Reconnaissance automatique de la parole : tout commence par la voix

Le 06/02/2019
Pierre Ponlevé

La reconnaissance automatique de la parole (RAP) est une notion vaste qui comprend la commande et la dictée vocale. Cette technologie permet d'analyser la voix humaine dans une logique d'amélioration continue grâce à des technologies comme celle du deep learning.

Il y a une vingtaine d'années, les logiciels permettant une reconnaissance de la voix se trompait sur quasiment un mot sur deux (43 %). Aujourd'hui, les meilleures solutions avoisinent les 90 % de fiabilité dans la retranscription.

La reconnaissance automatique de la parole (RAP), souvent improprement appelée reconnaissance vocale, est une technique qui permet d'analyser la voix humaine captée au moyen d'un microphone pour la transcrire sous la forme d'un texte ou d'un fichier exploitable sur ordinateur. Un système de RAP a pour but d'associer une séquence de mots à une séquence d'observations acoustiques, autrement dit un discours prononcé par un locuteur.

Dictée et commande vocale

La dictée vocale, commercialisée dès les années 80, est sans aucun doute l'application la plus populaire de la reconnaissance automatique de la parole. Elle permet de dicter oralement un texte qui sera retranscrit automatiquement par un processeur. Les premières solutions étaient lentes, maladroites et assez onéreuses. Aujourd'hui, il existe des solutions complètes qui proposent en plus de la retranscription de votre discours, d'autres fonctions comme la traduction ou la création de sous-titres pour une vidéo. Ces logiciels offrent des marges d'erreurs très raisonnables lors de la retranscription manuscrite - aujourd'hui ils atteignent 6% contre 3% environ pour un humain.

La dictée vocale est simple d'utilisation. La commande vocale, elle, nécessite une certaine prise en main pour que son fonctionnement soit optimal.

La commande vocale est, sans doute, le secteur le plus développé de la RAP. Le premier dispositif de commande vocale a été créé en 1971, c'est le Voice Command System. Il s'agit d'une calculatrice capable de reconnaître 24 ordres. On peut définir la commande vocale comme une interface d'entrée d'un système informatique permettant de passer des ordres à l'aide de messages vocaux.
Composants d'un système de reconnaissance automatique de la parole

Un système de reconnaissance automatique de la parole comporte typiquement 4 modules :

- le prétraitement acoustique qui va identifier les zones de parole dans l'enregistrement à transcrire et en extraire des séquences de paramètres acoustiques ;
- le modèle de prononciation qui associe les mots connus par le système à leurs représentations phonétiques ;
- le modèle acoustico-linguistique servant à prédire les phonèmes les plus probablement prononcés dans un énoncé audio, ainsi que la séquence de mots la plus probable dans cet énoncé ;
- le décodeur qui va combiner les prédictions des modèles acoustiques et linguistiques pour proposer la transcription en texte la plus probable pour un énoncé de parole donné.

Fonctionnement de la RAP

Le principe est assez simple. Une personne en parlant émet des variations de pression dans son larynx, les sons produits sont numérisés par le micro du logiciel afin d'être transmis sur le réseau. Ils sont ensuite transformés en vecteurs acoustiques. Le moteur de reconnaissance va alors analyser cette suite de vecteurs acoustiques en la comparant avec ceux qu'il a en mémoire (son modèle de langage) et proposer la suite qui lui paraît la plus probable. Il est donc nécessaire que la suite de vecteurs acoustiques se rapproche d'une de celles qui est mémorisée par le moteur de reconnaissance. Afin de créer cette base, il est primordial de développer ce que l'on appelle une grammaire.

Grammaire de règles et grammaire statistique

On distingue deux types de grammaires : les grammaires de règles et les grammaires statistiques.

- Les grammaires de règles sont des descriptions des phrases possibles, transformées par le moteur en une représentation acoustique qui sera comparée avec ce qui est prononcé. Si la phrase prononcée par l'utilisateur se rapproche acoustiquement d'une de celles présentes dans la grammaire, alors le moteur remonte cette phrase, sinon il indique qu'il n'a pas compris.
- Les grammaires statistiques découlent du calcul, à partir d'un corpus de phrases, des probabilités de suites de mots. Après compilation phonétique et acoustique, la grammaire statistique permet au processeur de proposer la suite de mots la plus probable en comparaison avec la parole prononcée.

