

La reconnaissance automatique de la parole (RAP)

Lingua francese per le relazioni internazionali e le pubbliche amministrazioni –
(RISE-SA), a.a. 2025-2026

Prof.ssa Alida Maria Silletti

SOURCES: <https://www.techniques-ingenieur.fr/base-documentaire/technologies-de-l-information-th9/documents-numeriques-technologies-d-acquisition-et-de-restitution-42310210/reconnaissance-automatique-de-la-parole-h3728/caracteristiques-de-la-communication-parlee-homme-machine-h3728v3niv10001.html> (Haton 2018) ; <https://core.ac.uk/download/pdf/15486009.pdf> (Mariani 1990) ; <https://journals.openedition.org/rac/27069> (Kneubühler 2022) ; <https://www.lebigdata.fr/intelligence-artificielle> ; <https://www.techno-science.net/glossaire-definition/Reconnaissance-vocale.html> ; <https://www.journaldunet.fr/web-tech/guide-de-l-intelligence-artificielle/1501849-reconnaissance-vocale/> ; <https://shs.hal.science/halshs-02917916/document> (Tancoigne *et al.* 2020) ; https://theses.hal.science/tel-00424699v1/file/These_LoicBARRAULT.pdf (Barrault 2008)

Aux origines de la reconnaissance automatique de la parole (RAP) (1)

- C'est une discipline quasi contemporaine de l'informatique, datant des années 1950, lorsqu'apparaît le premier système de reconnaissance de chiffres en électronique analogique, quoique imparfait
- Années 1960 : des méthodes numériques et des ordinateurs sont introduits grâce à l'évolution technologique et les systèmes de RAP sont implantés sur ordinateur : une facilitation des recherches, mais des résultats modestes à cause de diverses difficultés
- Début années 1970 : on comprend qu'il faut faire appel à des contraintes linguistiques dans le décodage automatique de phrases – la reconnaissance de la parole avait été auparavant considérée comme un problème d'ingénierie

Aux origines de la reconnaissance automatique de la parole (RAP) (2)

- Les progrès de la microélectronique permettent la miniaturisation et l'implantation de systèmes complexes de RAP sous forme logicielle ou sur une puce et leur utilisation dans des secteurs d'activité très variés, en lien avec le développement de la télématique vocale
- Pourtant, pour 1 heure d'enregistrement, la durée de transcription peut aller de 4 à 6 h, voire 30 h en fonction de la personne qui transcrit, des caractéristiques de l'enregistrement et de la transcription
- Depuis 1990 : la comparaison entre des méthodes basées sur les connaissances issues de l'intelligence humaine et des méthodes utilisant des bases de données de parole et des algorithmes d'apprentissage automatique tourne à l'avantage de ces dernières méthodes, qui obtiennent des résultats meilleurs lors d'essais comparatifs d'évaluation
- Depuis 2010 : des modules d'intelligence artificielle sont intégrés aux algorithmes de reconnaissance automatique de la parole – les algorithmes progressent rapidement

La reconnaissance de mots (1)

- L'analyse du signal vocal ne permet de déceler aucune séparation entre mots successifs : la parole est un « semi-continuum » caractérisé par des pauses qui correspondent à des types de sons ou de respiration et les mots n'ont pas de frontières
- Pour simplifier cette identification, depuis les années 1970, on a recours à la reconnaissance de mots prononcés artificiellement de façon isolée par seulement une personne, utilisant un petit vocabulaire (entre 20 et 50 mots), avec de courtes pauses entre les mots – la reconnaissance d'un mot relève de la reconnaissance de formes
- Un système de reconnaissance de formes comporte trois parties :
 - un capteur (c'est-à-dire un microphone) permet d'appréhender les fréquences sonores de la voix ;
 - une paramétrisation des formes par un analyseur spectral, qui traduit les fréquences sonores de la voix en texte exploitable par la machine, et par les technologies de l'intelligence artificielle ;
 - un niveau de décision pour classer une forme inconnue dans l'une des catégories possibles

