Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Mise en place d'un système robuste de reconnaissance automatique de la parole appliqué au domaine médical

Master

TRAITEMENT AUTOMATIQUE DES LANGUES

Spécialité

RECHERCHE ET DEVELOPPEMENT

par

Lucía ORMAECHEA GRIJALBA

Tuteur universitaire

Cyril GROUIN

Année académique

2019/2020

Sous la direction de

Benjamin LECOUTEUX

Didier SCHWAB

Pierrette BOUILLON

Abstract

The proper functioning of Automatic Speech Recognition (ASR) systems is a complex challenge in the context of Speech-to-Speech (STS) translation applied to the medical field. The current master's thesis presents research work aiming to build a robust speech recognition system as part of the BabelDr project, a STS translation tool that has been implemented in the Geneva University Hospitals (HUG) to facilitate doctor-patient interaction when no common language is shared. Currently, its speech recognition technology is based on a black box system provided by a private company. The main goal of this study is to break the dependency of an external device on the basis of open source tools that can evolve according to the needs of HUG. For this reason, we propose a French ASR system based on the Kaldi toolkit which performs automatic transcription in real time, using hybrid HMM-DNN acoustic models and linguistic modeling adapted to the medical discourse specific to emergency contexts. In light of the overall results observed, a significant improvement is noted compared to the black box approach previously employed.

Key words: *automatic speech recognition – acoustic modeling – language modeling – Kaldi – BabelDr – speech-to-speech translation*

Résumé

Le bon fonctionnement des systèmes de reconnaissance automatique de la parole s'avère un défi complexe dans le contexte de la traduction *speech-to-speech* utilisée dans le domaine médical. Ce mémoire présente un travail de recherche qui vise à construire un système robuste de reconnaissance vocale dans le cadre du projet BabelDr, un outil de traduction vocale quasi instantanée qui a été mis en place dans les Hôpitaux Universitaires de Genève (HUG) afin de favoriser l'interaction médecin-patient lorsqu'aucune langue n'est partagée. Actuellement, sa technologie de reconnaissance de la parole est issue d'un système boîte noire fourni par une société privée. Le but principal de cette étude est de rompre la dépendance à un dispositif externe en se basant sur des outils libres et qui pourront évoluer selon les besoins des HUG. Pour cela, nous proposons un système de reconnaissance vocale pour le français appuyé sur la boîte à outils Kaldi. Celle-ci permet d'effectuer une transcription automatique en temps réel, utilisant des modèles acoustiques hybrides HMM-DNN et une modélisation linguistique adaptée au discours médical caractéristique du contexte d'urgences. À la lumière des résultats globaux observés, une importante amélioration est constatée par rapport à l'approche boîte noire précédemment utilisée.

Mots clés : *reconnaissance automatique de la parole – modélisation acoustique – modélisation linguistique – Kaldi – BabelDr – traduction vocale quasi instantanée*

Remerciements

Je tiens tout d'abord à remercier mes tuteurs de stage, Benjamin Lecouteux et Didier Schwab, de m'avoir confié cette mission et de m'avoir donné l'opportunité de faire partie d'un projet tellement passionnant. Un grand merci pour leur temps, leur soutien, leur disponibilité constante et leur avis toujours précieux.

Je souhaiterais remercier Pierrette Bouillon et Johanna Gerlach, d'avoir placé leur confiance en moi et de rendre possible ma participation à BabelDr. Merci également d'avoir relu attentivement ce document et pour leurs sages conseils qui m'ont permis de le peaufiner.

Je remercie également mon encadrant, Cyril Grouin, qui m'a accompagnée à distance tout au long de la réalisation de mon stage et l'élaboration de mon mémoire. Merci bien pour son soutien, ses conseils avisés et sa relecture attentive.

Je tiens à remercier tous les membres du GETALP, pour leur convivialité et leur accueil cordial, et plus particulièrement aux membres de l'Ivy LIG, pour leur compagnie si agréable et tous les bons moments passés ensemble.

Un grand merci à mes collègues de promotion et à mes professeurs de l'INALCO, Sorbonne Nouvelle et Paris Nanterre, qui m'ont dotée d'une formation solide et m'ont permis d'élargir mon bagage de connaissances. Plus particulièrement, merci à Jean-Michel Daube et Serge Fleury, pour leur patience et disponibilité, à Damien Nouvel, pour son appui, et à Marie-Anne Moreaux, de m'avoir initiée à l'algorithmique.

J'aimerais exprimer ma gratitude à Rosa Fernández Urtasun et Pablo Ruiz Fabo, de m'avoir fait découvrir le domaine fascinant du traitement automatique des langues et de m'avoir renseignée avec gentillesse et patience.

Bien entendu, je n'oublie pas de remercier mes collègues musicaux Van Morrison, Carmen McRae, Miles Davis, Gary Stewart et Chris Rea (pour ne citer qu'eux), de m'avoir accompagnée sans relâche tout au long de ce périple.

Enfin, un énorme merci à mon père, *il mio consigliere*, et à ma mère, peu importe où elle est.

Table de matières

Introduction

Dans les services de santé actuels, et plus particulièrement dans les contextes d'urgence, les barrières linguistiques constituent un problème d'une importance indéniable, car la difficulté ou l'impossibilité d'une bonne interaction entre un patient et un médecin qui ne partagent aucune langue commune est susceptible d'entraîner des répercussions fatales. C'est précisément pour cette raison que des mécanismes fiables de traduction médicale spécialisée sont importants, afin de favoriser une communication adéquate et efficace entre les acteurs mentionnés ci-dessus.

Certes, des plateformes de traduction automatique telles que Google Translate ont été mises en place, et des outils de traduction spécialisés préétablis comme MediBabble ont également été construits. Toutefois, ils ne s'avèrent pas satisfaisants dans les échanges médicaux pour des raisons liées à l'absence de précision ou de flexibilité communicative, respectivement (Turner *et al*., 2019). C'est dans ce contexte que le projet BabelDr a émergé aux Hôpitaux Universitaires de Genève (HUG). Avec cet outil de traduction *speech-to-speech* spécialisé et basé sur des règles, le médecin peut facilement communiquer avec le patient en s'adressant oralement au système, qui renverra une traduction à ce que l'utilisateur a émis dans la langue cible souhaitée.

Un tel instrument de traduction nécessite que la phrase prononcée par le locuteur soit reconnue par un système de reconnaissance automatique de la parole. Actuellement, la technologie utilisée à cet effet par BabelDr est issue d'un système boîte noire fourni par une société privée. L'objectif de ce mémoire et de rompre la dépendance à un système de reconnaissance externe, en se basant sur des outils libres et qui pourront évoluer selon les besoins des HUG. À cet égard, nous avons conçu un système de reconnaissance automatique de la parole pour le français à l'aide de « Kaldi online » qui transcrit à la volée. Pour ce faire, nous avons utilisé des modèles acoustiques hybrides HMM-DNN et nous avons effectué une modélisation linguistique adaptée à une langue de spécialité telle que le discours médical dans le cadre d'urgences.

Cette étude a été effectuée au sein de deux laboratoires de recherche, qui travaillent en étroite collaboration avec l'unité d'urgences ambulatoires des HUG :

- Le département de Traitement de l'Information Multilingue (TIM), appartenant à la Faculté de Traduction et d'Interprétation (FTI) de l'Université de Genève (UNIGE). Sa recherche se focalise sur des domaines tels que la traduction automatique, la reconnaissance vocale multilingue et la lexicologie.
- Le Groupe d'Études en Traitement Automatique de la Langue et la Parole (GETALP), qui est rattaché au Laboratoire d'Informatique de Grenoble (LIG) et affilié à son tour à l'Université de Grenoble-Alpes (UGA). Ses axes de recherche incluent le traitement automatique de la parole, le développement de ressources lexicales multilingues ou la communication médiée par ordinateur.

Nous présenterons dans un tout premier temps un tour d'horizon sur les services de traduction disponibles à présent dans le domaine médical, ainsi que le rôle de BabelDr à ce sujet. Dans un deuxième temps, nous exposerons le fonctionnement et l'architecture typique d'un système de reconnaissance automatique de la parole et examinerons les difficultés auxquelles ils sont confrontés. La troisième partie sera dédiée aux méthodes utilisées pour construire notre système de reconnaissance de la parole appuyé sur la boîte à outils Kaldi, où nous détaillerons les corpus employés et la façon dont nous avons procédée pour effectuer la modélisation linguistique adaptée aux échanges médicaux. Par la suite, nous nous arrêterons sur les expériences que nous avons menées et les résultats issus de l'implémentation de nos systèmes. Enfin, nous nous attarderons sur la conclusion de notre étude, où nous proposerons des perspectives de travaux futurs.

1. Le projet BabelDr et son rapport avec la traduction dans le domaine médical

1.1. Les difficultés de la traduction et interprétation médicale d'aujourd'hui

De nos jours, les services de santé se trouvent plus que jamais en mesure de devoir assister des personnes d'origines géographiques distinctes, compte tenu de la circulation continue des gens, des processus de mondialisation et d'interconnexion de plus en plus importants ainsi que les phénomènes migratoires permanents. De telles tendances peuvent exceptionnellement favoriser l'émergence de crises sanitaires à dimension supranationale, touchant tous les groupes de population, comme l'a montré la pandémie de Covid-19 en 2020.

Il existe, en effet, diverses raisons qui expliquent ces déplacements (qui peuvent être, à leur tour, volontaires ou forcés), qu'elles soient d'ordre politique, économique, climatique ou encore liées au fait de fuir des situations instables sur les plans sociaux (comme des guerres). Ce qui les unit pourtant, c'est le fait qu'ils répondent normalement à une volonté de quitter une situation d'insatisfaction suivie d'une ultérieure recherche d'un accroissement du bien-être.

Nous constatons des exemples remarquables de ces mouvements internationaux dans le courant de l'histoire contemporaine, tels que les migrations nord-africaines vers l'Europe Occidentale ou l'établissement de la population d'origine sud-américaine aux États-Unis. Cependant, l'exemple le plus important de ces derniers temps est à coup sûr la récente crise des réfugiés, pour laquelle des centaines de milliers des personnes se déplacent et demandent l'asile dans l'espace européen. En 2019, 676 300 personnes ont sollicité une protection internationale dans les 27 pays membres de l'Union européenne, soit une hausse de 11,2% par rapport à 2018. Il faut noter qu'il s'agit de la première fois que ce nombre augmente depuis 2015 et que la Syrie, l'Afghanistan ou le Vénézuela rassemblaient le plus grand nombre de demandes (*Eurostat Statistics Explained*, 2019). Ces migrants doivent non seulement faire face à des frontières politiques et culturelles à leur passage, mais ils ont aussi à affronter des frontières linguistiques, qui supposent

notamment une difficulté de communication entre les nouveaux arrivants et la population autochtone.

Dans le domaine médical, cette barrière linguistique s'avère un problème qui peut avoir des conséquences importantes, d'autant plus que la difficulté ou l'impossibilité d'une bonne interaction entre un patient et un médecin qui ne partagent aucune langue commune est susceptible d'entraîner des répercussions fatales. Même si ces effets peuvent être observés chez le patient, l'émission d'un diagnostic médical incorrect peut également affecter le reste de la population locale dans le cas de maladies infectieuses (Hacker *et al*., 2015). C'est précisément pour cette raison que des mécanismes fiables de traduction et d'interprétation médicale sont importants, afin de favoriser une communication adéquate entre le spécialiste et le patient.

Il faut noter à cet égard que la traduction médicale traite une langue de spécialité, à savoir, « un sous-système linguistique tel qu'il rassemble les spécificités linguistiques d'un domaine particulier » (Dubois *et al*., 1994). D'un point de vue lexical, le discours médical s'accompagne d'un niveau technique élevé, qui fait appel à une terminologie gréco-latine (Soubrier, 2011), que ce soit sous la forme de bases lexicales (comme φάρμακον dans *pharmacologie*), préfixes (comme *adipo-* dans *adipose*) ou suffixes (comme -δερμία dans *pachydermie*). Il est également observable la présence des néologismes lexicaux (compte tenu des progrès vertigineux de la médecine) ainsi que de la néosémie, souvent corrélée à l'influx de l'anglais, et l'abréviation, issue des procédés de siglaison (comme ECG pour *électrocardiogramme*) ou acronymie (comme *sida*). D'un point de vue syntaxique et pragmatique, il faut souligner la phraséologie particulière que le spécialiste emploie lorsqu'il réfère aux symptômes ou détermine un diagnostic (Rouleau, 2007), d'autant plus que l'efficacité d'un traitement dépend également de la compréhension du patient dudit traitement.

De plus, la traduction dans le domaine médical possède un caractère hétérogène du point de vue pratique, puisque l'activité médicale se concrétise sous la forme de la recherche et la publication d'articles, la formation de professionnels ou l'assistance médicale dans les hôpitaux et les centres de santé. Par voie de conséquence, l'éventail des contextes communicatifs (tant écrits qu'oraux) dans lesquels les traductions médicales sont requises est extrêmement large et varié, ce qui implique une adaptation du degré de spécialisation du discours et de la formalité de la communication (Muñoz-Miquel, 2016). Cette difficulté est particulièrement accentuée dans le domaine des soins de santé susmentionné, car elle exige donc la tâche particulière de trouver le juste équilibre entre *précision* (c'est-à-dire que le message soit suffisamment précis et rigoureux) et *intelligibilité* (à savoir, qu'il soit suffisamment informatif de façon à être bien compris et interprété par le patient).

En plus de cet obstacle à la communication, il y a de même une conception pas nécessairement unitaire de la médecine, des maladies ou des traitements médicaux dans les différentes cultures existantes (Boujon *et al*., 2018). Ces asymétries se reflètent, par exemple, dans certains modèles d'interaction communicative. Songeons sinon à certains tabous liés à la sexualité ou à d'éventuelles restrictions religieuses en fonction des

croyances du patient, ou encore à la seule pression de parler ou exprimer clairement ses idées au médecin (effet « blouse blanche ») dans un état d'indisposition ou en situation d'urgence. Ce sont vraisemblablement des questions qui peuvent amener le patient à refuser de dévoiler des informations intimes ou à refuser les soins médicaux proposés (Priebe *et al*., 2011).

À la lumière des aspects que nous venons d'évoquer, il semble que la traduction spécialisée dans le domaine des soins de santé possède ses propres règles et contraintes discursives. Nous pouvons donc considérer la traduction médico-sanitaire comme une discipline ayant une identité propre ; il reste cependant à voir comment cette tâche peut être menée à bien.

1.2. La traduction automatisée spécialisée dans le cadre des services médicaux

Idéalement, le soignant communiquerait directement avec le patient dans sa langue maternelle. Pourtant, si le professionnel ne la maîtrise pas, la qualité de la communication est dangereusement réduite (Hudelson, 2019). À la lumière des contextes de plus en plus hétéroglossiques et interculturels, le besoin de services d'interprétation et de médiation entre les médecins et les patients paraît évident. C'est pourquoi, des plateformes de traduction spécialisées adaptées aux échanges médicaux sont nées au cours des dernières décennies afin de garantir une évaluation diagnostique adéquate.

L'un des mécanismes les plus prototypiques de la traduction médico-sanitaire est celui des services d'interprétation par téléphone, qui constituent une option dans le contexte de l'interaction patient-médecin. Néanmoins, ces services n'offrent pas une disponibilité immédiate pour toutes les langues pour lesquelles une traduction est requise (Spechbach *et al*., 2019). Il s'agit, par ailleurs, d'un service coûteux, qui n'est pas entièrement satisfaisant en raison de l'absence d'interaction directe et physique dans le dialogue interpersonnel (Bouillon *et al*., 2017). Certes, nous pourrions conclure que l'existence de membres de la famille ou d'amis pourrait compenser cette absence de communication en présentiel avec le patient, mais elle est également soumise à des contraintes, tel que le fait que l'accompagnant peut changer, écarter ou nuancer le discours du médecin, et donc ne pas garantir une traduction fidèle du message délivré initialement par le spécialiste (Ehsani *et al*., 2008). En plus, elle peut amener à une autolimitation de la personne soignée dans le contexte d'un sujet délicat, tel que le suicide, le cancer ou la sexualité (Hudelson, 2019).

L'intervention d'un interprète professionnel en direct paraît être la solution la plus cohérente dans ce contexte, dans la mesure où il possède la formation jugée nécessaire[1]

[1] La formation académique (médicale ou linguistique) d'un interprète médical est souvent controversée, en raison des compétences spécifiques qu'elle requiert, du point de vue des connaissances terminologiques et de l'exactitude de la traduction requise (Karwacka, 2014).

pour agir en tant que vase communicant entre les deux intervenants, et manque d'un attachement émotionnel au patient qui pourrait déclencher un certain biais dans la traduction du message. Cependant, il se peut que certaines personnes le perçoivent comme une intrusion dans leur intimité, ce qui pourrait influer sur leur relation avec le médecin responsable (Priebe *et al*., 2011). De plus, la disponibilité d'interprètes professionnels spécialisés dans le domaine médico-sanitaire est en soi limitée et, en outre, dans certains cas, il peut être difficile de trouver des spécialistes pour certaines combinaisons de langues source et cible.

L'existence de plateformes de traduction automatique (ou *machine translation*) en ligne s'avère une alternative de plus en plus envisageable, compte tenu de leur coût accessible (Spechbach *et al*., 2019) et des avances en termes de techniques d'apprentissage automatique (Dew *et al*., 2018). En revanche, elle comporte également certaines limites. Ce type de traduction est fortement dépendant des ressources de données à grande échelle et des corpus parallèles de qualité, ce qui n'est pas nécessairement disponible pour des langues peu dotées (Ruiz Costa-jussà *et al*., 2010). Par ailleurs, Google Translate constitue par exemple un service de traduction généraliste, il n'est pas destiné à travailler spécifiquement dans une langue de spécialité, ce qui signifie qu'il n'est pas un outil suffisamment précis pour le discours médical et par là même sa fiabilité peut être mise en question (Turner *et al*., 2019).

D'autre part, ces systèmes sont également susceptibles de produire de mauvaises traductions ou malentendus à cause de leur faible modélisation sémantique et contextuelle (Dew *et al*., 2018). Cela étant dit, il ne semblerait pas professionnel de dégager un diagnostic basé sur une évaluation reposant uniquement sur une telle plateforme ; tout au contraire, dans le domaine de la santé, la traduction automatique nécessite la post-édition et correction humaine pour améliorer et réparer le résultat final des traductions et éviter ainsi des effets dramatiques. Sans perdre de vue qu'elle peut parallèlement poser des problèmes éthiques en termes de respect de la vie privée des patients, en raison d'une protection insuffisante des données (Boujon *et al*., 2018) et de l'absence de couverture des informations sensibles.