En pratique, les grammaires statistiques sont plus aptes à répondre à des requêtes du type « que voulez-vous ? ». En revanche pour des demandes de formulaire (comme la saisie d'un numéro de carte bancaire), on utilisera des grammaires de règles afin d'éviter les erreurs de saisie. Les grammaires statistiques seront plus orientées vers des métiers où une interactivité directe avec le client est nécessaire (assurances, banques, etc.).

Pour les grammaires de règles, elles seront surtout mises en place dans les entreprises qui ont un besoin de renseignements. Si vous devez remplir un formulaire d'authentification sur un site internet, vous aurez affaire à une grammaire de règles. Des solutions combinant à la fois grammaire de règle et grammaire statistiques existent comme la célèbre application de commande vocale développée par Apple, Siri.

Monolocuteur vs multilocuteur

Il y a deux techniques de reconnaissance automatique de la parole couramment utilisées : monolocuteur et multilocuteur. La première sera plus utile en entreprise quand la seconde sera plus orientée grand public. Pour les entreprises, il semble plus avantageux d'utiliser la reconnaissance monolocuteur qui pourra s'adapter individuellement à la personne qui utilise le logiciel grâce à l'apprentissage profond. Prenons pour exemple une profession où le vocabulaire est parfois assez pointu : les bibliothécaires. Ils peuvent utiliser la reconnaissance monolocuteur et la faire évoluer grâce aux dictionnaires intégrés dans ces solutions. Concrètement, cela s'illustre par la possibilité d'ajouter au dictionnaire du logiciel un vocabulaire spécifique au métier de bibliothécaire ("catalogage", "classification", "périodique"…). Ces termes ne seront rajoutés qu'une seule fois. Ils seront ensuite reconnus par le logiciel, puis le bibliothécaire pourra les réutiliser sans avoir à se soucier de la transcription. Cela apporte un gain de temps non négligeable.

- La reconnaissance monolocuteur est une solution généralement stockée sur des serveurs locaux et qui nécessite d'enregistrer au préalable la voix de l'utilisateur pour que le logiciel s'en serve comme référence. On peut citer par exemple le logiciel Dragon NaturallySpeaking de Nuance qui utilise ce type de reconnaissance. Ces produits sont plutôt destinés à des usages précis B to B.
- La reconnaissance multilocuteur fonctionne avec n'importe quelle voix, mais nécessite une connexion internet pour comparer la requête avec une base de données stockée dans le cloud. Elle est utilisée dans des produits grand public comme les enceintes connectées des Gafam à l'instar d'Amazon et son Echo.

Cette reconnaissance est généralement intégrée dans les produits à destination des néophytes car elle ne fonctionne qu'avec des commandes "simples" comme le contrôle de la domotique d'une maison ou la consultation d'un agenda virtuel.

La reconnaissance multilocuteur est donc privilégiée pour le grand public et destinée à envahir le quotidien des gens.

Usages de la reconnaissance automatique de la parole

Les avantages que l'on attend de la reconnaissance automatique de la parole sont multiples. Elle libère complètement l'usage de la vue et des mains, contrairement à l'écran et au clavier, et laisse l'utilisateur libre de ses mouvements.

Par ailleurs, la vitesse de transmission des informations est naturellement plus rapide à la voix qu'avec l'écriture manuscrite. Enfin, tout le monde ou presque peut parler, alors que peu de gens sont à l'abri des fautes de frappe et d'orthographe.

Ces avantages sont à l'origine d'une grande variété d'applications comme :

- l'aide aux handicapés ;
- la messagerie ;
- l'avionique ;
- la commande de machines ou de robots ;
- le contrôle de qualité et la saisie des données ;
- l'accès à distance : téléphone, internet ;
- la dictée vocale.

La RAP permet de contrôler-commander un outil et de dicter des mots. Elle peut être utilisée dans le domaine industriel, mais également dans les secteurs de l'automobile (par exemple, entrer une destination dans le GPS) et de la domotique (par exemple, programmer la température de la maison). Elle constitue également une aide pour la communication, que ce soit pour les personnes handicapées (notamment la surdité partielle ou totale) ou les apprenants d'une langue.

Source : https://www.archimag.com/vie-numerique/2019/02/06/reconnaissance-automatique-parole-commence-par-voix