Voix & IA / Reconnaissance Automatique de la Parole

14 mai 2020
Mohamed Bouaziz

On voit ces dernières années la place de plus en plus grande de la communication vocale dans nos appareils intelligents. Aujourd'hui on peut activer notre smartphone via une simple commande vocale, on peut dicter à sa voiture la destination souhaitée et on peut même demander à un assistant vocal de commander à manger. Et ce n'est que la partie émergée de l'iceberg car les applications sont très nombreuses. Côté business, les avancées technologiques permettent de transcrire automatiquement une réunion ou un compte rendu vocal, d'estimer le niveau de satisfaction d'un client à travers les traits de son expression, de vérifier l'identité d'un client qui appelle son conseiller bancaire… Toutes ces prouesses sont les fruits de l'exploitation, par les capacités de l'intelligence artificielle, des informations offertes par notre voix.

En effet, la parole est un moyen de communication primordial chez l'homme. Elle est tellement riche en informations que les scientifiques essayent sans cesse de l'analyser afin d'en comprendre les différents aspects. Depuis les années 1950, de nombreuses équipes de chercheurs (informaticiens, phonéticiens, mathématiciens, linguistes…) se sont penchées sur un objectif commun : automatiser les processus d'interprétation de la parole, mais aussi de sa production. La reconnaissance automatique de la parole (RAP), la reconnaissance du locuteur et la synthèse de la parole ont particulièrement intéressé les académiques ainsi que les entreprises. Les résultats de ces problématiques représentent maintenant l'interface d'échange avec beaucoup de nos appareils intelligents, notamment avec nos assistants vocaux. Dans cet article, nous allons particulièrement nous intéresser à la **reconnaissance automatique de la parole**.

La reconnaissance de la parole consiste à transcrire automatiquement un contenu parlé afin obtenir la séquence de mots correspondante. Les premiers systèmes étaient capables de transcrire uniquement des mots isolés avec un vocabulaire réduit. Durant le dernier quart du XXème siècle, Les systèmes ont commencé à pouvoir transcrire une parole continue grâce notamment à la modélisation acoustique à base de modèles de Markov cachés (HMM) et à une modélisation stochastique de la langue. L'architecture concernée, présentée dans la figure 1, est utilisée jusqu'à nos jours, surtout dans des cas d'utilisation où les données annotées ne sont pas abondantes.

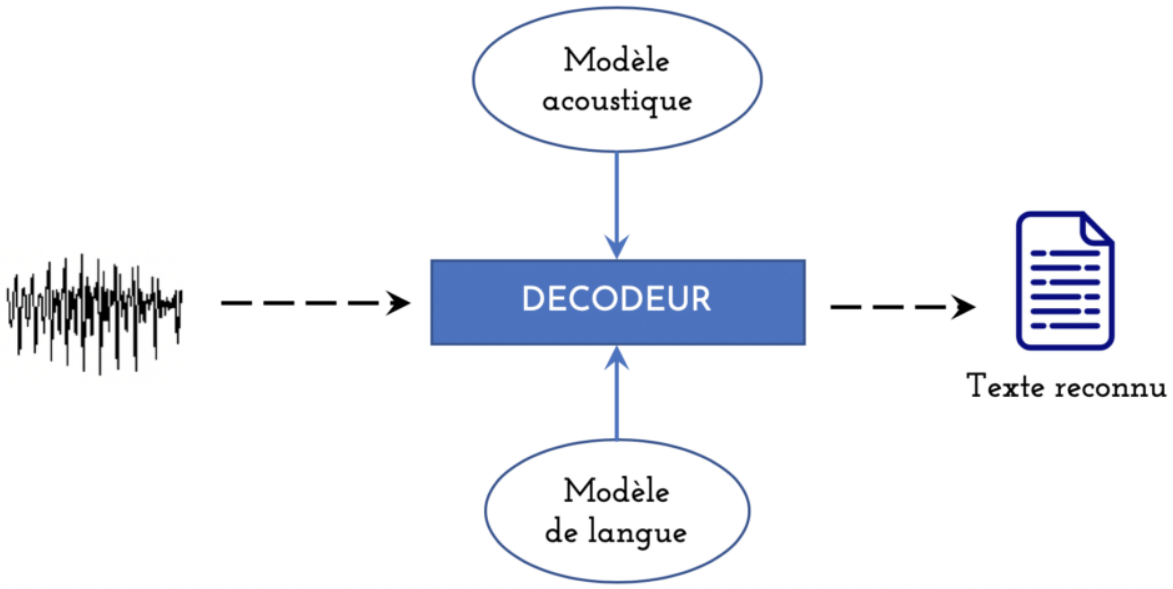

Figure 1 : Architecture classique (et hybride) d'un Système de Reconnaissance Automatique de la Parole (SRAP)

Deux modèles sont nécessaires à ce type d'approches.