Il est à noter, à ce propos, que dans les contextes d'urgence médicale, la correction d'un jugement clinique est cruciale. C'est notamment à la lumière de cet objectif que des plateformes de traduction médicale spécialisées telles que MediBabble (voir la Fig. 1.1), Canopy ou Universal Doctor sont apparues. Leur format est celui des *phraselators* (*Phraselator*, n.d.), c'est-à-dire, qu'elles se basent sur un ensemble fini d'énoncés pré-traduits (et donc préétablis), qu'il s'agisse de questions, d'ordres ou d'affirmations. Bien qu'ils offrent une traduction fiable et soient facilement reproductibles dans d'autres langues (Boujon *et al*., 2018), ce sont des outils limités du point de vue de la communication, à cause de la restriction des phrases traduisibles (Bouillon *et al*., 2007), de l'artificialité de l'interaction et du manque de souplesse dans le dialogue (Ahmed *et al*., 2017).

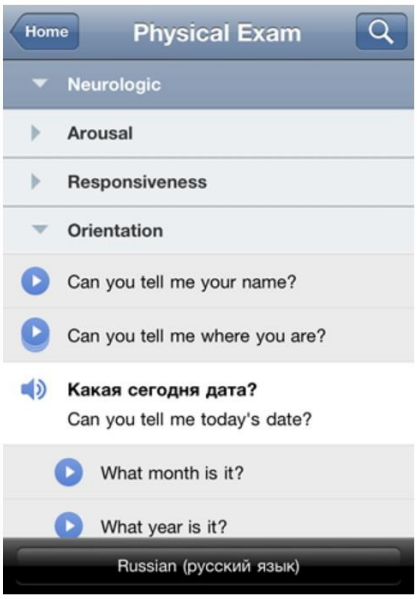

Fig. 1.1 – Exemple d'utilisation du phraselator MediBabble en anglais-russe
[Source : http://www.medibabble.com/screenshots.html].

C'est dans ce contexte que la plateforme présentée ci-dessous, BabelDr, est née, ayant pour but de trouver un bon équilibre entre la traduction automatique, bénéficiant d'une grande flexibilité communicative, et la traduction préétablie traditionnelle, caractérisée par sa précision et clarté (Turner *et al.*, 2019). Ce faisant, BabelDr vise à favoriser une attention médicale efficace dans des contextes d'urgence, et à éliminer autant que possible les barrières linguistiques entre le service médical et la personne soignée lors d'un échange clinique.

1.3. BabelDr : son origine, motivation et fonctionnement

Le projet BabelDr est né de la collaboration entre le département Traitement de l'Information Multilingue (TIM), appartenant à la Faculté de Traduction et d'Interprétation de l'Université de Genève, et l'unité d'urgences ambulatoires des Hôpitaux Universitaires de Genève (HUG). En 2019, le TIM a proposé au Groupe d'Études en Traitement de la Langue et la Parole (GETALP)[2] d'y participer pour aider à la création d'un système de reconnaissance automatique de la parole basé sur des outils libres. Ce faisant, il est possible de s'affranchir du système *speech-to-text* actuellement utilisé dans BabelDr, qui est issu de la société privée Nuance.

BabelDr apparaît en Suisse, qui a reçu près de 100 000 demandes d'asile depuis 2015 (*Swiss Refugee Council*, 2020), et il émerge plus notamment dans le contexte d'un

[2] Il est rattaché au Laboratoire d'Informatique de Grenoble (LIG), affilié à son tour à l'Université de Grenoble-Alpes (UGA).

noyau cosmopolite tel que Genève, où la présence de population allophone[3] marque une forte diversité linguistique. Celle-ci est bien constatable dans les HUG, où les patients parlent plus de 50 langues différentes (Janakiram *et al.*, 2019) et au moins 10% ne parle absolument pas français (Boujon *et al.*, 2018). C'est ainsi qu'une mauvaise compréhension est suspectée de mettre en danger la sécurité et la santé du patient, ce qui implique, par conséquent, un éventuel questionnement éthique de la part des services de santé. Dans ces circonstances, un système de traduction efficace s'avère nécessaire. C'est pourquoi, BabelDr propose un outil de traduction *speech-to-speech*[4] adapté au discours médical permettant au soignant d'énoncer sa phrase oralement, qui sera reconnue par un système de reconnaissance vocale et ensuite traduite dans la langue du patient. Plus précisément, dans le fonctionnement de BabelDr, nous pouvons distinguer les étapes suivantes (Bouillon *et al.*, 2016) :

1. Reconnaissance de la phrase en entrée au moyen d'un système de reconnaissance automatique de la parole. La technologie prévue à cet effet est fournie actuellement par l'entreprise Nuance.
2. Conversion de la phrase reconnue en son équivalent canonique (toujours dans la langue de départ). Cette étape sera plus détaillée par la suite (un exemple de cette phase est consultable sur le Tabl. 1.1).
3. Validation manuelle, effectuée par l'utilisateur, de la phrase canonique source.
4. Pré-traduction manuelle de la phrase source canonique vers la langue d'arrivée par des traducteurs professionnels.
5. Production orale de la traduction résultante à travers des systèmes de synthèse vocale.

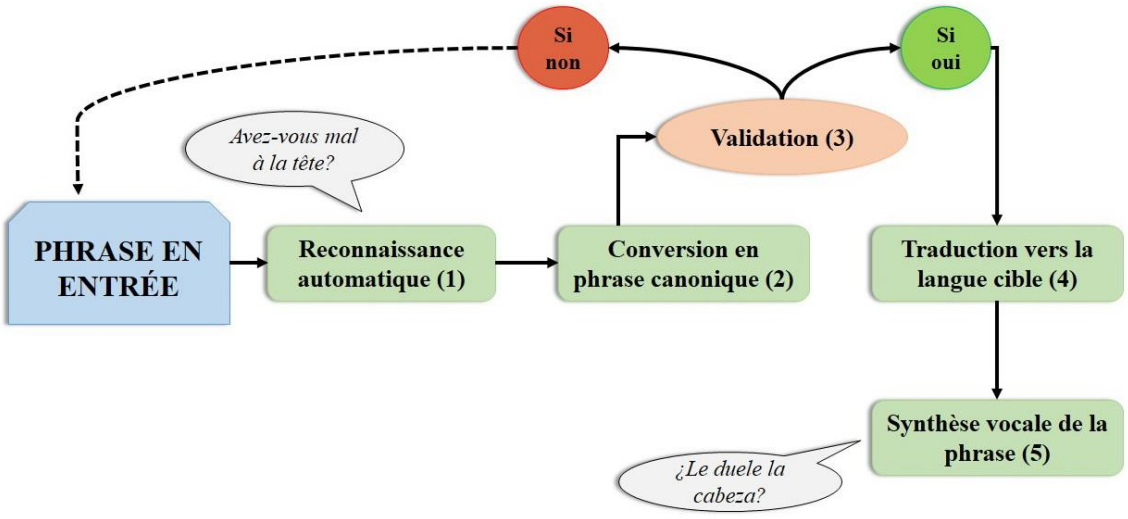

Fig. 1.2 – Schéma représentant l'ensemble d'étapes suivies par BabelDr.

[3] À savoir, les personnes dont leur langue maternelle est une langue étrangère dans la communauté où elles résident (Hudelson, 2019).

[4] Type de traduction aussi dite « parole-parole » ou connue sous le nom de « traduction vocale quasi instantanée ». Elle vise à traduire à haute voix (dans une langue cible) une phrase prononcée par un locuteur, de sorte que tant l'entrée et la sortie du système sont sonores.

Des comparaisons entre BabelDr et d'autres plateformes de traduction disponibles ont été menées dans le contexte d'interaction médicale avec des patients allophones. Une étude récente (Bouillon *et al*., 2017) a révélé les faibles scores d'*intelligibilité* (à savoir, si un message est compréhensible ou non, grammatical ou agrammatical) et d'*adéquation* (un message est sémantiquement correct, ambigu ou contradictoire) de Google Translate par rapport à BabelDr, dans un contexte d'entretien diagnostique avec des médecins francophones et des patients arabophones. Alors que le premier a atteint 38% dans les deux sections, il faut noter que le deuxième a obtenu 94% et 93% respectivement (Bouillon *et al*., 2017), ce qui montre son intérêt du point de vue de la langue de spécialité discutée ici.

Il est possible que les plateformes basées sur les mécanismes d'un *phraselator* comme celles déjà mentionnées ci-dessus offrent de meilleurs résultats que ceux dérivés d'une traduction automatique effectuée par Google Translate ; leur utilisation est pourtant moins confortable que celle dont BabelDr est dotée. C'est du moins ce qu'indique une autre expérience récente menée par les HUG, dans laquelle ils ont mis en évidence que le temps moyen requis par le spécialiste pour interagir avec l'interface de BabelDr était inférieur (20 secondes en moyenne) à celui exigé par MediBabble (30 secondes) (Boujon *et al*., 2018). Cela semble indiquer que le spécialiste a trouvé moins de difficultés à manipuler le premier et suggère donc une utilisation plus intuitive ou plus facilement manipulable.

Cette particulière maniabilité de BabelDr est due à l'incorporation du système de reconnaissance vocale automatique susmentionné, qui ergonomise l'utilisation de la plateforme et facilite une interaction médecin-patient (voir la Fig. 1.3). Actuellement, l'application est unidirectionnelle et effectue une traduction de la langue de départ, le français, vers la langue d'arrivée. C'est pourquoi, le patient doit répondre par oui ou non, avec des gestes ou en écrivant des informations (Janakiram *et al*., 2019) à ce que l'utilisateur (à savoir, le spécialiste) a énoncé. Grâce à une modélisation linguistique effectuée par des traducteurs, BabelDr permet aux médecins de trouver les phrases à énoncer en s'adressant au système oralement et en utilisant un large éventail de paraphrases et de variations stylistiques (Bouillon *et al*., 2016). Cette robustesse du système permet ainsi que les utilisateurs s'expriment plus librement (Rayner *et al*., 2018), car ils n'ont pas à rechercher la phrase souhaitée dans une liste préétablie à la manière des *phraselators* traditionnels.

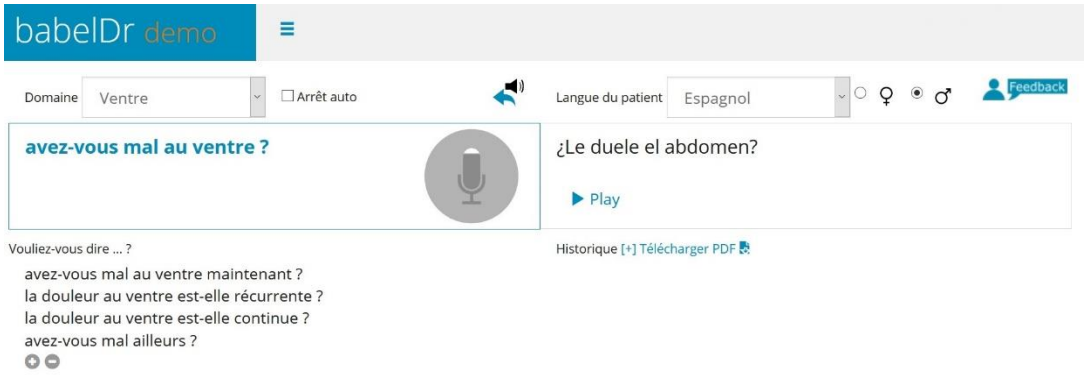

Fig. 1.3 – Exemple d'utilisation de l'interface BabelDr pour une traduction français-espagnol [Source : https://regulus.unige.ch/babeldrclient].

En tout état de cause, il faut également garder à l'esprit que ce n'est pas seulement une convivialité avec l'utilisateur qui est visée, mais aussi et surtout une bonne compréhension entre le médecin et le patient de manière à favoriser un diagnostic correct le plus rapidement possible. Et il convient de noter que, à la lumière de cette recherche de *fiabilité* (Mutal *et al*., 2019), BabelDr propose un système de vérification manuelle.

Ce mécanisme de sécurité consiste, d'une part, en la simplification de l'énoncé produit par le locuteur et l'inspection ultérieure de la phrase résultant de ce processus. En d'autres termes, une fois le résultat de la reconnaissance de la phrase source est obtenu, il sera attribué à une *core sentence* ou « phrase canonique », par un processus de *backtranslation* ou « rétro-traduction monolingue »[5]. Dans la version actuelle de BabelDr, cet objectif est atteint grâce à une méthode basée sur des règles. Pourtant, des approches basées sur des réseaux de neurones ont été également explorées.

Cette procédure se compose d'une simplification lexicale, syntaxique et sémantique (Mutal *et al*., 2019), qui vise à réduire le jargon médical, à éviter les éventuelles ambiguïtés et à accroître l'explicitude du message[6] (le Tabl. 1.1 propose des exemples issus du domaine anatomique « ventre »). C'est à ce moment que le médecin décide, manuellement, d'approuver ou de rejeter la rétro-traduction produite et, si le système reçoit l'autorisation correspondante, il cherchera la traduction de cette phrase canonique dans la langue cible[7]. Cette phrase sera ensuite émise par un système texte-parole ou, dans le cas de langues minoritaires ou de langues des signes, par des fichiers préenregistrés (Boujon *et al*., 2018).

[5] Même si la notion de « rétro-traduction » renvoie souvent à la réécriture d'une traduction dans la langue originale, dans le cas de BabelDr elle réfère à la conversion de la phrase transcrite par le système de reconnaissance en son équivalent canonique, toujours dans la langue de départ. La traduction vers la langue cible sera effectuée si et seulement si la phrase canonique est favorablement validée par le médecin.

[6] BabelDr compte actuellement environ 30 000 phrases canoniques, chacune étant associée à un domaine anatomique spécifique (ventre, tête ou traumatologie, entre autres).

[7] Tel que mentionné précédemment, cette phrase est pré-traduite par des traducteurs professionnels.

Phrase source	***Phrase canonique***
votre ventre vous fait-il mal ?	*avez-vous mal au ventre ?*
avez-vous des antécédents chirurgicaux au niveau de l'abdomen ?	*avez-vous eu une opération du ventre ?*
est-ce que vous pourriez me montrer votre carte d'assuré ?	*pouvez-vous me montrer la carte d'assurance ?*

Tabl. 1.1 – Alignement de phrases sources avec leurs respectives phrases canoniques (Mutal *et al*., 2019).

Nous ne pouvons pas ignorer que c'est en raison de cette caractéristique précise que, d'un point de vue davantage psychologique, les médecins interrogés dans les expériences citées ci-dessus ont affirmé se sentir moins contraints via BabelDr. Ceci était possible grâce au fait qu'ils n'ont pas trouvé autant d'obstacles lors de la réalisation de l'anamnèse (à savoir, le recueil d'informations sur les antécédents médicaux) et, ce faisant, l'interaction est devenue plus naturelle (Boujon *et al*., 2018). Ainsi, la convivialité ressentie à l'égard du système suggère avoir un impact sur la correction des diagnostics émis.

L'intégration de la reconnaissance vocale dans le domaine de la traduction médicale n'est pas une nouveauté. Preuve en est, par exemple, le projet S-MINDS (*Speaking Multilingual Interactive Natural Dialog System*), mené au complexe hospitalier *Kaiser Permanente* de San Francisco, en Californie. Ce projet répondait au pourcentage élevé de population LEP (*Limited English Proficiency* ou ayant une connaissance très limitée de la langue anglaise), dont les locuteurs hispanophones représentent une majorité (Ehsani *et al*., 2008). Étant donnés les risques en termes d'erreurs médicales, de faux diagnostics et de soins de santé médiocres (Priebe *et al*., 2011) auxquels ils étaient exposés, une plateforme de traduction a alors été conçue, intégrant un système de reconnaissance vocale automatique, espagnol-anglais et anglais-espagnol.

Le projet BabelDr vise à construire, certes, un système à l'image de S-MINDS, la différence étant qu'il vise à l'appliquer de manière multilingue. Les langues le plus couramment parlées par les patients des HUG sont l'espagnol, l'albanais, l'arabe, le farsi, le dari et le tigrinya. C'est pourquoi, l'application fournit des traductions vers ces six langues. De plus, l'incorporation de l'anglais, le turc et la langue de signes de la Suisse Normande ainsi que le développement d'une nouvelle version bidirectionnelle (permettant de transcrire et traduire la parole produite par le médecin ainsi que le patient) sont actuellement en préparation. À partir de cet intérêt pour le plurilinguisme, on comprend alors le principe de *portabilité* ou d'*adaptabilité* visé par les auteurs de ce projet, auquel s'ajoute une recherche de la *sécurité* et de la *confidentialité* des patients, qui repose sur l'utilisation de serveurs locaux sécurisés et le stockage local des données (Mutal *et al*., 2019).

La pertinence d'un système tel que BabelDr ne peut ainsi être niée dans le contexte de notre époque, où l'écart entre la nécessité et la disponibilité des services linguistiques dans le domaine médico-sanitaire est flagrant. L'impossibilité d'une intercommunication adéquate et bénéfique entre le patient et le médecin met en péril la santé du premier, dans

la mesure où des soins inadéquats ou inefficaces peuvent conduire à une détection incorrecte de pathologies ou à des procédures plus invasives (Ku & Flores, 2005). Cela répercute négativement sur les coûts de santé, mais surtout sur la sécurité de la personne traitée.

De cette manière, BabelDr émerge en tant qu'une réponse à un besoin. Les services d'urgence s'occupent de patients avec lesquels ils ne partagent pas forcément la même langue ; cela constitue un fait incontournable. Toutefois, il semble approprié qu'il y ait des moyens de surmonter cet obstacle et de garantir ainsi que les patients soient correctement soignés, quelle que soit leur langue.

Comme nous l'avons indiqué tout au long de ce chapitre, BabelDr vise à répondre efficacement à cette tâche. Il va de soi que la construction de systèmes efficaces de traduction vocale médicale semble être la solution la plus cohérente dans le domaine des soins de santé. C'est pourquoi, dans le présent mémoire, nous allons explorer la première étape du processus de traduction effectué par BabelDr, qui correspond à la reconnaissance automatique de la parole émise par l'utilisateur. Nous avons déjà signalé que le logiciel venant de Nuance est actuellement utilisé pour effectuer cette tâche. Cependant, avec ce travail, nous présentons un système de reconnaissance alternatif, basé sur des outils *open source* et appliqué sur le français, dans le but de rompre la dépendance d'un système externe et de le faire évoluer selon les besoins des Hôpitaux Universitaires de Genève.

2. La reconnaissance automatique de la parole

2.1. Brève introduction au domaine

La reconnaissance automatique de la parole est devenue une réalité de plus en plus répandue dans le monde d'aujourd'hui. La croissance et l'expansion des assistants vocaux tels que Siri (lancé en 2011) Cortana et Alexa (lancés en 2014), ou Google Assistant (en 2017) en sont la preuve, et constituent l'aboutissement, à présent, des fruits résultants de la recherche et du développement dans un domaine qui compte déjà 70 ans d'histoire et continue à travailler sans relâche pour relever de nouveaux défis.

L'origine de la parole, en revanche, remonte à des dizaines de milliers d'années et, dès ses débuts, elle semble s'être imposée en tant que le mécanisme de communication prédominant de l'espèce humaine. Certes, la capacité langagière est souvent décrite comme une faculté distinctive de notre nature vivante et humaine, mais cela ne nous a pas empêché de transférer certaines de nos connaissances linguistiques à la machine pour qu'elle puisse les reproduire. L'apparition de l'écriture il y a environ 6000 ans a signifié la création d'un instrument permettant de transcrire manuellement la parole et, par là même, la conservation et la transmission de l'information. Aujourd'hui, cette tâche est présentée comme un processus qui peut être automatisé grâce au traitement informatique de la parole.