La reconnaissance de mots (2)

- Ce système comporte également des phases :
 - par une phase préalable d'entraînement-apprentissage, un sujet locuteur prononce l'ensemble des mots du vocabulaire, souvent plusieurs fois, pour créer en machine le dictionnaire de référence – le signal correspondant est traité au niveau « acoustique » et l'information résultante est conservée en mémoire, avec son étiquette correspondante ;
 - par la phase suivante de reconnaissance, un sujet locuteur prononce un mot du vocabulaire qui est comparé aux mots de référence : la forme correspondante est comparée avec toutes les formes de référence conservées en mémoire – si le mot prononcé n'est pas reconnu, il est rejeté du vocabulaire ;
 - l'algorithme de reconnaissance permet de choisir le mot le plus ressemblant, par calcul d'un taux de similitude entre le mot prononcé et les diverses références – lorsqu'un sujet locuteur prononce deux fois un même mot, les spectrogrammes correspondants ne seront jamais exactement les mêmes

Les difficultés du traitement automatique de la parole continue

- Absence de séparateurs, de silences entre les mots, contrairement aux blancs dans le langage écrit
- Chaque son élémentaire (phonème) est modifié par son contexte proche, à savoir le phonème qui le précède et celui qui lui succède (coarticulation) – lorsqu'un phonème est prononcé, la prononciation du phonème suivant est préparée par un mouvement du conduit vocal – et par son contexte plus large, à savoir la place du phonème dans la phrase
- La parole présente une très grande variabilité
 - intra-locuteur, concernant le mode d'élocution (voix chantée, criée, murmurée, enrhumée, enrouée, sous stress, bégaiement ...)
 - interlocuteur (timbres différents, voix masculines, féminines, voix d'enfants ...)
 - due au moyen d'acquisition du signal (type de microphone), ou à l'environnement (bruit)
- Il faut étudier et traiter une grande quantité de données pour obtenir un son élémentaire, en dépit des différents contextes, des différents modes d'élocution, des différents sujets locuteurs, des différents environnements
- Le même signal renferme différents types d'informations (les sons eux-mêmes, la structure syntaxique de la phrase, sa signification, mais aussi l'identité du sujet locuteur et son état émotionnel (joie, colère...))

De la reconnaissance de mots isolés à la parole continue

- Travailler sur les mots isolés limite la communication entre un être humain et une machine
- Donc, seule la parole naturelle et continue assure le niveau d'expression nécessaire pour des applications complexes
- Il faut prendre en compte plusieurs informations :
 - acoustico-phonétiques : elles régissent la transcription phonétique du message avec les informations phonologiques, qui rendent compte des variations individuelles (ex. accent) et des phénomènes phonétiques d'altérations des sons (ex. liaisons) ;
 - lexicales : elles sont liées aux mots, même si les mots n'apparaissent pas explicitement dans le signal acoustique ;
 - prosodiques : la prosodie concerne le rythme, l'intensité et la mélodie de la voix – une sorte de « ponctuation » de la parole ;
 - syntaxiques : la syntaxe est liée à la structure des phrases ;
 - sémantiques : elles relèvent de la signification des mots et donc de la compréhension du sens de la phrase prononcée ;
 - pragmatiques : elles portent sur le contexte de l'univers et de la conversation