- Le modèle acoustique apprend à reconnaitre les séquences de phonèmes (qui forment les mots) présentes dans un dictionnaire de prononciation. L'apprentissage est effectué sur plusieurs heures d'enregistrements audio transcrites manuellement.
- Le modèle de langue apprend (sur des données textuelles) les probabilités d'un (très grand) ensemble de séquences de mots possibles.

Un décodeur (principalement, un algorithme de recherche dans un graphe) combine les connaissances acoustiques et linguistiques afin de transcrire automatiquement l'enregistrement en entrée. On note qu'une étape d'extraction de caractéristiques est généralement effectuée a priori afin d'obtenir une représentation en spectre de ces enregistrements.

Le modèle acoustique représente le composant le plus important dans cette architecture. Il consiste en un ensemble de HMM modélisant (généralement) des phonèmes et dont les probabilités d'émission sont représentées par des mixtures de gaussiennes (voir la figure 2). Un HMM estime donc la probabilité d'observer une forme acoustique sachant un phonème donné.

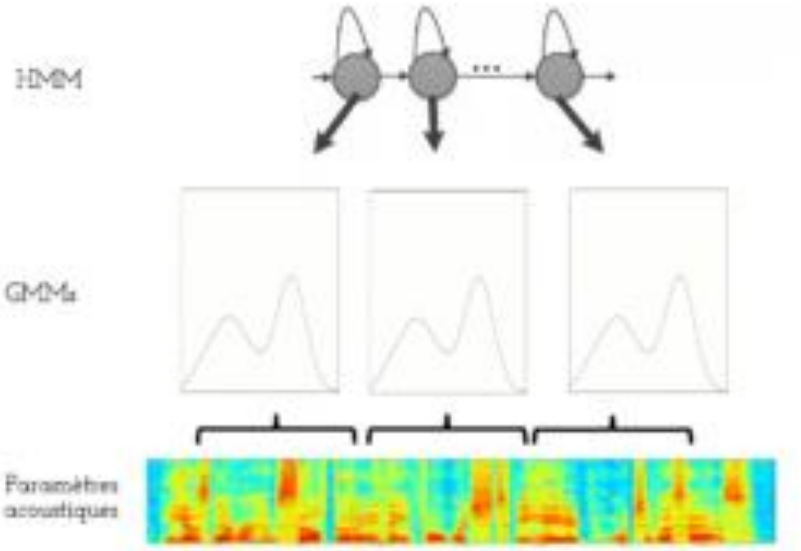

Figure 2 : Modélisation acoustique HMM-GMM

L'adoption des avancées du Deep Learning dans le domaine de la RAP était relativement progressive. Les scientifiques ont tout d'abord opté pour une approche hybride. Comme présenté dans la figure 3, cette approche consiste à remplacer les GMM par des réseaux de neurones profonds (DNN). Elle garde donc la même architecture générale de la figure 1. Contrairement aux GMM, un seul DNN est appris pour estimer les probabilités de tous les états.

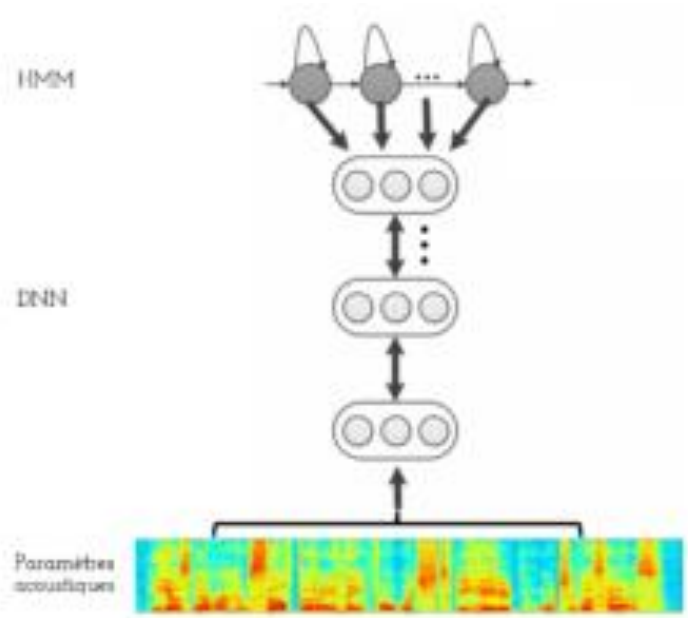

Figure 3 : Modélisation acoustique HMM-DNN

Les modèles HMM-DNN possèdent une meilleure capacité de généralisation. Leurs entrées peuvent aussi être enrichies par des informations relatives au locuteur, à la réverbération, etc. Toutefois, des paramètres acoustiques normalisées, une fenêtre de trames acoustiques relativement large et un nombre suffisant de couches sont nécessaires pour garantir son efficacité.