Le domaine de la reconnaissance automatique de la parole intéresse la science depuis le milieu du XXe siècle (Rabiner & Juang, 2004). D'ailleurs, grâce au progrès de la technologie en matière de ressources informatiques et l'augmentation progressive de la taille des données d'apprentissage (Huang *et al.*, 2014), les systèmes sont devenus de plus en plus sophistiqués. Il est intéressant de noter que le grand public, aujourd'hui souvent muni de dispositifs de reconnaissance vocale dans sa poche, a également accompagné la science dans cette fascination pour que la machine soit capable de comprendre ce que nous disons. Des films classiques comme *2001 : A Space Odissey* ou la saga *Star Wars* (Llisterri, 2003; Rabiner & Juang, 2004) sont souvent évoqués comme des exemples qui ont révélé, dans une perspective fictive mais de moins en moins invraisemblable, des personnages incarnés sous la forme de machines parlantes, capables

de reconnaître le discours humain et d'y répondre en conséquence. Il ne semble donc pas étrange que ce soit précisément quelques années plus tôt, dans les années 50, que le domaine du traitement de la parole a commencé à fleurir, tel que nous le verrons dans la section suivante.

Mais avant de commencer par un bref historique du domaine, il s'avère tout d'abord nécessaire d'éclaircir une définition de ce que nous entendons par parole et ce que nous « reconnaissons » exactement lorsque nous opérons un traitement de la parole. D'une manière générale, la parole est la capacité humaine à communiquer ou transmettre un sens oralement au moyen d'un message. Tel message, pour être effectivement compris comme une production de la parole humaine, est nécessairement *linguistique*, dans la mesure où il a été émis dans un code, une langue spécifique, et doit donc être adapté, à son tour, aux règles et conventions grammaticales partagées par la communauté linguistique à laquelle il appartient.

Nous pourrions alors conclure que toute production orale humaine qui véhicule un sens est linguistique. Mais pas exclusivement. Ledit message est de même accompagné d'autres aspects :

- *Paralinguistiques*, c'est-à-dire, les éléments non verbaux associés à l'élocution, tels que le volume, l'intonation ou le rythme. Ils se révèlent des indicateurs éventuels de l'attitude ou de l'état émotionnel du locuteur (Besacier, 2018) et sont susceptibles d'avoir un impact sur le sens de ce qui est censé être communiqué.
- *Extralinguistiques*, dans la mesure où ils sont extérieurs à la structure de la langue et font référence à des facteurs extrinsèques qui influencent son usage. Ils peuvent, à leur tour, être regroupés en :
  - *Sociolinguistiques*, qui réfèrent à l'influence des aspects sociaux tels que le sexe, l'âge, le niveau d'éducation ou l'origine ethnique sur l'usage de la langue par le locuteur (Moreno Fernández, 1998).
  - *Géographiques*, relatifs à l'influx de la variation dialectale sur les réalisations linguistiques.
  - *Stylistiques*, liés au registre de langue ainsi qu'à la variation qui résulte du contexte situationnel dans lequel le message est exprimé (de plus soutenu à plus relâché).

Toutes ces composantes s'imbriquent entre elles pour définir la parole. Cependant, la reconnaissance automatique de la parole utilise l'élément strictement linguistique pour accomplir sa tâche, dans la mesure où elle cherche principalement à effectuer une transcription intégrale du message (correspondant à l'entrée sonore) en mots graphiques.

Contrairement à ce que l'on pourrait croire, la reconnaissance de la parole humaine est une tâche ardue. En effet, le signal de parole est continu, la variabilité et l'instabilité sont inhérentes au discours oral et la nature mixte de la discipline se reflète dans la convergence des champs comme la physiologie, la linguistique ou l'informatique. Ces

facteurs rendent le travail problématique et complexifient l'objectif qui est visé par le domaine traité.

Hors du secteur de la recherche, la reconnaissance vocale n'est souvent pas un objectif en soi, mais plutôt une phase d'un processus qui cherche à résoudre un problème plus complexe, que ce soit la compréhension de la parole ou la traduction de la parole. Nous proposons dans la section suivante un rappel historique des travaux de reconnaissance automatique de la parole qui permettra de mieux comprendre ses principaux axes d'intérêt et les projets dans lesquels ces technologies se sont matérialisées.

2.2. Aperçu historique de la reconnaissance automatique de la parole

La genèse de l'automatisation de la reconnaissance vocale est apparue aux États-Unis en 1952 avec *Audrey*, une machine analogique créée aux Laboratoires Bell, capable de reconnaître des nombres à un seul chiffre prononcés par un locuteur entre deux pauses (Pieraccini, 2012). À cette époque, le fonctionnement du système était guidé par des théories acoustico-phonétiques basées sur le comportement des formants, dont le but était de distinguer les voyelles émises dans un acte verbal, et donc de différencier un chiffre d'un autre. Parmi les autres expériences similaires, citons *Shoebox*, présentée en 1961 par la société IBM, qui visait à aller plus loin et à créer une calculatrice à commande vocale basée sur une reconnaissance numérique et verbale limitée. En 1962, sur la base des expériences de reconnaissance de la parole menées à l'Université de Kyoto au Japon, un dispositif hardware qui opérait la segmentation de la parole a été conçu dans le but d'identifier les éléments phonémiques constituants d'une phrase (Rabiner & Juang, 1993).

Mais ce n'est qu'au début des années 1970 que la reconnaissance vocale a connu une explosion aux États-Unis, grâce au financement de *Advanced Research Projects Agency* (ARPA), dépendant du Département de la Défense, pour le programme de recherche *Speech Understanding Research* (SUR). Cela a facilité l'avancement sur des projets tels que *Hearsay* ou encore *Harpy*, un système de reconnaissance vocale qui utilisait un vocabulaire de 1000 mots et qui fournissait des résultats de transcription acceptables (Rabiner & Juang, 2004).

Pendant ce temps, les compagnies IBM et AT&T Bell ont développé des stratégies commerciales différentes. IBM s'est concentré sur *Tangora*, un système de reconnaissance dépendant du locuteur et axé sur la correspondance administrative. La société, qui cherchait à atteindre la reconnaissance d'un large vocabulaire (Le Blouch, 2009), a montré que les statistiques commençaient à présenter un intérêt significatif du côté de la modélisation linguistique, qui était jusqu'alors traitée par des systèmes à base de règles. Les laboratoires AT&T Bell ont toutefois opté pour le développement d'applications commerciales destinées au grand public dans le cadre des services

téléphoniques. L'accent était davantage mis sur la création de systèmes robustes à la variabilité inter-locuteur, ce qui a mené à la conception de nouveaux algorithmes de classification (Rabiner & Juang, 2004).

Les années 1970 ont marqué le début de la familiarisation avec les processus stochastiques visant à modéliser la parole, comme en témoigne le système *Dragon* (Huang *et al*., 2014), introduit en 1975 et rattaché par la suite à la société *Dragon Systems*. Il faudra cependant attendre les années 1980 pour assister à une véritable éclosion des méthodes de modélisation statistique, et avec elles, des modèles (ou chaînes) de Markov cachés (HMM, *Hidden Markov Models*), qui se distinguent par leur robustesse par rapport à la liaison entre les traits acoustiques et les phones. La publication de littérature scientifique sur leur méthodologie et leur fonctionnement a en effet permis aux HMM de gagner en notoriété et en acceptation, à tel point qu'ils ont été adoptés globalement par la grande majorité des laboratoires de recherche en reconnaissance vocale (Rabiner & Juang, 1993). Depuis la convergence vers cette technique, toujours en vigueur aujourd'hui, l'intérêt s'est orienté vers la reconnaissance de la parole continue basée sur des grands vocabulaires (LVCSR, *Large Vocabulary Continuous Speech Recognition*). La création de systèmes robustes capables de reconnaître la chaîne parlée semblait de plus en plus atteignable

La révolution technique provoquée par la mise en œuvre de HMM a signifié, avec l'arrivée des années 1990, un nouvel investissement dans les technologies de la parole par ARPA, qui a cherché à relever des défis inexplorés à l'époque, tels que la reconnaissance de la parole spontanée ou la transcription des informations diffusées. Les résultats se sont concrétisés dans des projets tels que *Byblos* ou *Decipher* (Rabiner & Juang, 2004), mais ce n'était pas la seule voie de la croissance dans le domaine, car des solutions commerciales sont également apparues. La société *Dragon Systems*, citée ci-dessus, a lancé *Dragon Dictate* en 1990, qui disposait d'un vocabulaire de 5000 mots et qui exigeait de l'utilisateur une pause entre les mots pour effectuer la transcription correspondante. Des années plus tard, en 1997, la société a connu un grand succès suite au lancement de *Dragon NaturallySpeaking*, une version améliorée du précédent, dont le vocabulaire comprenait 23000 mots.

Dans ce contexte de peaufinage et innovation continues, il faut tenir également compte de l'apparition de logiciels tels que *HTK Speech Recognition Toolkit*, conçu au département d'ingénierie de l'Université de Cambridge (Young *et al*., 1995), qui a notamment facilité la tâche de construction de HMM et donc la création de systèmes de reconnaissance vocale. Cette tendance s'est poursuivie au tournant du siècle avec la gestation et le développement de nouvelles plateformes *open source* de reconnaissance vocale telles que *Kaldi*, *Sphinx* ou *Julius* (Huang & Deng, 2010). En outre, les progrès accélérés des algorithmes d'apprentissage automatique liés à la sophistication continue des infrastructures informatiques ont également favorisé la création et l'expansion de solutions commerciales telles que *Google Voice Search*, paru en 2008, qui permettait la recherche vocale sur des téléphones portables.

L'augmentation de la capacité de calcul et l'accès à des collections de données de plus en plus importantes ont ravivé l'intérêt pour les réseaux neuronaux, ce qui a entraîné une large utilisation de méthodes d'apprentissage non supervisées (c'est-à-dire, où les données ne sont pas étiquetées au préalable). L'explosion au cours des dernières années de l'apprentissage profond et des réseaux de neurones profonds (DNN, *Deep Neural Networks*) s'est avéré une voie alternative à l'approche proposée par les HMM[8], et a démontré sa grande performance dans les tâches de reconnaissance vocale. Cela a notamment multiplié la présence des systèmes de reconnaissance de parole dans nos vies.

Le sous-titrage automatique sur Youtube ou les systèmes de messagerie vocale font partie des technologies disponibles aujourd'hui, tout comme les assistants vocaux déjà très répandus. Siri, Cortana, Alexa et Google Assistant (appartenant respectivement à Apple, Microsoft, Amazon et Google) sont désormais embarqués dans un plus grand nombre d'applications. Au cours des dernières années, nous avons pu constater l'ubiquité émergente de tels dispositifs par le biais de la technologie fleurissante de l'*Internet-of-Things*. Aussi appelée « Internet des objets », elle réfère à l'interaction entre des appareils quotidiens reliés à Internet (tels que les haut-parleurs, les télévisions, les lampes, les réfrigérateurs, les aspirateurs ou les montres) avec des utilisateurs humains, au moyen d'assistants intelligents.

L'idée de la « maison intelligente » semble, en effet, avoir de nombreux avantages, dans la mesure où elle offre des possibilités insoupçonnées auparavant et favorise les économies énergétiques ou l'assistance aux personnes. Toutefois, il convient également de noter qu'elle entraîne des problèmes de confidentialité : les assistants vocaux, de manière à opérer la reconnaissance, doivent se connecter à des serveurs sur Internet auxquels ils envoient le signal acoustique reçu et ainsi être en mesure de l'interpréter.

En conclusion, le domaine du traitement de la parole a été confronté à des défis de plus en plus sophistiqués depuis sa naissance ; le passage de l'identification des chiffres jusqu'à la reconnaissance de la parole continue basée sur des grands vocabulaires en est la preuve. Dans un processus constant d'essai-erreur, elle a su surmonter les obstacles qui se sont dressés sur son chemin, tout en s'adaptant à la disponibilité des ressources de chaque époque et en imaginant sans cesse des méthodes alternatives d'amélioration des systèmes.

De tout cela est né le développement technologique que nous connaissons et contrôlons aujourd'hui. Néanmoins, il ne fait aucun doute que l'histoire de la reconnaissance vocale n'est pas terminée et qu'il reste encore de nombreux défis à relever, tels que la reconnaissance de parole appliquée à des domaines de spécialité (tel que le domaine médical) ou la création de systèmes de reconnaissance des langues peu dotées. Il s'avère intéressant, en tout cas, que le contexte actuel ressemble de plus en plus au désormais classique et fictif HAL 9000, qui impressionna le public à la fin des années 60 par sa reconnaissance impeccable de la voix humaine et du traitement de la langue.

[8] Cette division n'est pas totalement exclusive : il existe également des approches hybrides DNN/HMM qui fournissent des résultats très satisfaisants en termes de taux d'erreur (Huang *et al*., 2014).

2.3. Fonctionnement d'un système de reconnaissance vocale

2.3.1. Principes de base

Nous avons vu que l'objectif d'un système de reconnaissance vocale est de transcrire la parole émise par un locuteur, ce qui implique une conversion de l'entrée sonore au texte écrit correspondant. Or, des questions peuvent être posées à ce sujet. Nous pouvons effectivement nous interroger sur la manière dont les sons sont reconnus dans le signal sonore, ou comment est-ce qu'ils peuvent être « traduits » dans la bonne séquence des mots prononcée par le locuteur. Un bon départ pour éclaircir ces questions nous le fournit une formulation mathématique, qui nous permettra de distinguer plus nettement les différents problèmes rencontrés par cette discipline.

Nous avons déjà évoqué que le système reçoit un signal sonore qui est ensuite décodé. Après un processus de discrétisation, cette entrée est représentée sous la forme d'une séquence de symboles, qui est produite au fur et à mesure que le locuteur parle. Dans notre formulation mathématique, nous les identifions comme l'ensemble **A**, dont les symboles appartiennent à un alphabet donné, $\boldsymbol{\alpha}$:

$$A = \{ (a_1, a_2, \ldots, a_m) \mid a_i \in \alpha, i = 1, 2, \ldots, m \}$$

Le résultat de la conversion se concrétise sous la forme d'une transcription du message sonore, présente dans l'ensemble **W**, qui désignera une série finie de mots appartenant à un vocabulaire donné, $\boldsymbol{\omega}$:

$$W = \{ (w_1, w_2, \ldots, w_n) \mid w_i \in \omega, i = 1, 2, \ldots, n \}$$

Avec ces éléments et grâce à la formule de Bayes, nous pouvons énoncer l'équation qui sous-tend un système de reconnaissance vocale :

$$P(W|A) = \frac{\mathbf{P}(\mathbf{W}) \times \mathbf{P}(\mathbf{A}|\mathbf{W})}{\mathbf{P}(\mathbf{A})}$$

Où **P(W)** représente la probabilité que la séquence de mots **W** soit générée, **P(A|W)** est la probabilité que lorsqu'un locuteur émet **W**, la séquence acoustique **A** soit observée, et où **P(A)** est la probabilité moyenne que **A** soit observée (Jelinek, 1998). Vu que la séquence acoustique **A** est fixée, **P(A)** est une valeur constante ; nous pouvons donc la supprimer de la formule. Nous obtenons ainsi :

$$\hat{W} = \arg\max_{\mathbf{W}} \mathbf{P}(\mathbf{W}) \times \mathbf{P}(\mathbf{A}|\mathbf{W})$$

Qui met en évidence que le système de reconnaissance vise à rechercher la séquence de mots $\hat{\mathbf{W}}$ qui maximise le produit **P(W) × P(A|W)**. Il faut noter que ces probabilités sont incarnées par les différents constituants d'un système de la parole :

- La probabilité **P(A|W)** sera calculée par les modèles acoustiques (Fig. 2.1).
- La probabilité **P(W)** sera calculée par les modèles de langue ou grammaires, qui modéliseront la probabilité d'occurrence d'un mot ou d'un ensemble de mots. Notons que cet ensemble de mots fait partie d'un dictionnaire ou modèle de prononciation, qui établit l'association entre chaque mot et son équivalent phonétique (Adda-Decker & Lamel, 2000).

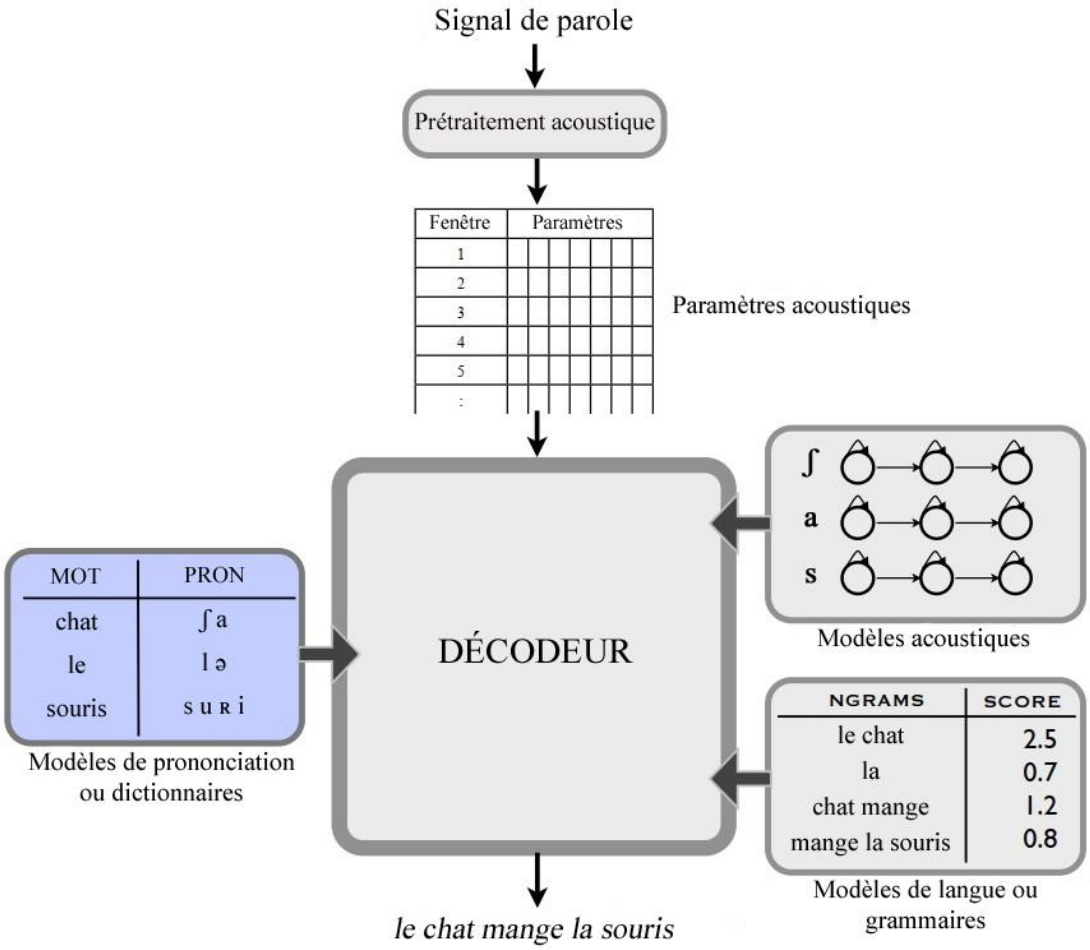

Fig. 2.1 – Architecture typique d'un système de reconnaissance de la parole (inspirée de (Jyothi, 2019)).