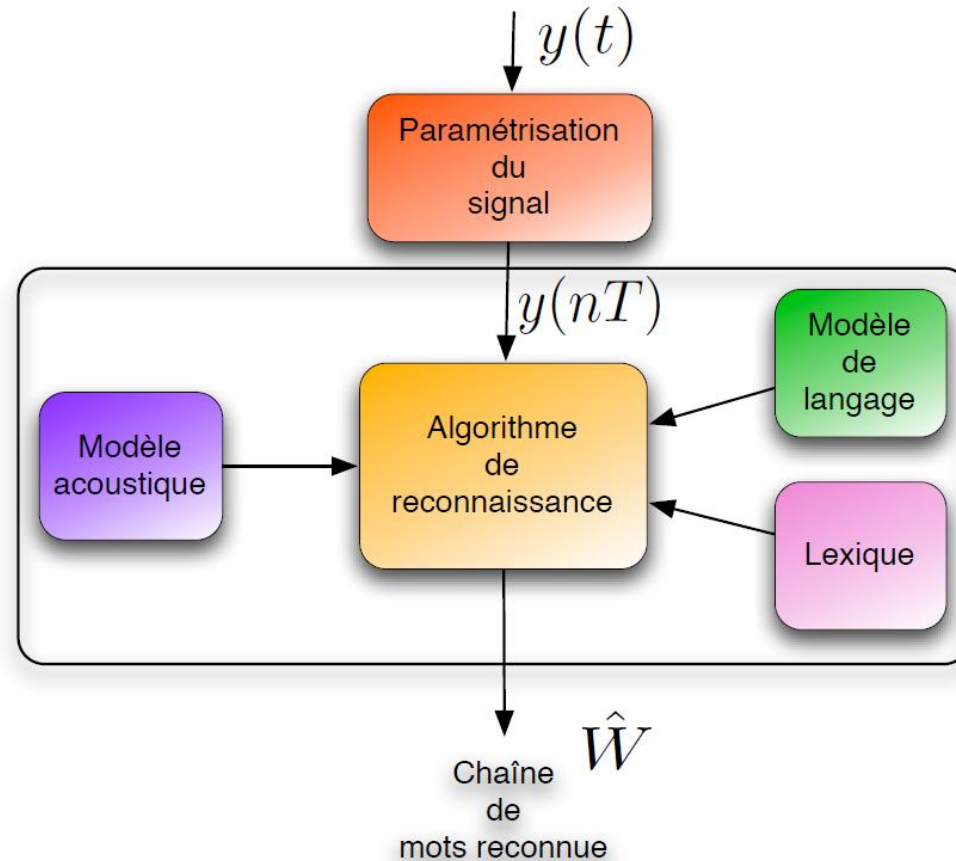
Quelles méthodes pour la reconnaissance de la parole continue ?

- Les systèmes de reconnaissance actuellement disponibles sont performants dans des conditions d'utilisation contrôlées, selon la complexité et la difficulté de la tâche envisagée
- En moyenne, les taux d'erreur mesurés en laboratoire peuvent aller de 0,3 % (pour des suites de chiffres) à 5 % (pour un vocabulaire de 20 000 mots en parole continue), à 8 % (pour des lettres épelées), jusqu'à 55 % pour des conversations téléphoniques spontanées...
- Le taux d'erreur s'accroît lorsque les conditions d'apprentissage et d'utilisation d'un système sont différentes (selon le type et le niveau de bruit)
- Il existe trois catégories de sources de variabilité de la parole, selon leur provenance :
 - l'environnement du sujet locuteur : il s'agit du bruit corrélé à la parole ou additif, donc extérieur (ex. bruit ambiant) ;
 - le sujet locuteur lui-même, selon son état et son mode d'expression : essoufflement, stress, modification de sa propre voix en cas d'ambiance très bruitée, rythme d'élocution, fatigue ;
 - les conditions d'enregistrement liées au type de microphone, distance au microphone, canal de transmission (distorsion, écho, bruit électronique)

Les réseaux neuronaux profonds

- Depuis 2006, les modèles acoustiques de reconnaissance sont améliorés grâce aux modèles neuronaux profonds (*Deep Neural Networks*)
- Ces réseaux sont inspirés du fonctionnement du cortex animal et sont capables d'apprendre des fonctions beaucoup plus complexes qu'auparavant
- Leur développement est dû à la convergence de trois conditions :
 - l'existence de très grandes bases de données acoustiques étiquetées nécessaires à l'apprentissage de ces modèles (*Big Data*), avec des millions d'heures de parole, d'où des systèmes de reconnaissance disponibles dans de nombreuses langues (plus de cent pour Google) ;
 - des capacités de calcul en augmentation ;
 - l'amélioration des algorithmes d'apprentissage de ces modèles (*Deep Learning*)

