ 70 000 voix pour quelque 1800 heures d'audio validées. Suivent l'allemand (849 heures ; 13 500 voix), le français* (623 heures ; 12 900 voix) et l'espagnol (351 heures ; 20 100 voix).

Le défi de la diversité

Le corpus dans son ensemble comprend 7335 heures validées, pour 9283 enregistrées. Mozilla est donc proche de son objectif de 10 000 heures, considéré comme « la quantité de données nécessaire pour être en mesure de produire un système de reconnaissance vocale de qualité ».

La fondation se penche désormais sur un autre défi : la diversité. Beaucoup de langues disponibles sur Common Voice comptent encore moins d'une centaine de voix au répertoire. Le luganda (parlé en Ouganda) en fait partie. Même chose pour le iakoute (Sibérie) ou le bas-engadinois (Suisse).

Levier probable de mise en action de ce plan : un investissement de 1,5 million de dollars en provenance de NVIDIA.

** Pour la France, le corpus associe, entre autres, des contributions individuelles, des extraits de débats de l'Assemblée nationale, des livres du projet Gutenberg et des extraits de pièces de théâtre sous licence le permettant.*

Photo d'illustration © Visual Generation – Adobe Stock