De cette manière-là, le décodeur, représenté par $\arg\max_{\mathbf{W}} \mathbf{P}(\mathbf{W}) \times \mathbf{P}(\mathbf{A}|\mathbf{W})$, peut alors effectuer la transcription, et rechercher ainsi l'hypothèse optimale correspondant aux paroles du locuteur. Nous examinerons plus en détail chacun des éléments constitutifs d'un système de reconnaissance automatique vocale dans la section suivante.

2.3.2. Structure prototypique d'un système de reconnaissance de la parole

2.3.2.1. L'extraction de paramètres acoustiques

De manière à opérer une reconnaissance automatique de la parole, il est tout d'abord nécessaire de procéder à un prétraitement du signal acoustique, l'objectif étant de convertir l'entrée sonore en une représentation discrète et de générer ainsi les paramètres acoustiques ou *acoustic features* (Jyothi, 2017).

L'existence d'une telle étape s'explique par des raisons particulières. D'une part, le processus de discrétisation est nécessaire puisqu'il n'est pas possible de travailler directement sur l'entrée sonore brute[9]. De cette manière, le signal de parole, obtenu via un capteur de pression acoustique (à savoir, un microphone), subit un processus d'échantillonnage, d'où on obtient une séquence d'éléments qui émerge en tant que représentation discrétisée de la parole (Jelinek, 1998). Ou plus précisément, cela suppose la génération d'une séquence de vecteurs des paramètres acoustiques à des intervalles réguliers (entre 20 et 30 millisecondes de manière générale), qui sont porteurs d'informations non redondantes sur le signal.

Par ailleurs, cette étape est également guidée par l'idée que cette représentation doit extraire des informations véritablement pertinentes sur la parole humaine. Cela se traduit par l'utilisation de paramètres acoustiques tels que *Linear Predictive Coding* (LPC), qui sont fondés sur la théorie « source filtre ». Cette théorie part du principe que la production des sons se divise en deux ; elle différencie la « source », à savoir, la transformation du courant d'air en voix (et opérée par les organes de phonation) du « filtre », qui réfère à la transformation de la voix en parole et résulte de l'agencement des articulateurs (Vaissière, 2006). De cette manière, il est possible d'expliquer la production des différents sons à la lumière de la relation entre la physiologie et les formants.

Toutefois, cette phase répond également à l'objectif de rendre cette représentation similaire au système perceptif humain, ce qui découle des *Mel Frequency Cepstral Coefficients* (MFCC), qui constituent les paramètres acoustiques les plus répandus dans le domaine de la reconnaissance de la parole. Son succès s'explique par l'intérêt de ses concepteurs, Davis et Mermelstein, à imiter le comportement auditif humain. C'est ainsi qu'avec l'application de la transformée de Fourier et la conversion du spectre à l'échelle de Mel, il est possible d'émuler le fonctionnement logarithmique et non linéaire de l'oreille humaine. Il faut noter, à cet égard, que l'ouïe dont nous sommes dotés permet d'identifier plus finement les changements de fréquence dans les basses fréquences, où juste une échelle de 20-30 Hz est suffisante pour percevoir des différences. Pourtant, elle

[9] Il faut savoir qu'il y a de nouvelles approches basées sur des réseaux de neurones parus dès 2015 qui travaillent directement sur le signal. Au lieu d'extraire des caractéristiques sur le signal avec des techniques préétablies, le système est conçu pour apprendre à extraire des représentations qui soient les plus efficaces pour la tâche à résoudre (Besacier, 2018). De cette manière-là, le traitement du signal ne constitue pas une étape de prétraitement, mais se rattache à l'entraînement des modèles acoustiques et leur modélisation.

devient de plus en plus insensible aux changements dans des fréquences supérieures, qui nécessitent des périodes fréquentielles plus larges pour que l'oreille perçoive des variations. De cette façon, le calcul des MFCC met à disposition une représentation discrète en accord avec ce phénomène.

2.3.2.2. Les modèles acoustiques

Une fois que les paramètres acoustiques ont été extraits, ils serviront d'entrée pour la prochaine étape, relative aux modèles acoustiques, dont la sortie correspondra à une séquence de phones.

Avant de passer en revue le processus qui mène à terme cette opération, il paraît nécessaire d'éclaircir au préalable la notion de *phone* et de quelle manière elle se distingue de celle de *phonème*. Notons, à cet égard, que les phones sont reliés à l'étude de la phonétique, à savoir, de la réalisation concrète de la parole, ou tout simplement aux sons produits par un locuteur lorsqu'il émet un message. Les phonèmes, en revanche, constituent l'unité minimale fonctionnelle du champ de la phonologie, qui vise à examiner plutôt le système de sons d'une langue (aussi nommé « système phonologique ») et comment les phonèmes établissent des oppositions entre les mots de son lexique (Vaissière, 2006). Nous pouvons donc en déduire un caractère plus théorique ou plus abstrait de cette discipline.

Pour illustrer cette différence, il suffit de songer à un exemple relié à la langue courante. En français, /d/ et /ʁ/ sont des phonèmes distincts, dans la mesure où la substitution de /d/ par /ʁ/ produit deux mots différents, tels que *dent* (/dɑ̃/) et *rang* (/ʁɑ̃/). En revanche, « le r uvulaire parisien, prononcé [ʁ] et le r apical roulé [r] (le [r] dit bourguignon) sont deux variantes régionales d'un seul et même phonème /ʁ/ » (Vaissière, 2006), ce qui les place dans le domaine phonétique et donc dans le cadre des phones.

En effet, nous constatons qu'un mot (concrétisé à travers les graphèmes appartenant à une langue particulière) peut être représenté sous la forme d'une séquence de phonèmes ou encore de phones. D'ailleurs, ce que les modèles acoustiques visent à faire, c'est de transformer les vecteurs de paramètres acoustiques en une séquence de phones qui soit cohérente avec la représentation donnée en entrée[10]. Pour accomplir cette tâche, différentes techniques ont été utilisées, comme celles offertes par les chaînes de Markov cachées (HMM), introduites dans les années 1980 et toujours en vigueur de nos jours.

Les modèles acoustiques permettent d'associer un phone spécifique à un ensemble de vecteurs de caractéristiques acoustiques, de sorte qu'ils doivent déterminer les phones qui correspondent à chaque trame, mais également à quel instant il faut passer d'un phone à l'autre. Cette problématique est modélisée par les chaînes de Markov cachés (HMM) selon une approche basée sur des automates probabilistes à états finis. D'après un calcul mathématique, on peut déterminer la probabilité de transition d'un phone à un autre

[10] Comme nous le constaterons plus tard, cela ne se réalise pas dans le cas des systèmes appelés *End-to-End*, où les paramètres acoustiques trouvent leur correspondance directe dans les graphèmes.

(Rabiner & Juang, 1986). De plus, on obtient la probabilité d'émission de chaque phone, qui correspond aux transitions dans chaque HMM (Jelinek, 1998).

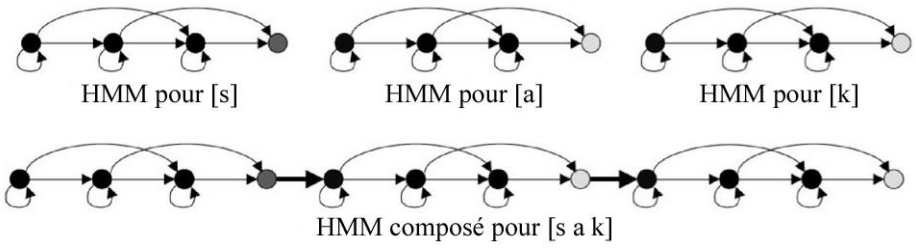

Fig. 2.2 – Exemple simplifié du fonctionnement des modèles de Markov cachés pour la séquence de phones [s a k].

Pour modéliser la distribution de probabilités des paramètres acoustiques pour un phone nous pouvons avoir recours actuellement à plusieurs techniques. Parmi les plus connues, nous retrouvons les *Gaussian Mixture Models* (GMM), qui s'associent avec les HMM de manière à tirer une hypothèse sur la séquence de phones la plus probable sur la base d'un jeu de données d'entraînement.

Au cours des dernières années, de nouveaux modèles sont également apparus dans le but de ressortir les probabilités de distribution autrement. Ces approches HMM-DNN, utilisant de réseaux de neurones, sont souvent connus sous le nom de « modèles hybrides » et présentent plusieurs avantages par rapport aux modèles dits HMM-GMM. Plus particulièrement, les réseaux de neurones, à la différence des GMM, peuvent apprendre des représentations acoustiques plus complexes, ce qui se traduit par une meilleure performance du système de reconnaissance (Maas *et al*., 2017). C'est précisément pour cette raison que nous utiliserons ce modèle dans l'étude que nous présentons ici. Les modèles hybrides HMM-DNN sont de manière générale combinés avec certaines techniques d'adaptation acoustique telles que SAT (*Speaker Adaptive Training*) ou fMLLR (*feature space Maximum Likelihood Linear Regression*), qui permettent de normaliser les paramètres acoustiques de façon à améliorer une reconnaissance de la parole indépendante du locuteur (Miao *et al*., 2014).

En fin de compte, nous constatons que l'objectif principal est d'aligner de la manière la plus efficace possible des échantillons du signal de parole aux phones émis par un locuteur. Pourtant, cette étape n'est pas suffisante pour aboutir à une transcription du message prononcé ; il restera encore à convertir la séquence de phones obtenue en la séquence de mots la plus cohérente (du point de vue sémantique et syntaxique) avec ladite entrée, ce qui s'effectuera ensuite avec les modèles de prononciation et de langue.

Au cours des dernières années, plusieurs expériences ont proposé une approche alternative à la reconnaissance automatique de la parole et ont rompu avec l'architecture typique des systèmes. La nouvelle direction établie par certains systèmes basés sur les réseaux de neurones connus sous le nom de « bout-en-bout » (*End-to-End neural systems*) ne peut pas être ignorée. Dans ces systèmes, l'objectif est d'associer directement les paramètres acoustiques aux caractères (graphèmes), ce qui suppose une vraie économie

technique, car la représentation intermédiaire en phones est omise (Wang *et al*., 2019) et les modèles de prononciation et les grammaires sont supprimés. Il faut noter qu'une grande quantité de données d'entraînement est tout de même nécessaire pour obtenir des résultats acceptables (Jyothi, 2017). En tout état de cause, ce type d'approches ne fera pas l'objet de notre étude.

2.3.2.3. Les modèles de prononciation ou dictionnaires

Pour procéder à la production d'une séquence de mots à partir d'une séquence de phones, un système typique de reconnaissance de la parole a recours à un modèle de prononciation, aussi nommé dictionnaire. Ce composant fournit l'association entre un nombre fini de mots, représentés sous forme de graphèmes, et leur équivalent, ou équivalents, phonétisés correspondants.

Même s'il existe des dictionnaires qui contiennent uniquement la prononciation phonématique (Adda-Decker & Lamel, 2000) et donc standard de chaque entrée lexicale, de manière générale, ce type de modèles incluent les prononciations alternatives produites par les locuteurs d'une langue qui s'écartent de celle qui est canonique. Ces réalisations sont souvent reliées à l'accent[11] ou à la vitesse de parole du locuteur, de même qu'au contexte phonétique du mot. L'intérêt d'insérer ces variantes dans le modèle de prononciation réside dans le fait que, parfois, la différence spectrale entre les variantes et la forme standard peut être significative, ce qui suppose qu'elle est susceptible de ne pas être reflétée dans les modèles acoustiques. C'est la raison pour laquelle l'inclusion de ces variantes améliore la capacité des modèles acoustiques à les reconnaître (Adda-Decker & Lamel, 2000).

La phonétisation du dictionnaire peut se faire selon différentes conventions de représentation des phones (qui doivent par ailleurs être coïncidentes avec celles obtenues en sortie des modèles acoustiques). La plus célèbre, et celle qui sera d'ailleurs utilisée dans notre étude, est la notation IPA (*International Phonetic Alphabet*), qui vise à normaliser un inventaire phonétique unique et opérationnel pour toutes les langues. Une autre représentation bien connue est celle proposée par le SAMPA (*Speech Assessment Methods Phonetic Alphabet*), la particularité étant de n'utiliser que le jeu de caractères imprimables sur 7 bits de la convention ASCII (*American Standard Code for Information Interchange*), pour répondre à la difficulté d'utiliser l'IPA par des ordinateurs.

Les difficultés liées à un modèle de prononciation ne touchent pas seulement la phonétisation ; tout au contraire, la couverture lexicale se révèle également une question problématique. Tout dictionnaire contient un nombre fini de mots, ce qui peut signifier que certains mots prononcés par un locuteur tout-venant ne fassent pas partie de la liste établie. Ces mots sont appelés des OOV (*Out-of-Vocabulary*) et le but de tout modèle de prononciation est de minimiser leur taux d'occurrence (Adda-Decker & Lamel, 2000). Ce n'est pas une tâche facile, car ils doivent faire face à l'apparition d'acronymes, de

[11] Nous y reviendrons plus en détail dans la section 2.5.

néologismes, de pseudo-mots (tels que *bof, ben*) ou encore de noms propres, dont la transcription peut être difficile si leur origine est étrangère.

De manière générale, les modèles de prononciation ne sont pas issus d'un entraînement de données. Or, il faut noter qu'au fil des dernières années, de nouveaux outils basés sur des réseaux de neurones ont été proposés et permettent d'élargir des dictionnaires déjà existants ou de les créer de zéro. Ainsi, sur la base d'un apprentissage des modèles graphème-phone (ou graphème-phonème), des outils qui atteignent de bons résultats comme *Phonetisaurus G2P* ou *g2p-seq2seq* (issu de CMU Sphinx) sont sortis.

Toutefois, ce n'est pas la seule façon dont la construction de modèles de prononciation s'est concrétisée. Il existe, à titre d'exemple, des modèles basés sur des caractéristiques articulatoires (*articulatory feature based pronunciation models*) qui envisagent de fournir des solutions alternatives à la difficulté touchant la variabilité de prononciation. Ainsi, au lieu de prendre les phones en tant qu'unités de représentation des mots, on prend un ensemble plus abstrait de caractéristiques articulatoires telles que la nasalité, le voisement ou le degré de ouverture des lèvres (Livescu *et al*., 2016), l'objectif étant de modéliser plus finement les variations de prononciation des formes de surface et de garantir une meilleure performance dans les systèmes de reconnaissance centrés sur la parole spontanée (Jyothi, 2017).

2.3.2.4. Les modèles de langue ou grammaires

Le dernier constituant est le modèle de langue ou grammaire, qui vise à modéliser les régularités d'une langue naturelle de façon à prédire la séquence la plus probable de mots lors du décodage (Adda-Decker & Lamel, 2000). Outre son utilisation dans le champ de la reconnaissance de la parole, ces modèles se révèlent des composants fondamentaux dans la traduction automatique, la reconnaissance optique de caractères ou la correction orthographique automatique (Jyothi, 2017). Nous présentons ci-dessous les deux types de modélisation linguistique les plus connus : les modèles stochastiques, plus récents et plus couramment utilisés aujourd'hui, et les modèles à base de connaissance, plus anciens, qui ont servi de référence pour les premiers.

Modèles stochastiques ou probabilistes

Dans le domaine de la reconnaissance automatique de la parole, les modèles de langue les plus répandus sont les modèles appelés *n-grams*, de nature *probabiliste* ou *stochastique*, où les *n-grams* constituent des séquences de symboles (mots, catégories syntaxiques, entre autres) et les modèles correspondants s'utilisent pour prédire chacun des symboles de cette séquence étant donnés les $n - 1$ symboles précédents (Young *et al*., 1995). Cette approche vise à modéliser les contraintes linguistiques (tant d'un point de vue syntaxique que lexico-sémantique) à partir des événements observés dans un corpus d'apprentissage (Estève, 2002). Le succès connu par ces systèmes peut s'expliquer par des raisons techniques, notamment grâce à la simplicité de leur utilisation et à leur bas coût bas de calcul lors du décodage. Mais aussi pour sa grande couverture des phrases

pouvant être exprimées dans une langue. À présent, il existe plusieurs boîtes à outils qui servent à la construction des modèles de langue de type statistique, telles que SRILM (Stolcke, 2004) ou KenLM (Jyothi, 2017).

De façon générale, les modèles de cette nature se fondent sur l'historique des mots du mot courant (*word n-gram models*). Si *n = 2*, le modèle prend en compte le mot précédent (modèle *2-gram*), alors que que si *n = 3*, le modèle prend en compte les deux mots précédents (modèle *3-gram*) et ainsi de suite. Certes, on peut effectivement baser un modèle stochastique sur des symboles représentant des mots, mais il est également possible de considérer à leur place une séquence de classes (*class n-gram models*), supposant que certains mots se comportent d'une manière équivalente. Ces classes peuvent être de différente nature, que ce soit syntaxiques, morphologiques ou même résultants d'une classification automatique.

Ce faisant, nous pouvons aboutir à une modélisation utilisant moins de données d'entraînement, car l'utilisation des classes permet d'établir une généralisation de l'information octroyée par les données et plus important encore : il rend possible la modélisation des séquences qui n'ont pas été observées au niveau de mots (Estève, 2002). Les modèles *n-gram* basés sur des mots pallient ce problème avec l'utilisation des modèles de lissage (un exemple en est le modèle Katz). Ces techniques visent à altérer les probabilités de façon à éviter qu'une séquence non observée dans les données d'apprentissage (et présente dans le corpus de test) soit affectée la probabilité 0 (Jyothi, 2017).

La combinaison des modèles *class n-grams* et *word n-grams* a été testée par interpolation linéaire, tout en obtenant une réduction du nombre d'erreurs dans la transcription. Néanmoins, il convient de signaler que les modèles stochastiques se confrontent à plusieurs difficultés. L'existence d'un volume suffisant de données d'apprentissage en accord avec le type de parole à décoder (Young *et al.*, 1995) n'est pas toujours disponible, ce qui engendre des modèles peu robustes en raison d'une manque d'informations statistiques. Certes, sa couverture est très grande, d'autant plus qu'il accepte toutes les phrases d'une langue, voire les dysfluences typiques à la parole spontanée ; pourtant, il faut noter que sa précision est limitée, dans la mesure où le système accepte aussi des séquences agrammaticales ou pas cohérentes sémantiquement et/ou syntaxiquement avec la langue traitée (Estève, 2002).