Le fonctionnement d'un système de RAP

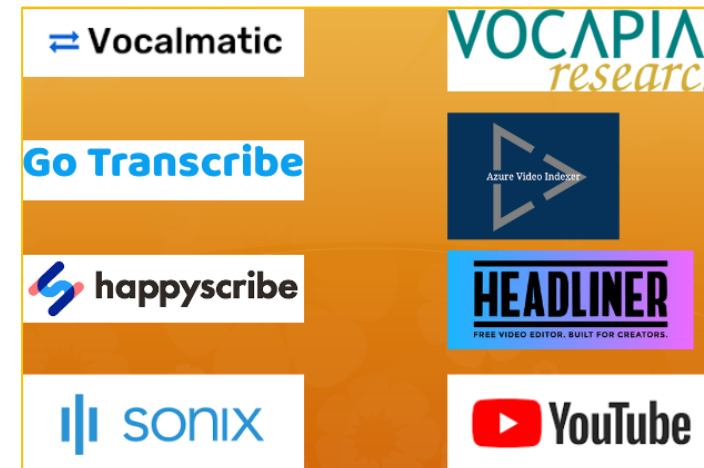


L'avenir de la RAP

- Les travaux actuels les plus avancés de RAP concernent des systèmes de dialogue *via* le téléphone, la reconnaissance de la parole spontanée et la transcription d'émissions de radio ou télévision
- Les performances obtenues dépendent du type de tâche considérée : taille et difficulté du vocabulaire, sujets locuteurs, conditions d'enregistrement
- Malgré ces avancées, les systèmes actuels restent imparfaits : un important effort de recherche est nécessaire sur le plan de la robustesse des méthodes de reconnaissance et de la conception de systèmes de dialogue

Les logiciels speech-to-text

- Ils appartiennent au domaine du traitement automatique du langage (TAL), à l'interface entre la linguistique et l'informatique
- Ils assurent le passage de la parole au texte dans le traitement automatique de sources orales
- Septembre 2020 : un rapport de recherche est publié par des spécialistes sur la transcription automatique en langue française (Tancoigne *et al.* 2020)
- Ces spécialistes comparent 8 logiciels *speech-to-text* de transcription automatique : Go Transcribe, Happy Scribe, Headliner, Sonix, Video Indexer, Vocalmatic, Vocapia, YouTube à partir de quatre extraits de fichiers audio en langue française :
 - un texte lu ;
 - un cours magistral enregistré en situation ;
 - un entretien avec deux personnes ;
 - une réunion associative avec de nombreuses personnes



L'étude de Tancoigne et al. (2020)

- Objectifs :

- évaluer les fonctionnalités des plateformes en termes de sécurité et confidentialité des données, tarification, interopérabilité, simplicité d'utilisation, outil d'édition ;
- évaluer les transcriptions obtenues pour un classement et une compréhension des erreurs générées par les logiciels et pour effectuer des estimations du potentiel de gain de temps de transcription par fichier et par logiciel

- Résultats :

- la qualité de la transcription dépend du type de fichier soumis en entrée (des discours planifiés VS la parole spontanée) ;
- tous les logiciels échouent à retranscrire le fichier le plus complexe, la réunion associative, à cause de la présence de nombreux chevauchements, d'extraits inaudibles et de bruits de fond ;
- un temps de réécoute et de correction reste indispensable ;
- le gain de temps final observé peut aller jusqu'à 75 % par rapport à une transcription manuelle

L'étude de Tancoigne et al. (2020) : YouTube

- Le seul logiciel 100 % gratuit
- Ses données sont hébergées par Google
- Il présente des balises temporelles
- La ponctuation y est absente
- L'alternance des personnes qui parlent n'est pas respectée
- 80-85 % des mots transcrits sont partagés avec le texte de référence