Kaldi est une boite à outils qui a implémenté avec excellence des techniques basées sur l'approche hybride. Elle est considérée comme un outil de référence dans plusieurs benchmarks et est fréquemment utilisée dans des solutions commerciales.

L'approche hybride donne toujours des performances à l'état de l'art sur beaucoup de corpus de données de RAP. Malgré les avancées apportées en termes de performance, cette approche conserve une architecture relativement complexe et requiert une certaine connaissance linguistique et phonétique de la langue traitée. Ces inconvénients ont longtemps motivé les chercheurs à concevoir des systèmes basés sur du Deep Learning en end-to-end.

Cet objectif n'était pas facile. Et ce n'est quasiment qu'en 2019 que des solutions end-to-end ont pu rejoindre les performances des approches hybrides. Proposée par Baidu Research, Deep Speech [1, 2] fait partie des premières approches end-to-end qui ont prouvé leur efficacité. L'architecture, schématisée dans la figure 4, combine des couches de convolution, des couches récurrentes bidirectionnelles et des couches complètement connectées et adopte une fonction de coût de type CTC (Connectionist Temporal Classification).

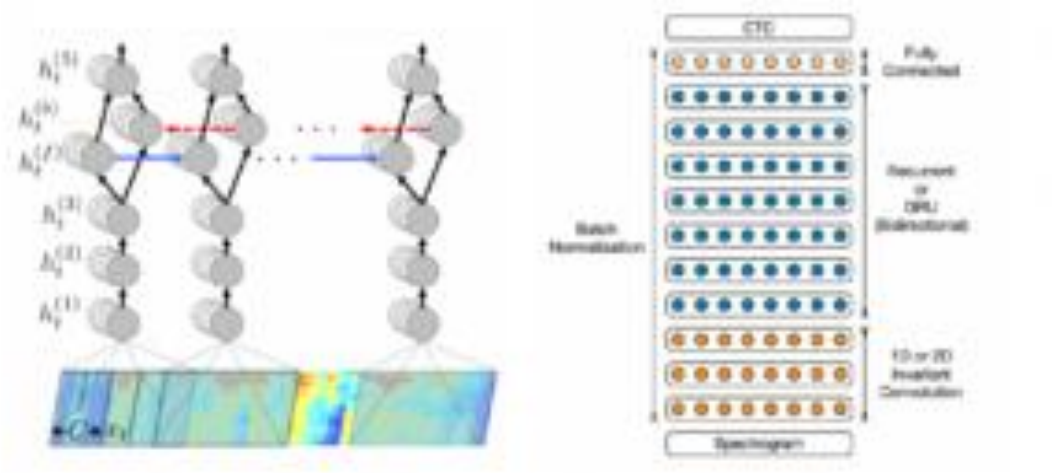

Figure 4 : Architectures Deep Speech 1 (à gauche) et Deep Speech 2 (à droite)

Plus récemment, les chercheurs de Facebook AI Research ont quant à eux opté pour une architecture sequence-to-sequence, de plus en plus utilisée. Dans un travail présenté à la conférence INTERSPEECH 2019 [3], l'équipe adopte un encodeur composé entièrement de couches de convolutions 2D et un décodeur qui consiste simplement en une couche récurrente. L'approche, implémentée sur l'outil open-source « wav2letter » (conçu par la même équipe), atteint des performances très proches de l'état de l'art sur le corpus LibriSpeech.

D'autres avancées ont également été révélées dans la même conférence. Des scientifiques du laboratoire de recherche japonais NTT ont appliqué une architecture sequence-to-sequence récente, appelée « Transformer », qui s'est montré encore plus efficace dans plusieurs tâches de transformation de séquences. En entraînant cette architecture par un apprentissage multitâche basé à la fois sur la CTC et l'attention, le système réalise de meilleurs résultats que ceux de plusieurs autres approches état-de-l'art sur deux datasets standards [4]. Cette solution est disponible en open-source à l'aide de l'outil ESPNet.
Les approches end-to-end évoluent de plus en plus et pourront devenir la nouvelle référence dans la RAP à l'instar de nombreuses autres tâches. De nouvelles pistes prometteuses ont également vu le jour ces quelques dernières années comme le self-supervised learning. Quelques travaux ont commencé, à générer des représentations latentes plus discriminantes ou à apprendre jointement la reconnaissance et la production de la parole, etc. Ces approches visent à réduire le besoin en données annotées et à étendre les technologies de RAP à beaucoup de langues faiblement dotées.

Source : https://www.aquiladata.fr/insights/voix-ia-reconnaissance-automatique-de-la-parole/