Modèles à base de connaissance

En plus des modèles de langue probabilistes, il existe aussi des modèles dits *à base de connaissance*, qui se composent généralement de grammaires formelles et requièrent des experts en linguistique pour les constituer manuellement. Les principes de base trouvent leur origine dans la classification chomskyenne des grammaires ou langages formelles, pour laquelle toute grammaire, *G*, est définie par quatre éléments constituants (ou quadruplet) : $G = (T, N, R, S)$, où :

- *T* représente l'ensemble des symboles non terminaux (c'est-à-dire, les mots possibles des énoncés).
- *N* fait référence à l'ensemble des symboles non terminaux.
- *R* constitue l'ensemble fini des règles de production.
- $S \in N$ représente le symbole de départ.

Le langage généré par une grammaire *G*, noté *L(G)*, est l'ensemble des suites de symboles terminaux qui permet de produire la grammaire (Estève, 2002). Chomsky a établi une hiérarchie entre quatre grands types de grammaires, suivant les contraintes imposées par les règles de dérivation et chacune d'elles s'associant à un type d'automate capable de reconnaître le langage qu'elle produit. Les voici dans un ordre décroissant :

- **Type 0** : *grammaires générales*, associées à la machine de Turing.
- **Type 1** : *grammaires contextuelles*, associées aux automates linéaires bornés.
- **Type 2** : *grammaires hors-contexte*, associées aux automates à pile.
- **Type 3** : *grammaires régulières*, associées aux automates à états finis.

Bien que les grammaires régulières (ou à états finis) soient les grammaires les moins complexes selon cette hiérarchie (Pullum & Gazdar, 1982), leur utilisation est privilégiée dans le domaine de la reconnaissance de la parole, où leur représentation sous forme d'automates à états finis est exploitée pour la constitution de modèles de langues. Ce type d'automates sont constitués d'un certain nombre d'états (ou nœuds) et d'arcs étiquetés où il existe *a minima* un état initial et un état final (la Fig. 2.3 en fournit un exemple).

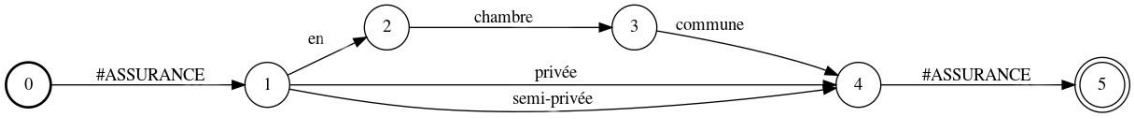

Fig. 2.3 – Exemple d'un automate à états finis.

De cette manière, « une phrase est acceptée par l'automate (et appartient donc au langage engendré par la grammaire associée à cet automate) si l'analyse de cette phrase a produit un chemin dans cet automate commençant sur un état initial et se terminant sur un état final » (Estève, 2002). Il existe parallèlement une variante aux automates à états finis : les transducteurs à états finis, qui fonctionnent à la manière des automates, la différence étant qu'ils produisent des symboles en sortie suivant l'entrée des arcs (la Fig. 2.4 en fournit un exemple).

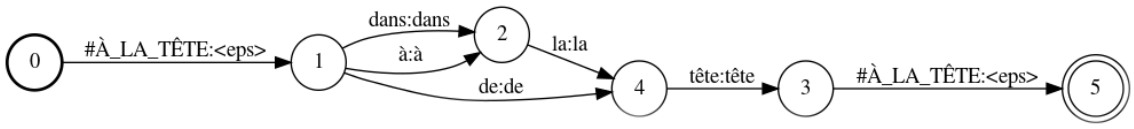

Fig. 2.4 – Exemple d'un transducteur à états finis.

Même si les modèles stochastiques du type *n-gram* s'avèrent performants dans le domaine de la reconnaissance automatique de la parole, l'approche par grammaires

régulières ne peut pas être ignorée. Un avantage de ce type d'approche renvoie au fait de ne pas nécessiter de grands volumes de données d'entraînement, à la différence des modèles stochastiques. En outre, il faut souligner que son emploi favorise l'insertion directe de l'information linguistique (lexicale, sémantique et syntaxique principalement) dans le modèle, et permet d'étendre la distance des contraintes entre les mots, généralement limitée à *n* mots pour un modèle probabiliste *n-gram* (Estève, 2002).

En revanche, ce type d'approche formelle présente également des problèmes lors du décodage. Les grammaires régulières ont beau être très complètes et détaillées, elles ne permettent pas de fournir une couverture intégrale d'une langue[12]. Cela peut s'avérer problématique si le discours à transcrire appartient par exemple à une parole spontanée et contient de nombreuses variations stylistiques, tournures ou dysfluences. Celles-ci ne sont vraisemblablement pas visées par un langage généré avec une grammaire à états finis.

Nous pouvons donc constater qu'il existe une certaine complémentarité entre les deux approches : les avantages des approches probabilistes constituent les inconvénients de celles formelles et vice versa. C'est notamment pour cette raison que des approches mixtes ou interpolées ont été envisagées, dans le but de réunir les points forts de chacune des deux méthodes. Ce faisant, il est donc possible de capturer des événements non rencontrés dans les données d'entraînement tout en gardant une vaste couverture de la langue visée, ce qui favorise ultérieurement l'amélioration de la performance d'un système de reconnaissance de la parole.

2.3.2.5. Le décodeur

Nous avons pu constater que la reconnaissance de la parole envisage de générer automatiquement une séquence des mots étant donné un signal de parole acoustique (Adda-Decker & Lamel, 2000). Par le biais des opérations effectuées par les modèles que nous venons de détailler, il est possible d'atteindre cet objectif, grâce au dernier élément constituant cette structure, le décodeur. Il reçoit en entrée les paramètres acoustiques ainsi que les modèles acoustiques, de prononciation et de langue, et produit en sortie la chaîne de mots la plus probable correspondant au signal de parole reçu, tout en terminant la tâche prévue par le système.

2.4. Mesures d'évaluation

L'évaluation de la qualité des systèmes de reconnaissance automatique de la parole est effectuée sur la base des transcriptions qu'ils produisent. Comme nous l'avons constaté dans la section précédente, la production la plus générale d'un système de reconnaissance vocale sont les mots. D'où la fréquence d'utilisation du WER (*Word*

[12] À savoir, de fournir l'ensemble total de phrases grammaticales d'une langue (Pullum & Gazdar, 1982).

Error Rate), taux d'erreur des mots, en tant que critère de calcul de la performance d'un système de reconnaissance vocale. En supposant que nous avions, par exemple, un taux de 10% de WER, cela signifierait qu'un mot sur dix n'aurait pas été correctement identifié.

Notons qu'au fur et à mesure que le décodage se déroule, le système propose une *hypothèse* de ce que le locuteur a émis comme message. Ainsi, si nous disposons de la transcription humaine fournie pour l'enregistrement correspondant, c'est-à-dire, la *référence*, il sera alors possible de calculer dans quelle mesure la transcription automatique se rapproche ou s'éloigne du message original.

Pour calculer le WER, il est nécessaire que le nombre total d'erreurs soit pris en compte. On en distingue trois types :

– Les *insertions*, abrégées I, qui correspondent à l'ajout d'un mot dans l'hypothèse :

Réf.	est-ce	que	cela	vous	***	convient
Hyp.	est-ce	que	cela	vous	en	convient
Rés.	**C**	**C**	**C**	**C**	**I**	**C**

Tabl. 2.1 – Exemple d'une erreur d'insertion.

– Les *substitutions*, raccourcies S, qui réfèrent aux mots transcrits dans l'hypothèse qui diffèrent de leur équivalent original :

Réf.	est-ce	que	cela	vous	convient
Hyp.	est-ce	que	cela	vous	convainc
Rés.	**C**	**C**	**C**	**C**	**S**

Tabl. 2.2 – Exemple d'une erreur de substitution.

– Les *délétions*, abrégées D[13], qui constituent le processus inverse des insertions : ils correspondent aux mots présents dans la référence et absents dans l'hypothèse.

Réf.	est-ce	que	cela	vous	convient
Hyp.	***	***	cela	vous	convient
Rés.	**D**	**D**	**C**	**C**	**C**

Tabl. 2.3 – Exemple de deux erreurs de délétion.

Comme nous le constatons, tout mot correct est désigné par la lettre C, de sorte que nous disposons de tous les éléments nécessaires pour représenter la formule de calcul correspondante. Le WER est obtenu en divisant le nombre total d'erreurs par le nombre

[13] Le mot « délétion » constitue effectivement un calque lexical de l'anglais. Il serait plus approprié de parler de « suppression » dans ce cas précis. Pourtant, pour une raison de clarté des abréviations, nous allons garder la première désignation.

total de mots contenus dans la référence, N, puis en multipliant le résultat par 100, afin d'obtenir le pourcentage final :

$$\text{WER} = \frac{\mathbf{I + S + D}}{\mathbf{N}} \times \mathbf{100}$$

Cependant, il faut noter que le WER entraîne certaines limites, dans la mesure où elle est globale. Juste en regardant les mots transcrits en sortie, il s'avère difficile de tirer des hypothèses sur quels phones, en fonction de leur contexte, sont moins bien reconnus, par exemple. D'ailleurs, toutes les erreurs générées ont la même importance d'après le WER (Galibert *et al*., 2016) et se trouvent donc sur la même échelle, ce qui ne favorise pas une bonne détection des problèmes les plus cruciaux du système. En outre, il semble compliqué d'éclaircir quels sont les mots qui sont les plus susceptibles d'être confondus les uns avec les autres[14].

Ce sont précisément ces lacunes qui entravent l'identification des éventuelles failles du système utilisé, ce qui a notamment inspiré l'utilisation d'autres métriques d'évaluation telles que le PER. En référence au taux d'erreur du phone ou, dans certains cas, du phonème (*Phone* ou *Phoneme Error Rate*), il se fonde sur les mêmes types d'erreurs que le WER (Moses *et al*., 2016) et fournit une vue plus directe de la prédiction faite par le système lors du processus de décodage.

Il ressort clairement de l'avancée de la technologie que le WER a réussi à être réduit à des taux qui sembleraient autrefois inconcevables. La sophistication progressive des systèmes, liée à l'impact de l'apprentissage profond au cours des dernières années, ont permis de réduire le WER à 5%, tel que certaines expériences menées sur la parole semi-préparée en l'anglais le démontrent (Besacier, 2018).

Même si le WER est la métrique d'évaluation préférée du domaine de la reconnaissance automatique de la parole depuis une trentaine d'années (Galibert *et al*., 2016), une mesure qui la complète et le SER (*Sentence Error Rate*). Elle représente le pourcentage des phrases qui contiennent au moins une erreur. Étant donnée la forte probabilité de trouver une erreur minimale dans un bon nombre de phrases, il en résulte que le taux de SER est généralement plus élevé que celui du WER.

Mais il convient finalement d'indiquer qu'il existe encore d'autres mesures d'évaluation qui sont de plus en plus répandues aujourd'hui, notamment en raison de la montée de la reconnaissance vocale *End-to-End*. Ces approches visent à construire un système avec une seule architecture, qui prend en entrée le signal et produit en sortie une séquence de caractères (Besacier, 2018). Le modèle CPC (*Connectionist Temporal Classification*) en est un exemple, où chaque trame de parole est étiquetée comme étant un symbole particulier, en l'occurrence, un caractère (Graves *et al*., 2006). D'où donc l'existence du CER (*Character Error Rate*) qui se concentre sur l'exactitude de chaque caractère et présente, par là même, un taux généralement inférieur à celui du WER. Quoiqu'utilisé en reconnaissance automatique de la parole, il faut noter que son

[14] L'ajout d'une mesure ensembliste telle qu'une matrice de confusion serait à coup sûr utile à ce propos.

implémentation originale se trouve dans le domaine de la reconnaissance optique des caractères.

2.5. Robustesse d'un système de reconnaissance automatique de la parole

La reconnaissance automatique de la parole humaine s'avère un défi complexe, ce qui influence directement la robustesse des systèmes. Idéalement, un système de reconnaissance automatique de la parole vise à transcrire le discours oral de tout locuteur humain, peu importe le style de parole utilisé et quelles que soient les conditions acoustiques. Malencontreusement, un tel système n'existe pas encore.

2.5.1. Segmentation du signal

Il ne faut pas oublier que les systèmes de reconnaissance vocale automatique sont principalement concernés par la nature éminemment continue du signal de parole. À la différence de son équivalent écrit, le discours oral ne comporte pas de segmentation explicite entre les mots (Besacier, 2018), ni de ponctuation ou de casse. Les mots, à leur place, sont enchaînés les uns après les autres, ce qui en fin de compte rend difficile la bonne identification de chaque unité lexicale par le système.

Il est à noter que l'être humain, recevant la même entrée que la machine, est capable de segmenter instinctivement les mots d'une langue connue. Toutefois, il est à noter que le traitement de la parole effectué par l'individu (en anglais *Human Speech Processing*) est régi par d'autres mécanismes de perception, déjà appréhendés après quelques mois d'exposition linguistique. À titre d'exemple, il est bien connu que les patrons prosodiques sont décisifs pour la discrétisation des unités lexicales en anglais (Cutler & Carter, 1987). La division désormais classique entre syllabes fortes et faibles aide perceptivement le locuteur natif, qui opérera une segmentation lexicale lorsqu'il retrouve l'occurrence d'une syllabe forte suivie d'une syllabe faible (Cutler & Norris, 1988).

En plus de ce problème, il faut ajouter la difficulté liée à la variabilité et à l'instabilité inhérentes à la parole, qui est notoire tant dans la variation du message exprimé que dans le support technique qui le transmet. Ils sont souvent qualifiés de facteurs négatifs qui perturbent le processus de reconnaissance, mais ils n'en restent pas moins des aspects propres, et donc inéluctables, à un acte parole et à son mode de transmission.

2.5.2. Variation intra-locuteur et contextuelle

La variabilité de la parole se reflète, également, sur le plan intra-locuteur. Chaque locuteur a un *idiolecte* (du composé grec *idios*, « propre » et *leksis*, « langage »), à savoir, une façon caractéristique de parler. Même si l'idiolecte est conditionné par des facteurs

physiologiques ou anatomiques (qui peuvent avoir un impact sur la fréquence fondamentale ou le timbre de la voix), il est observé principalement sur le choix du lexique (et la créativité lexicale), l'utilisation de certaines tournures ou dans la prononciation et l'intonation adoptées. Dans la mesure où l'idiolecte constitue un patrimoine exclusif et distinctif de chaque individu, il n'est pas étonnant qu'il s'agisse d'un défi pour la reconnaissance automatique de la voix indépendante du locuteur.

Il faut garder à l'esprit que l'idiolecte de chaque locuteur n'est pas statique, mais plutôt sujet à des variations suivant le contexte communicatif dans lequel il se trouve, ce qui aura pour conséquence qu'il privilégie un style de parole plus soutenu ou plus familier. Il ne faut pas omettre que ce que tous ces échanges partagent (pour autant qu'ils ne correspondent pas à une simple lecture à haute voix), c'est le fait qu'ils sont régis par les normes du discours oral, qui diffèrent notamment de celles du discours écrit. Rappelons à cet égard que parler et écrire constituent deux activités différentes. La dimension phono-auditive est caractéristique du discours conversationnel (Narbona, 1996), mais, plus important encore : le message, tant qu'il a un minimum de spontanéité, est construit au fur et à mesure de son émission, ce qui le différencie du discours écrit, qui résulte d'un processus consciencieux de confection textuelle. Ainsi, des phénomènes tels que des incorrections syntaxiques (Adda-Decker & Lamel, 2000) ou les dysfluences verbales, c'est-à-dire, les hésitations, les répétitions ou les tics de langage sont connaturels à la production du discours parlé et se révèlent des éléments très habituels.

La dimension phonique de la parole entraîne également l'apparition de phénomènes inhérents au discours parlé et qui peuvent également entraver la reconnaissance. Notons, par exemple, que les mots fréquents (souvent des mots grammaticaux) ont tendance à être réalisés d'une façon plus rapide et plus affaiblie que les mots plus rares (Vaissière, 2006). Par ailleurs, les phénomènes reliés à la coarticulation sont incontournables, dans la mesure où les sons sont dépendants de leur contexte. D'ailleurs, l'absence d'une coupure nette entre les phones est à l'origine que, par exemple, dans le mot *structure*, l'arrondissement des lèvres soit déjà présent durant la réalisation du [s], en raison de l'influence du [y]. Ou du fait que dans les mots *quitter* et *cadre*, [ki] soit plus strident que [ka], à cause de la présence d'une voyelle antérieure.

2.5.3. Variation inter-locuteur

Toutefois, outre l'idiolecte de chaque individu, son discours est également associé à une variation inter-locuteur, ce qui signifie qu'il est rattaché à une certaine variété dialectale ou à un accent régional[15]. C'est ainsi qu'il est possible d'inférer l'origine géographique du locuteur, dans la mesure où cette variation se manifeste généralement sous la forme de variations phonétiques de la prononciation, du placement de l'accent

[15] Il y a effectivement plein de littérature écrite à propos de la différence entre dialecte et accent. Cependant, nous allons considérer ces deux termes en tant qu'équivalents, d'autant plus qu'il n'est pas du ressort de cette étude d'élucider la différence entre les deux.

lexical ou des clichés mélodiques (Vieru-Dimulescu *et al*., 2008) qui s'écartent souvent de la réalisation de la langue socialement reconnue comme standard[16].

L'identification perceptive et la caractérisation des géolectes et des accents régionaux a beaucoup intéressé les linguistes, phonéticiens et dialectologues depuis les premières contributions de Jules Gilliéron et son *Atlas linguistique de la France* au début du XXème siècle. Mais il faut noter que son rôle a été moins important dans le domaine de la reconnaissance vocale, où l'on lui a plutôt attribué un effet négatif dans la performance de la tâche de transcription (Vieru-Dimulescu *et al*., 2008). Aujourd'hui, les modèles acoustiques utilisés ont tendance à être entraînés avec de la parole essentiellement standard (Bartkova & Jouvet, 2004), ce qui suppose un taux d'erreur plus élevé lorsqu'on évalue le système avec de la parole moins plus hétérogène. Mais il est essentiel de préciser à ce propos que la variation sociolinguistique est inhérente à toute langue (Tatman, 2017), ce qui veut dire que sa réalisation orale est suspectée, d'un point de vue descriptif, de ne pas être standard.

2.5.4. La problématique des accents

Sur la base de ce qu'une communauté linguistique entend par langue standard (ou variété socialement reconnue comme « non accentuée »), ses locuteurs peuvent percevoir un écart plus ou moins important par rapport à cette norme dans un accent plus ou moins marqué, plus ou moins masqué. Le choix d'une réalisation ou l'autre dépend souvent des questions extralinguistiques liées à la valorisation ou à la stigmatisation de chaque variété, ainsi qu'à des tendances vers la standardisation ou, inversement, vers la revendication de son propre accent régional.

Un tel choix peut effectivement être fait par le locuteur natif d'une langue, mais il n'en va pas de même lorsqu'il s'agit d'un accent étranger, où cette liberté de choix n'est pas aussi « disponible ». Dans ces cas, le locuteur étranger est influencé par l'inventaire phonétique, la structure syllabique ou l'accent lexical de sa langue maternelle (L1) dans la perception et la production de la langue apprise (L2) (Vieru-Dimulescu *et al*., 2011).

C'est ainsi que, dans la langue française, on peut très bien percevoir un accent allemand dans le voisement de la consonne sifflante ([s] → [z]) ou dans l'assourdissement des occlusives sonores ([b] → [p], [d] → [t], [g] → [k]), ou un accent arabe dans l'antériorisation du [e], causée par l'existence d'un système phonologique à trois voyelles (/a/, /i/, /u/) dans l'arabe classique (Vieru-Dimulescu *et al*., 2008). L'accent anglais, de son côté, est observé dans l'articulation apico-alvéolaire du [t], typiquement lamino-dentale en français (Vaissière, 2006), et dans la substitution du [r] grasseyé par le [ɹ] alvéolaire.

[16] La problématique liée à la notion de langue standard est en effet controversée. Certains auteurs ont suggéré qu'elle correspond à la conception d'une langue idéale (Moreno Cabrera, 2000), qui ne trouve pas sa réalisation chez les locuteurs et se rattache donc à un niveau plutôt abstrait. Il convient de noter que, pour des raisons d'extension, il s'agit d'un sujet que nous n'aborderons pas au cours de cette étude.

Pour ce qui concerne la production orale en français des locuteurs natifs de langues romanes, il est courant chez les italiens la postériorisation du [y] en [u], ainsi que l'antériorisation du schwa, [ə], en [e] (Vieru-Dimulescu *et al*., 2008), vu que dans le système phonologique italien, l'arrondissement n'est pas un trait phonématique distinctif. L'espagnol, de son côté, reflète son système phonologique à travers la neutralisation de [v] → [b] ou via l'assourdissement de la consonne sifflante ([z] → [s]). Même si un accent étranger est souvent ressenti à la vue de tout un éventail de facteurs, un seul élément suffit parfois pour arriver à distinguer et identifier un accent précis. Tel est le cas de la fricativisation des occlusives sonores en espagnol ([b] → [β], [d] → [ð], [g] → [ɣ]) (Vasilescu *et al*., 2018), qui est un phénomène allophonique dans des contextes intervocaliques.

Nous pouvons donc constater que, dans les cas où le locuteur est confronté à un son inconnu, il tend à le substituer par celui qu'il juge le plus proche dans son inventaire phonétique natif. Ainsi, le « degré » d'accent du locuteur dépendra, dans une large mesure, de sa capacité à s'adapter et à imiter la prononciation de la langue cible (Bartkova & Jouvet, 2004). Pour autant, cette particulière aptitude n'est pas forcément partagée par tous les apprenants d'une langue ; d'ailleurs, il est très courant qu'un locuteur non natif laisse des traces accentuelles lorsqu'il s'exprime.

C'est pour cette raison que la variation sociolinguistique implique la nécessité que la reconnaissance automatique vocale adapte son architecture de manière à intégrer tout ce recueil d'informations accentuelles et xénophoniques dans ses systèmes. L'intérêt pour ces questions a été exprimé dans les campagnes d'évaluation en reconnaissance des langues comme NIST (*National Institute of Standards and Technology*). De plus, elle s'est traduite par l'utilisation de diverses stratégies telles que l'introduction de modèles acoustiques contenant des accents spécifiques (Tatman, 2017) ou la génération de dictionnaires contenant des variantes de prononciation (Vieru-Dimulescu *et al*., 2011).

Il reste difficile de relever ces défis en raison du manque de disponibilité d'enregistrements contenant du « discours accentué », ce qui est encore plus évident dans le cas des langues peu dotées manquant de ressources numériques comme le swahili (Besacier *et al*., 2012). La prodigalité de la langue standard dans le domaine de la reconnaissance automatique de la parole est responsable d'une certaine asymétrie linguistique. Dans la mesure où les systèmes de reconnaissance visent à réduire autant que possible les erreurs de transcription et vu que leurs modèles acoustiques ont été entraînés avec de la parole standard, ils seront de préférence utilisés pour évaluer la parole qui est similaire aux données utilisées pour l'entraînement (Tatman, 2017), ce qui défavorise les variétés dialectales ou accentuées.

Comme souligné ci-dessus, la variation sociolinguistique (et donc dialectale et/ou accentuelle) est intrinsèque à chaque langue. À l'heure actuelle les sociétés s'orientent vers la création de communautés de plus en plus cosmopolites, ce qui accentue encore la coexistence de différents accents régionaux et étrangers et, par conséquent, la variabilité linguistique. Ce sont notamment des questions qui sont perçues dans la croissance des

phénomènes migratoires, mais tout également dans l'existence de noyaux de population allophone comme celles observées dans la région du Québec ou en Suisse.

2.5.5. Facteurs acoustiques et techniques

Parallèlement, il faut tenir compte des éventuels écueils causés par des facteurs acoustiques et techniques. L'acoustique de la pièce lors de la production orale peut avoir des effets négatifs sur la reconnaissance vocale dans le cas où l'on se trouve dans des environnements bruyants. Il faut mettre en évidence les variations liées au type de capteur (microphone) et à la ligne téléphonique ; que ce soit par téléphone fixe, portable ou via des logiciels de vidéoconférence, le signal de parole est compressé et donc potentiellement dégradé (Besacier, 2018).

Enfin, il convient de noter la nature complexe de la discipline de la reconnaissance automatique de la parole, dans laquelle il s'avère nécessaire de comprendre le processus d'extraction du signal de parole, mais aussi de donner des réponses à des questions telles que « comment est-ce que la parole humaine est perçue et produite ? » ou « quelles sont les rapports entre les sons ? ». Cela explique donc la convergence des domaines tels que le traitement du signal, l'acoustique, la linguistique, la physiologie, la psychologie, la théorie de l'information ou l'informatique (Rabiner & Juang, 1993) sous l'ensemble de la reconnaissance vocale ; mais cela implique, en revanche, une connaissance suffisamment multidisciplinaire des professionnels qui y travaillent, dont la disponibilité peut être limitée.

Nous observons donc que tout cet ensemble de facteurs (liés à la variation intra et inter-locuteur, à la prise en compte des accents ou encore aux difficultés techniques et acoustiques) est susceptible d'altérer, d'une manière ou d'une autre, l'efficacité du processus de transcription de la parole. Cela confère à la capacité d'adaptation à la variabilité (dans toutes ses dimensions) et, par conséquent, à la robustesse, un rôle fondamental dans le domaine de la reconnaissance automatique de la parole.

3. Ressources et méthodologie

Nous présentons ici la méthodologie que nous avons suivie pour la création d'un système de reconnaissance vocale automatique en français pour le domaine médical basé sur des outils libres.

3.1. Présentation du logiciel utilisé

Pour créer notre système de reconnaissance, nous avons eu recours à la boîte à outils Kaldi, qui est un outil *open source* écrit en C++ dédié à la reconnaissance vocale (Povey *et al*., 2011). L'idée de sa création est née en 2009, lors d'un atelier organisé à l'Université John Hopkins ; le projet s'est matérialisé l'année suivante à l'Université des Technologies de Brno. L'objectif était de créer un système facilement utilisable (avec la mise à disposition de recettes complètes) et modifiable (grâce à une licence Apache v2.0), mais aussi de le rendre « autosuffisant » et capable de rompre la dépendance avec d'autres boîtes à outils comme HTK.

Les algorithmes d'entraînement et de décodage acoustique sont basés sur des transducteurs à états finis pondérés (WFST, *Weighted Finite State Transducers*), et leur modification présuppose l'usage de la librairie externe OpenFST. Plus précisément, Kaldi utilise 4 niveaux différents de transducteurs à cet effet (Horndasch *et al*., 2016) :

- Une grammaire ou modèle de langue *G*, qui modélise les probabilités qu'une certaine séquence de mots ait été émise.
- Un dictionnaire ou modèle de prononciation *L*, qui associe les mots, représentés sous forme de graphèmes, avec leur représentation en phones correspondante.
- Un transducteur dépendant du contexte *C*, qui établit une correspondance entre les phones indépendants avec les phones dépendants du contexte.
- Un transducteur HMM nommé *H*, qui associe les phones dépendants du contexte avec des états HMM, qui permettent à Kaldi de déterminiser les transducteurs et d'entraîner les probabilités de transition.

Décodeur	*Transducteur*	*Entrée*	*Sortie*	*Modèles associés*
HCLG	**H**	état HMM	phone dépendant du contexte	Modèles acoustiques
	C	phone dépendant du contexte	phone	Modèles acoustiques
	L	phone	mot	Modèles de prononciation
	G	mot	mot	Modèles de langue

Tabl. 3.1 – Tableau esquissant les 4 niveaux de transducteurs utilisés par Kaldi, avec leurs entrées et sorties correspondantes, ainsi que les modèles qui leur y sont associés. HCLG représente l'union de tous les transducteurs, à savoir, le décodeur.

Plus particulièrement, pour la création de notre système de reconnaissance automatique de la parole appliqué au français, nous avons procédé à la modification et à l'ajustement des composants suivants (comme observé sur le schéma ci-dessous) :

– Modèles acoustiques (transducteurs *H* et *C*). Cela a impliqué un entraînement acoustique basé sur les données décrites dans la section 3.3.1. La description du type de modèles utilisés est détaillée dans 3.2.
– Modèles de langue (transducteur *G*). Nous avons mis en place une modélisation linguistique adaptée au discours médical afin de garantir la robustesse de notre système pour son utilisation prévue. La démarche effectuée pour les générer et le type de méthodes suivies s'expliquent dans la section 3.4.

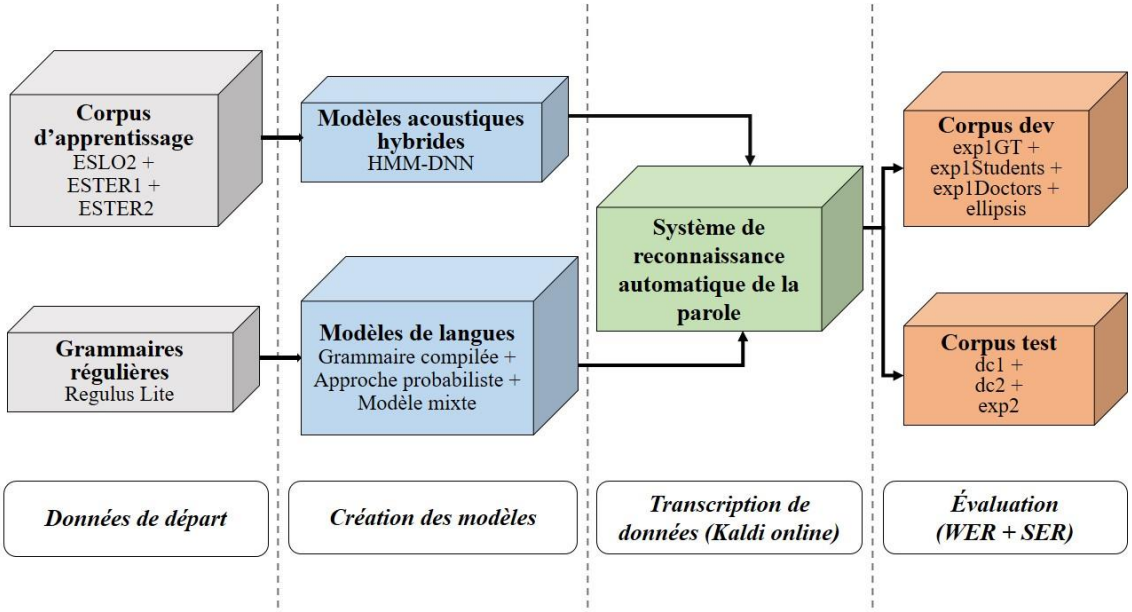

Fig. 3.1 – Schéma montrant les phases que nous avons suivies pour construire et évaluer notre système de reconnaissance automatique de la parole à l'aide de Kaldi.

3.2. Description générale du système HMM-DNN

Kaldi est principalement orienté vers la modélisation acoustique ; or, il s'avère être un outil très flexible du point de vue de la configuration de l'extraction des caractéristiques acoustiques, la création des modèles acoustiques ou le type de décodeur. Pour notre étude, nous avons décidé de travailler avec un système hybride, HMM-DNN, entraîné avec des *lattice-free* MMI (*Maximum Mutual Information*) (Povey *et al*., 2016). Il faut également signaler que :

- La topologie des modèles acoustiques est fondée sur un TDNN (*Time Delay Neural Network*), suivi d'une pile de 16 TDNN factorisés (Povey *et al*., 2018). Cette architecture a été utilisée considérant son bon fonctionnement pour les systèmes hybrides constaté dans l'état de l'art.
- Le vecteur de caractéristiques acoustiques est une concaténation de MFCC à 40 dimensions sans troncature cepstrale (MFCC-40) et d'un i-vecteur à 100 dimensions pour l'adaptation au locuteur (Dehak *et al*., 2010).
- Les échantillons sonores ont été perturbés aléatoirement en vitesse et en amplitude pendant le processus d'entraînement de données (Ko *et al*., 2015).
- Nous avons effectué l'apprentissage des modèles acoustiques en suivant la recette de *tedlium*[17]. Les couches TDNN ont une dimension cachée de 1536 avec une dimension linéaire *bottleneck* de 160 dans les couches factorisées.
- L'extracteur i-vecteur est entraîné à partir de toutes les données acoustiques (que ce soit de la parole perturbée ou de la parole normale) en utilisant une fenêtre glissante de 10 secondes.

Il est à noter que le système de reconnaissance utilisé s'appuie sur « Kaldi online », qui fonctionne en temps réel (configuration en ligne) et produit des hypothèses à la volée. Il a été entraîné avec plus de 350 heures de parole en français issue des corpus ESLO2, ESTER1 et ESTER2.

3.3. Description des corpus

Pour mettre en œuvre notre système de reconnaissance de la parole, nous nous sommes servis de la partition classique en trois corpus audio représentatifs, différents et exclusifs (Le Blouch, 2009), à savoir, sans aucune imbrication entre eux :

- Corpus d'apprentissage.
- Corpus de développement (souvent raccourci comme *corpus de dev*).
- Corpus de test.

[17] Elle est disponible sur : https://github.com/kaldi-asr/kaldi/tree/master/egs/tedlium.

3.3.1. Corpus audio d'apprentissage

Dans la mesure où notre système de reconnaissance automatique de la parole vise à être appliqué dans le contexte d'interaction médecin-patient, il semble important qu'il soit robuste face à la variabilité intra et inter-locuteur. C'est pour cette raison que nous utilisons un ensemble de données représentant une diversité suffisamment couvrante d'accents, registres et aspects sociolinguistiques (comme l'âge ou le sexe du locuteur) pour entraîner nos modèles acoustiques. Ce corpus sonore est issu de trois bases de données pour l'étude de la parole française : ESTER1, ESTER2 et ESLO2.

3.3.1.1. ESTER1

La campagne d'Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophoniques (ESTER) est financée par le Ministère de Recherche dans le cadre de l'appel au projet Technolangue, sous l'impulsion de l'Association Francophone de la Communication Parlée, du Centre d'Expertise Parisien de la Délégation Générale de l'Armement et de l'ELDA (*Evaluations and Language resources Distribution Agency*) (Le Blouch, 2009).

L'objectif de la première campagne d'évaluation ESTER, menée entre 2003 et 2005, était d'évaluer automatiquement les systèmes de transcription automatique pour la langue française dans le cadre des journaux radiophoniques. Pour constituer son corpus d'entraînement, ils ont utilisé des enregistrements issus de différentes stations de radio, effectuées en 1998, 2000, 2003 et 2004. Cette collection de données, qui a été orthographiquement transcrite par des annotateurs humains, rassemble 100 heures de parole (voir le Tabl. 3.2). Elle inclue 2172 locuteurs différents, tout en regroupant des voix des hommes et femmes, enfants ainsi que des locuteurs français non natifs (Galliano *et al.*, 2006).

Corpus	***Parole transcrite***
France Info	12h
France Inter	37h
RFI	27h
RTM	22h
France Culture	1h
Radio Classique	1h
Total	**100h**

Tabl. 3.2 – Corpus audio ESTER1 transcrit manuellement.

3.3.1.2. ESTER2

De son côté, ESTER2 reprend les objectifs de ESTER1, tout en incluant d'autres types de données, notamment la parole accentuée et la parole spontanée. Pour cette

nouvelle campagne, un corpus d'apprentissage d'environ 100 heures d'émissions radiophoniques est distribué (collecté entre 2002 et 2004), ainsi qu'un ensemble de 6h de données (daté 2007), identifiées comme le corpus de développement (Gravier *et al*., 2008). Tous les enregistrements ont été manuellement transcrits et contiennent des émissions d'actualité, parfois dans un registre plus conversationnel (voir le Tabl. 3.3).

Corpus	***Corpus d'apprentissage***	***Corpus de développement***
France Inter	26h	2h
RFI	69h	0h40
Africa nº 1	10h	2h20
TVME (ex RTM)	–	1h
Corpus EPAC	13h	–
Total	**100h**	**6h**

Tabl. 3.3 – Corpus audio ESTER2 transcrit manuellement.

3.3.1.3. ESLO2

Le corpus ESLO2 (Enquête Sociolinguistique à Orléans) a été recueilli par le CORAL (Centre Orléanais de Recherche en Anthropologie et Linguistique) entre 2008 et 2012 (Eshkol-Taravella *et al*., 2011) et s'inscrit dans le projet Variling appartenant à l'Agence Nationale de Recherche (ANR). Il a été créé dans le but de constituer une collection de données qui représente l'ensemble des variations linguistiques qui existent dans le français parlé. Des aspects rapportés à la sociolinguistique, à l'innovation lexique, au registre ainsi qu'aux géolectes (tenant compte du contact entre dialectes régionaux, mais aussi avec d'autres langues) ont été pris en compte (Serpollet *et al*., 2007).

À cette fin, ce corpus recueille des données venant de différentes sources, telles que des entretiens face-à-face, des interviews avec des personnalités, des enregistrements de conversations téléphoniques ou des émissions radiophoniques. ESLO2 vise, à son tour, à créer un corpus comparable à ESLO1, qui a été collecté entre 1968 et 1971 et regroupe plus de 300 heures de parole, pour pouvoir ainsi prendre en compte la variation diachronique en français. Pour notre étude, nous avons notamment utilisé 150 heures de parole transcrite issue du corpus ESLO2, ce qui fait au total plus de 350 heures de parole comme corpus audio d'entraînement (voir Tabl. 3.4).

Corpus	***Parole transcrite***
ESTER1	100h
ESTER2	106h
ESLO2	150h
Total	**356h**

Tabl. 3.4 – Résumé des données utilisées comme corpus d'entraînement.

3.3.2. Corpus audio de développement

Pour vérifier le comportement de notre système de reconnaissance de parole, il est indispensable de se servir d'un corpus de développement. De cette manière, il est possible d'effectuer des ajustements et réglages qui ultérieurement pourront être appliqués sur le corpus final de test. Pour ce faire, nous avons utilisé du corpus sonore venant des expériences effectuées au sein des HUG sur BabelDr visant à évaluer le système de reconnaissance vocale précédent basé sur des technologies Nuance. Cela nous permettra, en outre, de comparer la performance entre ledit système et celui que nous présentons dans ce mémoire.

Il est à noter que le jeu de données rassemble des locuteurs différents (qui varient en âge, sexe), des accents différents (tant régionaux qu'étrangers) ainsi que des conditions d'enregistrements diverses (où l'acoustique de la salle peut être soumise à un entourage bruyant). Toutes les données ont été transcrites orthographiquement par des annotateurs humains. Les sous-ensembles constituant notre corpus de développement sont les suivants (et ils sont tous récapitulés dans le Tabl. 3.5) :

- *exp1Students*, qui comporte des données collectées avec l'ancienne version de BabelDr (utilisant un client Flash[18]), et inclue des dialogues complets échangés entre des étudiants en médecine et des patients standardisés.
- *exp1Doctors,* qui contient aussi des données issues du client Flash et intègre des échanges entre des médecins et des patients standardisés.
- *exp1GT*, qui comprend l'extraction des phrases prononcées par des médecins lors de l'utilisation de Google Translate.
- *ellipsis,* incluant des données artificielles collectées avec une autre plateforme en ligne qui utilise la même technologie que le client BabelDr actuel (Rayner *et al*., 2018). Il contient des paires des phrases, où la première est incomplète (elliptique) alors que la deuxième correspond à son équivalent complet.

Corpus	***Mots***	***Phrases***	***Locuteurs***	***Parole***
exp1Students	3078	573	4	0h26
exp1Doctors	1710	328	6	0h20
exp1GT	1899	321	9	0h16
ellipsis	10071	1642	5	1h13
Total	**16758**	**2864**	**24**	**2h15**

Tabl. 3.5 – Ensemble des données de développement.

[18] La notion de client fait référence au logiciel qui envoie des demandes à un serveur. L'ancien client web de BabelDr avait été créé via Flash.

3.3.3. Corpus audio de test

Pour ce qui concerne le corpus de test, le jeu de données que nous avons choisi correspond aux données collectées avec la version actuelle de BabelDr (qui utilise un client web JavaScript). À la manière du corpus de développement, il existe de la variabilité inter-locuteur et des conditions d'enregistrement différentes. Il est divisé en trois sous-parties :

- *dc1* et *dc2*, qui intègrent des collections de données.
- *exp2*, comprenant des dialogues complets entre des médecins et des patients standardisés.

Corpus	*Mots*	*Phrases*	*Locuteurs*	*Parole*
exp2	5388	928	12	0h57
dc1	2981	622	1	0h58
dc2	6797	1158	1	0h24
Total	**15166**	**2708**	**14**	**2h19**

Tabl. 3.6 – Ensemble des données de test.

3.4. Génération des modèles de langue

3.4.1. Approche basée sur des grammaires formelles

Une conception appropriée des modèles de langue (*G*) est cruciale pour le bon fonctionnement d'un système de reconnaissance de la parole adapté au domaine médical. Dans des situations d'urgence, la rapidité et l'efficacité d'un diagnostic correct peuvent être compromises lorsque le spécialiste ne parle pas la langue du patient qu'il traite. C'est pourquoi les outils de traduction doivent contribuer à lever ces obstacles.

Nous avons déjà indiqué les particularités du discours médical en tant que langue de spécialité dans le cadre de l'échange médecin-patient. Elles sont notoires d'un point de vue lexical, dans la mesure où le médecin fait usage d'une terminologie spécifique durant l'entretien (technicismes, néologismes, acronymes), mais aussi d'un point de vue syntaxique, car la phraséologie typique est visible dans la manière dont le médecin interroge le patient sur les symptômes ou la manière dont il établit le diagnostic (Rouleau, 2007).

Outre ces facteurs, il existe également d'autres contraintes discursives de nature pragmatique, pour lesquelles le spécialiste doit implicitement être suffisamment précis, mais en même temps suffisamment informatif de manière à être bien compris par le patient. Par ailleurs, la dimension d'oralité d'un tel échange suppose l'apparition de phénomènes comme les ellipses lors de la communication, qui sont susceptibles de rendre

plus complexe la tâche de reconnaissance, puisqu'un énoncé correctement reconnu peut facilement être mal interprété (Rayner *et al*., 2018).

Pour injecter de telles informations dans notre système, utiliser un corpus quelconque pour entraîner les modèles de langue ne semble pas une approche adéquate. Il est pourtant souhaitable qu'ils soient représentatifs du type de discours utilisé lors d'un échange médical pour contribuer à la robustesse du système. C'est pourquoi nous avons décidé de nous appuyer sur une modélisation linguistique élaborée par des experts traducteurs et basée sur des grammaires régulières.

Ce type de modélisation, connue sous le nom général de modèles à base de connaissance (évoquée précédemment dans 2.3.2.4), présente l'avantage de ne pas nécessiter de grands volumes de données pour l'entraînement, à la différence des modèles stochastiques. Par ailleurs, ils permettent d'insérer de l'information sémantique et syntaxique directement dans le modèle, tout en étendant la distance des contraintes entre les mots, limitée à un nombre réduit de mots dans des modèles probabilistes *n-gram* (Estève, 2002). En revanche, ce type d'approche ne fournit pas une couverture intégrale de la langue modélisée ; c'est pourquoi, nous nous sommes également servis d'approches alternatives moins contraintes utilisant le langage généré par la grammaire de départ utilisée.

Plus précisément, cette grammaire a été écrite dans le formalisme *Regulus Lite* par des traducteurs professionnels et elle est principalement utilisé pour la construction des systèmes de traduction vocale (Rayner *et al*., 2016). Son contenu est exprimé sous la forme d'un ensemble de phrases modélisées par des grammaires régulières, où nous pouvons distinguer :

– Une *grammaire principale*, G_{MAIN}, composée d'un ensemble fini de phrases qui visent à modéliser le discours médical tenant compte de différentes paraphrases. Chacune de ces phrases est identifiée par le terme `Utterance`[19] :

```
Utterance
Source $avez_vous ( mal | des douleurs ) ( quelque part | à un
endroit )
Target/french avez-vous mal quelque part ?
EndUtterance
```

Fig. 3.2 – Exemple de `Utterance` *contenant la variable* `TrPhrase $avez_vous`.

```
Utterance
Source je suis $$personne ?aujourd'hui
Source c'est moi ?( qui suis ) $$personne ?aujourd'hui
Target/french je suis $$personne
EndUtterance
```

Fig. 3.3 – Exemple de `Utterance` *contenant la variable* `TrLex $$personne`.

[19] Les exemples ont été légèrement modifiés aux fins d'exposition.

– Une *sous-grammaire*, G_{SUB}, représentée par deux types de classes de mots, `TrPhrase` et `TrLex`, qui peuvent être intégrées dans la grammaire principale à travers des symboles non terminaux précédés de « $ » (dans la Fig. 3.6, le symbole non terminal est `$AVEZ_VOUS`). Chaque classe de mots est identifiée par un symbole de désambiguïsation précédé de « # » en entrée et un mot vide (`<eps>`) en sortie (dans la Fig. 3.7, le symbole de désambiguïsation est `#AVEZ_VOUS`).

  - Les classes de type `TrPhrase` représentent des patrons de phrases.

```
TrPhrase $avez_vous
Source ( avez-vous | vous avez )
EndTrPhrase
```

Fig. 3.4 – Exemple d'une variable (`$avez_vous`) *appartenant à la classe de mots du type* `TrPhrase`.

  - Les classes de type `TrLex` représentent des paradigmes lexicaux.

```
TrLex $$personne source="( votre | le | un ) (docteur |
médecin ) ?( de service )" Target/french="le docteur"
```

Fig. 3.5 – Exemple d'une variable (`$$personne`) *appartenant à la classe de mots du type* `TrLex`.

Nous observons que chaque construction combine un ou plusieurs patrons linguistiques `Source` et au maximum une ligne `Target`. Chaque `Target` spécifie la phrase canonique (Rayner *et al*., 2018) reliée à tous les lignes `Source` qui lui y sont associées[20]. Cette phrase sera utilisée par le médecin pour vérifier le résultat du système de reconnaissance vocale. Si elle est approuvée, elle sera utilisée comme source pour la traduction vers la langue cible choisie.

Nous voyons, en effet, que cette modélisation se fait au moyen de grammaires régulières ; cependant, étant donné leur format source, elles ne peuvent pas être directement utilisées par notre système en tant que modèles de langue. Il est donc nécessaire de convertir les grammaires en transducteurs finis et de les compiler ensuite. De cette manière, nous parvenons à les rendre exploitables par Kaldi. Les travaux présentés dans (Horndasch *et al*., 2016) ont fourni une base pour cette démarche[21].

Pour ce faire, nous avons d'abord procédé à une normalisation, qui nous a permis d'uniformiser le contenu et de distinguer la grammaire principale des sous-grammaires[22] :

[20] Les classes de mots du type `TrPhrase` n'ont pas de `Target`, puisque leur phrase canonique correspondante est présente au niveau des `Utterance`.

[21] L'ensemble d'outils que nous avons créés pour mettre en place la conversion des grammaires sources sous un format lisible par Kaldi est disponible sur : https://github.com/lormaechea/Grammar-Tools-ASR.

[22] Pour des raisons de compréhensibilité, nous ne tiendrons compte que de la première `Utterance` (Fig. 3.2) présentée ci-dessus.

```
#0 $AVEZ_VOUS (mal|des douleurs) (quelque part|à un endroit)
```

Fig. 3.6 – Normalisation effectuée pour la grammaire principale, G_{MAIN}.

```
#AVEZ_VOUS (avez-vous|vous avez) #AVEZ_VOUS
```

Fig. 3.7 – Normalisation effectuée pour la sous-grammaire, G_{SUB}.

Une fois que les fichiers normalisés ont été créés, il faut suivre les étapes décrites ci-dessous pour les rendre utilisables dans Kaldi :

1. Il est d'abord nécessaire de convertir chaque `Source` en un transducteur à états finis (FST). Grâce à un automate à pile, nous les avons transformés en FST (Fig. 3.8 et Fig. 3.9) dans un format de texte brut. Ensuite, nous les avons compilés séparément avec l'exécutable `fstcompile`, faisant partie de la librairie OpenFST.

```
0    1    #0           <eps>
1    2    $AVEZ_VOUS   $AVEZ_VOUS
2    3    des          des
2    4    mal          mal
3    4    douleur      douleur
4    5    quelque      quelque
4    6    à            à
5    7    part         part
6    8    un           un
8    7    endroit      endroit
7
```

Fig. 3.8 – Représentation en FST source de G_{MAIN}.

```
0    1    #AVEZ_VOUS   <eps>
1    2    avez-vous    avez-vous
1    3    vous         vous
3    2    avez         avez
2    4    #AVEZ_VOUS   <eps>
4
```

Fig. 3.9 – Représentation en FST source de G_{SUB}.

2. Par la suite, tous les FST compilés appartenant à G_{MAIN} sont récursivement unifiés sous un seul fichier compilé en utilisant `fstunion` (Fig. 3.10). Pour ce qui touche G_{SUB}, nous avons procédé à une unification différenciée de chaque classe de mots, toujours à l'aide de `fstunion` (Fig. 3.11).

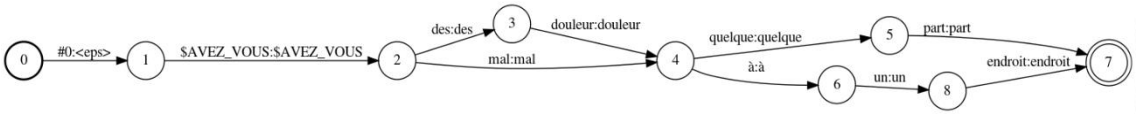

Fig. 3.10 – Résultat de la compilation de G_{MAIN} *avec* `fstunion`.

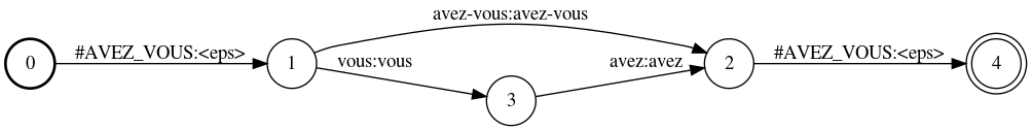

Fig. 3.11 – Résultat de la compilation de G_{SUB} *avec* `fstunion`.

3. Lors de l'intégration des classes de mots (appartenant à G_{SUB}) dans la grammaire principale (G_{MAIN}) il faut remplacer récursivement tous les symboles non terminaux

existants (commençant par « $ ») par les terminaux correspondants utilisant `fstreplace`.

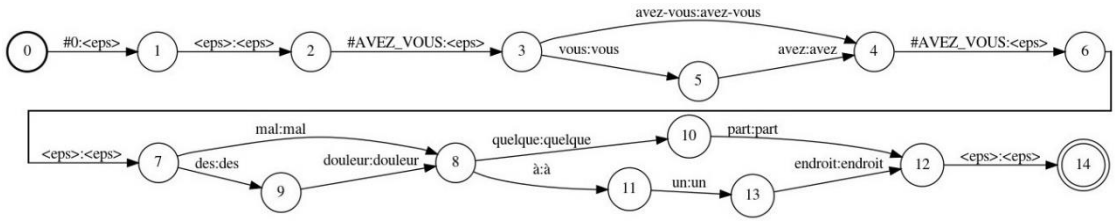

Fig. 3.12 – Jonction de la grammaire principale avec la sous-grammaire, qui résulte de l'opération de `fstreplace`.

Sur cette base, il en résulte un fichier `G.fst` qui contient la grammaire principale, G_{MAIN}, y compris les classes de mots, G_{SUB}. Néanmoins, elle n'est pas encore opérationnelle sur Kaldi. Les prochaines phases doivent également être suivies :

4. Supprimer toutes les transitions vides (`<eps>:<eps>`, pas de symboles en entrée et en sortie) avec `fstrmepsilon`.
5. Déterminiser le FST résultant avec `fstdeterminize`. Cela veut dire que chaque état du FST obtenu aura au maximum une transition d'un certain symbole en entrée ; par ailleurs, il n'y aura aucun état ayant `<eps>` comme symbole en entrée (Mohri *et al*., 2002).
6. Par la suite, minimiser le FST obtenu avec `fstminimizeencoded`, l'un des exécutables OpenFST supplémentaires inclus dans Kaldi. Si `fstminimize` était utilisé à sa place, il pousserait automatiquement les étiquettes au départ lorsque cela est possible, ce qui pourrait entraîner un décalage de la combinaison des étiquettes d'entrée et de sortie dans le FST.
7. Enfin, trier les arcs du transducteur résultant par état avec `fstarcsort` (Fig. 3.13) et continuez avec la procédure habituelle sur Kaldi.

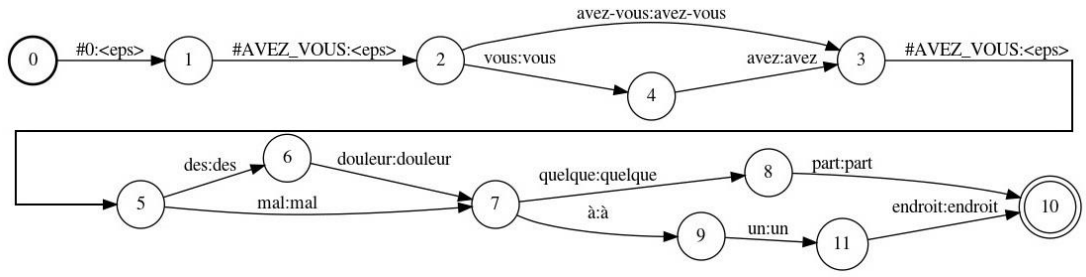

Fig. 3.13 – Résultat de l'implémentation de `fstrmepsilon`, `fstdeterminize`, `fstminimizeencoded` *et* `fstarcsort` *sur la grammaire compilée.*

De cette façon, et en utilisant la bibliothèque externe OpenFST, nous avons pu représenter notre *G* basée sur des grammaires formelles à travers des transducteurs à états finis. Nous vérifierons la performance d'un système de reconnaissance vocale utilisant cette approche dans le chapitre suivant.

3.4.2. Approche basée sur des modèles n-gram

Les modèles à base de connaissances constituent l'une des approches possibles dans le cadre de la modélisation linguistique ; or, il existe des alternatives disponibles. L'une des plus répandues est celle des modèles probabilistes de type *n-gram*. Sa principale différence avec les grammaires formelles renvoie à la création d'un modèle de langue plus « permissif », capable d'accepter un plus large éventail de possibilités comme hypothèses dans le système de reconnaissance de parole. Cela est dû en grand partie à la longueur de contraintes lexico-syntaxiques, puisqu'un modèle statistique ne peut qu'intégrer des contraintes sur de courtes distances (Estève, 2002).

Par ailleurs, les modèles statistiques nécessitent une vaste quantité de données d'apprentissage pour créer des systèmes robustes. Toutefois, la disponibilité de ces corpus est souvent restreinte, d'autant plus s'ils sont reliés à un domaine spécifique. À cet égard, nous avons décidé de créer des modèles probabilistes de type *n-gram* basés sur le langage produit par la grammaire précédente, à savoir, *L(G)*. Vu que les phrases suspectées d'être prononcées sont *a priori* couvertes par la grammaire formelle dans 3.4.1, nous l'avons utilisée pour générer notre corpus d'apprentissage basé sur le discours médical. Pour ce faire, nous avons suivi les étapes suivantes :

1. Tout d'abord, nous avons produit tout le langage généré par la grammaire que nous avions compilée précédemment[23].
2. Cette grammaire génératrice a donné lieu à un ensemble de plus de 4500 millions de phrases, que nous avons ensuite divisé en 10 partitions, pour le rendre plus maniable.
3. Chacune de ces sous-parties a été conséquemment transformée en un modèle de langue 3-gram à l'aide de la boîte à outils SRILM (Stolcke, 2004), le résultat étant des fichiers ARPA. Le lissage repose sur la technique Kneser-Ney (Heafield *et al.*, 2013).
4. Nous avons ensuite fait converger toutes les partitions, ce qui a généré le fichier ARPA définitif représentant les probabilités des séquences des mots suivant le corpus d'apprentissage[24].
5. Finalement, pour le rendre opérationnel sur Kaldi, il suffit de le compiler avec l'utilitaire `arpa2fst`.

Ce faisant, nous obtiendrons un fichier `G.fst` qui pourra désormais être utilisé en tant que modèle de langue probabiliste dans notre système de reconnaissance de la parole.

[23] Pour ce faire, nous avons eu recours à `fstprint-nbest-strings`. Cet exécutable fait partie de la collection `openfst-utils` et vise à effectuer des manipulations sur des FST avec la librairie OpenFST. Il est disponible sur : https://github.com/benob/openfst-utils.

[24] Ceci est possible grâce à l'exécutable `ngram` et l'option `mix-lm`, qui font partie de SRILM.

3.4.3. Approche basée sur des modèles mixtes

La dernière stratégie que nous avons suivie pour créer notre *G* est basée sur l'interpolation de modèles. Les modèles mixtes sont utilisés en tant que technique d'adaptation des modèles linguistiques, et sont généralement obtenus en fusionnant un modèle spécifique avec un modèle généraliste par interpolation linéaire (Estève, 2002). En l'occurrence, nous avons procédé à la jonction de notre modèle probabiliste *n-gram* avec un modèle généraliste entraîné avec 3 milliards de mots venant de plusieurs corpus français (Elloumi, 2019) :

Corpus	*Phrases (M)*	*Tokens (M)*	*Vocabulaire (M)*
EUbookshop	18	432	1,71
Giga	18	57	0,23
Wikipedia	20	502	2
MultiUN	13	404	0,41
OpenSubtitles2016	90	534	0,87
News-Commentary11	30	661	1,43
LeMonde	13	368	1,12
TED2013 + Wit3	0,16	2	0,06
GlobalVoices	0,37	7	0,18
Europarl-v7	2,24	60	0,13
Trames	0,21	0,79	0,03
DGT	3,1	61,7	0,28
Total	**208,08**	**3089,49**	**5,14**

Tabl. 3.7 – Ensemble de corpus utilisés pour la création du modèle de langue généraliste.

Pour effectuer une telle interpolation, la démarche est simple :

1. Tout d'abord, nous récupérons le fichier ARPA que nous avons créé pour la génération de notre modèle *n-gram*, ainsi que l'ARPA représentant notre modèle généraliste. Nous les mélangeons avec la définition de 0,50 comme coefficient d'interpolation linéaire.
2. Ensuite, nous compilons le fichier ARPA résultant en utilisant l'exécutable `arpa2fst`.

Nous obtiendrons ainsi un fichier `G.fst` qui pourra être utilisé comme modèle de langue mixte dans notre système de reconnaissance de la parole.

4. Expériences et résultats

4.1. Introduction aux systèmes de reconnaissance vocale évalués

Pour évaluer la performance du système de reconnaissance vocale que nous avons conçu, nous avons procédé au décodage des corpus de développement et de test que nous avons présentés dans la section précédente (dans 3.3.2 et 3.3.3, respectivement). Rappelons que ce corpus sonore s'inscrit dans le domaine d'échange médical et qu'il contient plusieurs locuteurs et conditions d'enregistrement.

Dans cette section, nous présentons les résultats que nous avons obtenus grâce à la mise en place de nos modèles acoustiques hybrides HMM-DNN, ainsi que de nos modèles de langue adaptés au discours médical (approche basée sur les grammaires formelles, approche probabiliste *n-gram*, approche mixte d'interpolation linéaire). Ainsi, nous avons mis en évidence les taux d'erreur auxquels ils aboutissent, ainsi que les limitations de chacun d'eux.

Toutefois, nous ne nous sommes pas limités exclusivement à une évaluation interne. Nous avons voulu également comparer nos résultats avec ceux des systèmes *speech-to-text* actuellement utilisés par BabelDr dans les HUG. La société qui les fournit, Nuance, a construit deux systèmes de reconnaissance qui sont similaires dans leur modélisation linguistique à ceux que nous avons créés. Il est donc *a priori* judicieux de comparer leurs résultats avec ceux de Kaldi. Les systèmes Nuance sont les suivants :

- Nuance-GLM (*Grammar-based Language Model*), qui, sous une approche formelle, utilise un ensemble de grammaires compilées divisées selon le domaine anatomique (tête, poitrine ou abdomen, entre autres). Notons que ces grammaires utilisent les mêmes fichiers source *Regulus Lite* que nous avons utilisés.
- Nuance-NTE (*Nuance Transcription Engine*), qui a recours à un modèle de langue probabiliste qui combine un modèle généraliste avec un modèle appris sur les données produites par la grammaire précédente. Il trouve son analogue dans notre système à base de modèles mixtes.

En ce qui concerne l'évaluation de nos résultats, nous avons utilisé deux mesures :

- WER (*Word Error Rate*). Cette mesure nous a permis de connaître les performances de nos systèmes en termes de taux d'erreur sur les mots. Nous avons déjà indiqué que le WER est une mesure d'évaluation globale, ce qui suppose que toute erreur a le même poids. L'intérêt de la reconnaissance de la parole appliquée au domaine médical est en effet de produire une transcription cohérente avec la référence. Mais avant tout, il est davantage important que la valeur sémantique véhiculée par la phrase reconnue soit la même que celle que la référence transmettait. Les erreurs liées, par exemple, à l'oralité (telles que les répétitions et les dysfluences) ou celles qui retombent sur les mots outils n'affectent guère le sens de la phrase reconnue. C'est précisément pour cette raison que nous avons procédé à une normalisation préalable de la référence et de l'hypothèse.
- BERT (*Bidirectional Encoder Representations from Transformers*). En raison dudit intérêt sémantique, nous avons également décidé d'effectuer une évaluation basée sur les plongements lexicaux. Rappelons que la phrase reconnue par le système de transcription est ensuite associée à sa phrase canonique correspondante. De sorte qu'il s'avère pertinent de vérifier si, sémantiquement, la phrase reconnue est adaptée à la forme canonique la plus proche. Il est à noter que les résultats de cette expérience ont été calculés par le département TIM de l'Université de Genève sur le corpus de développement[25].

4.2. Résultats issus de la transcription effectuée par un système basé sur une grammaire formelle contrainte (Kaldi-G)

Nous montrons dans cette section les résultats relatifs à la transcription effectuée à l'aide du système utilisant notre grammaire régulière compilée. Pour simplifier sa désignation, nous l'appellerons dorénavant Kaldi-G.

Dans les deux tableaux suivants, nous observons les résultats en termes de WER des corpus de développement (Tabl. 4.1) et de test (Tabl. 4.2). Il est possible de distinguer tant le taux individuel de chaque sous-corpus, que le WER global pondéré suivant le nombre de mots présents dans la référence de chaque sous-corpus. Le nombre de mots reconnus par Kaldi-G, ainsi que celui d'insertions, *I*, délétions, *D*, et substitutions, *S*, est également affiché. Les résultats de Nuance-GLM sont indiqués entre crochets à droite de ceux de Kaldi-G ; ils nous aideront à comparer la performance de chaque système.

[25] Je tiens à remercier Jonathan Mutal de m'avoir transmis les résultats concernant ces expériences.

Kaldi-G [Nuance-GLM]					
Corpus dev	**Mots reconnus**	**I**	**D**	**S**	**WER (%)**
exp1GT	1383/1899	76	151	289	27,17 [34,01]
exp1Doctors	1348/1710	68	106	188	21,17 [30,13]
exp1Students	2409/3078	99	181	389	21,73 [34,67]
ellipsis	7961/10071	304	632	1174	20,95 [31,15]
Total	**13101/16758**	**WER total ponderé (%)**			**21,82 [31,99]**

Tabl. 4.1 – Résultats du décodage de l'ensemble de dev issus du système Kaldi-G. Ceux de Nuance-GLM se trouvent entre crochets.

Kaldi-G [Nuance-GLM]					
Corpus test	**Mots reconnus**	**I**	**D**	**S**	**WER (%)**
dc1	1823/2981	134	486	538	38,85 [61,41]
dc2	5333/6797	174	463	827	21,54 [43,10]
exp2	4229/5388	125	547	487	21,51 [37,28]
Total	**11385/15166**	**WER total ponderé (%)**			**24,93 [45,31]**

Tabl. 4.2 – Résultats du décodage de l'ensemble de test issus du système Kaldi-G. Ceux de Nuance-GLM se trouvent entre crochets.

Comme le montrent les tableaux ci-dessus, notre système Kaldi-G améliore considérablement les résultats dérivés de Nuance-GLM. Dans le cas du corpus de développement, le WER diminue dans chacun des sous-corpus, ce qui se traduit par une réduction globale de plus de 10% en termes de WER. Pour ce qui concerne le corpus de test, nous observons une baisse encore plus importante du taux d'erreur, ce qui suppose une amélioration de plus d'un 20% du WER.

Les hypothèses renvoyées par Kaldi-G révèlent que les phrases appartenant au langage engendré par la grammaire compilée, *L(G)*, sont mieux reconnues que celles qui en diffèrent. D'ailleurs, la plupart des erreurs présentes dans Kaldi-G tendent à être corrélées avec des phrases qui diffèrent à des degrés divers du discours modélisé par *G*. Cela expliquerait pourquoi l'hypothèse s'écarte parfois de ce qui a été énoncé dans la référence, puisque le système vise à s'adapter exactement aux phrases qui peuvent effectivement être produites par *G* (Ormaechea Grijalba *et al*., 2020).

Ainsi, ces phénomènes peuvent se manifester sous la forme de suppressions (Fig. 4.1) ou insertions (Fig. 4.2), où la phrase prononcée contient ou manque d'un mot qui

n'est pas prévu dans le contexte lexico-syntaxique de la phrase la plus probable faisant partie de *L(G)* :

REF	votre	douleur	elle	est	faible
HYP	votre	douleur	***	est	faible
	C	**C**	**D**	**C**	**C**

Fig. 4.1 – Exemple de suppression du mot « elle » dans l'hypothèse, car cet élément est absent dans la grammaire dans le contexte syntaxique donné.

REF	pouvez-vous	***	montrer	où
HYP	pouvez-vous	me	montrer	où
	C	**I**	**C**	**C**

Fig. 4.2 – Exemple d'insertion d'un élément supplémentaire dans l'hypothèse. L'élément « me » est prévu par la grammaire dans le contexte donné, alors que son absence n'est pas prise en compte.

Nous pouvons constater que la présence des erreurs précédentes ne répercute pas sur le contenu sémantique de la phrase source. Toutefois, des phrases en entrée qui ne sont pas incluses dans le langage engendré par *G* peuvent subir des modifications plus évidentes sur la séquence transcrite. Cela peut potentiellement amener à des faibles changements sémantiques (Fig. 4.3, Fig. 4.4 et Fig. 4.5) :

REF	vous	toussez	sec
HYP	vous	toussez	***
	C	**C**	**D**

Fig. 4.3 – Exemple de suppression de l'élément « sec », qui ne fait pas partie de G dans le contexte syntaxique donné, à savoir, « vous toussez ».

REF	avez-vous	des	battements	de	cœur	anormaux
HYP	avez-vous	des	battements	de	cœur	irréguliers
	C	**C**	**C**	**C**	**C**	**S**

Fig. 4.4 – Exemple de substitution de l'élément original « anormaux », non prévu par G dans le contexte donné, par « irréguliers ».

REF	les	démangeaisons	sont-elles	liées	à	votre	alimentation
HYP	les	démangeaisons	sont-elles	liées	à	certains	aliments
	C	**C**	**C**	**C**	**C**	**S**	**S**

Fig. 4.5 – Exemple de substitution de la série « à votre alimentation », absent dans G sachant le contexte gauche donné, par la séquence la plus proche contenue dans G : « à certains aliments ».

Mais elle est susceptible de produire également des séquences où le sens de la référence est modifié (Fig. 4.6 et Fig. 4.7) :

REF	nous	allons	organiser	des	examens
HYP	nous	allons	réaliser	des	examens
	C	**C**	**S**	**C**	**C**

Fig. 4.6 – Exemple II de substitution du mot « organiser », absent dans G dans le contexte gauche et droite donné, par « réaliser », faisant partie de G.

REF	contre	la	fièvre
HYP	contre	la	grippe
	C	**C**	**S**

Fig. 4.7 – Exemple de substitution de l'élément « fièvre » par « grippe », prévu dans le contexte syntaxique donné par G.

Nous observons que le système Kaldi-G s'avère très performant lors du décodage de phrases qui appartiennent au modèle de langue donné. Pourtant, il dévoile en même temps la rigidité d'un système strictement fondé sur des grammaires formelles, où toute différence par rapport au langage reconnu par *G* est susceptible d'entraîner une reconnaissance divergente de ce qui a été initialement émis.

Nous avons déjà mentionné les fortes contraintes lexico-syntaxiques de ce type d'approche (comme expliqué dans 3.4.1), ainsi que leur difficulté (ou impossibilité) d'aboutir à une couverture intégrale de la langue modélisée. Parfois, ce trait peut être périlleux, dans la mesure où il peut ressortir des phrases parfaitement grammaticales, mais non cohérentes avec le sens exprimé dans la référence (les dernières figures en sont la preuve), ce qui peut remettre éventuellement en question la confiance donnée au système. C'est pourquoi nous avons décidé d'assouplir les contraintes de nos modèles linguistiques par le biais d'une approche probabiliste des *n-grams*. Les résultats correspondants sont présentés ci-dessous.

4.3. Résultats issus de la transcription effectuée par un système basé sur un modèle probabiliste *n-gram* (Kaldi-Ngram)

Afin de diminuer les limitations dérivées du système Kaldi-G, nous avons décidé de créer un modèle de langue *n-gram* entraîné sur le langage engendré par *G* (comme indiqué dans 3.4.2), qui est constitué de plus de 4 milliards de phrases. Cela nous permet de réduire la longueur des contraintes lexico-syntaxiques et donc de favoriser une transcription correcte lorsque le locuteur s'éloigne de l'ensemble des paraphrases prévues par la grammaire contrainte. Les résultats obtenus au moyen de ce système, Kaldi-Ngram, sont détaillés ci-dessous.

Dans les tableaux qui suivent, nous présentons les résultats en termes de WER calculés à partir des corpus de développement (Tabl. 4.3) et de test (Tabl. 4.4). Il est possible de distinguer le taux individuel de chaque sous-corpus et le WER global pondéré suivant le nombre de mots présents dans la référence de chaque sous-corpus. Le nombre de mots reconnus par Kaldi-Ngram, ainsi que celui d'insertions, délétions et substitutions est également inclus. Notons qu'il n'existe aucun système analogue Nuance pour cette version.

Kaldi-Ngram					
Corpus dev	**Mots reconnus**	**I**	**D**	**S**	**WER (%)**
exp1GT	1607/1899	47	49	196	15,38
exp1Doctors	1295/1710	103	104	208	24,27
exp1Students	2451/3078	110	148	369	20,37
ellipsis	8805/10071	147	251	868	12,57
Total	**14158/16758**	**WER total ponderé (%)**			**15,51**

Tabl. 4.3 – Résultats du décodage de l'ensemble de dev issus du système Kaldi-Ngram.

Kaldi-Ngram					
Corpus test	**Mots reconnus**	**I**	**D**	**S**	**WER (%)**
dc1	2166/2981	110	319	386	27,34
dc2	5733/6797	250	198	616	15,65
exp2	4510/5388	200	252	426	16,30
Total	**12409/15166**	**WER total ponderé (%)**			**18,18**

Tabl. 4.4 – Résultats du décodage de l'ensemble de test issus du système Kaldi-Ngram.

Nous constatons une diminution du WER de la version « non contrainte » Kaldi-Ngram, par rapport au système Kaldi-G, qui passe de 21,82% à 15,51% pour le corpus dev, et de 24,93% à 18,18% pour le corpus test. Cela se reflète dans la meilleure reconnaissance des phrases partiellement écartées des productions de la grammaire, et qui par là même n'étaient pas correctement transcrites avec le premier système présenté. Avec l'usage d'un système moins rigide, celles-là reçoivent un traitement plus favorable, vu que le sens véhiculé sur la référence est conservé dans l'hypothèse, comme ces exemples en témoignent :

REF	avez-vous	des	battements	de	cœur	anormaux
HYP	avez-vous	des	battements	de	cœur	anormaux
	C	**C**	**C**	**C**	**C**	**C**

Fig. 4.8 – Exemple I ressorti de la reconnaissance effectuée par Kaldi-Ngram.

REF	les	démangeaisons	sont-elles	liées	à	votre	alimentation
HYP	les	démangeaisons	sont-elles	liées	***	votre	alimentation
	C	**C**	**C**	**C**	**D**	**C**	**C**

Fig. 4.9 – Exemple II ressorti de la reconnaissance effectuée par Kaldi-Ngram.

REF	nous	allons	organiser	des	examens
HYP	nous	allons	organiser	des	examens
	C	**C**	**C**	**C**	**C**

Fig. 4.10 – Exemple III ressorti de la reconnaissance effectuée par Kaldi-Ngram.

REF	contre	la	fièvre
HYP	contre	la	fièvre
	C	**C**	**C**

Fig. 4.11 – Exemple IV ressorti de la reconnaissance effectuée par Kaldi-Ngram.

En effet, nous constatons que les phrases éloignées des paraphrases prévues par *G* ont une meilleure reconnaissance avec Kaldi-Ngram, ce qui suppose un taux d'erreur plus faible par rapport à Kaldi-G. Cependant, cette version est également soumise à des limites. D'une certaine manière, l'expansion exponentielle des phrases reconnaissables par le système rend difficile l'association ultérieure de chaque phrase avec sa canonique correspondante par une méthode basée sur des règles. Rappelons que c'est précisément la phrase canonique qui fonctionne en guise de système de vérification de la reconnaissance vocale et qui sert à son tour d'entrée pour la postérieure traduction vers la langue cible. C'est pourquoi, d'autres méthodes d'association basées sur des modèles neuronaux (comme des plongements lexicaux) pourraient être envisagées.

En tout cas, si nous nous tenons strictement à l'étape de la reconnaissance en tant que telle, il est indéniable que la version probabiliste *n-gram* surpasse la version basée sur une approche formelle. Il reste à voir si ces résultats peuvent encore être améliorés sur la base d'un modèle d'interpolation linéaire.

4.4. Résultats issus de la transcription effectuée par un système basé sur un modèle mixte (Kaldi-Mix)

Certes, les résultats obtenus par la version Kaldi-Ngram réduisent substantiellement le taux d'erreur obtenu par la version basée sur une approche strictement formelle. Toutefois, il convient de noter que la couverture de la grammaire, qu'elle soit plus ou moins contrainte, est toujours limitée. C'est pourquoi nous avons décidé de mettre en œuvre un troisième système qui mélange un modèle de langue général avec le modèle spécifique *n-gram* ci-dessus par interpolation linéaire.

Dans les tableaux suivants, nous observons les résultats en termes de WER obtenus des corpus de développement (Tabl. 4.5) et de test (Tabl. 4.6). Nous distinguons le taux individuel de chaque sous-corpus et le WER global. Le nombre de mots reconnus par Kaldi-Mix, ainsi que celui d'erreurs est également affiché. Les résultats de Nuance-NTE, sont indiqués entre crochets à droite de ceux de Kaldi-Mix, de façon à comparer la performance de chaque système.

Kaldi-Mix [Nuance-NTE]					
Corpus dev	**Mots reconnus**	**I**	**D**	**S**	**WER (%)**
exp1GT	1707/1899	23	46	123	10,11 [16,10]
exp1Doctors	1315/1710	74	137	184	23,10 [24,19]
exp1Students	2475/3078	121	161	321	19,59 [15,01]
ellipsis	8889/10071	113	363	706	11,74 [23,21]
Total	**14386/16758**	**WER total ponderé (%)**			**14,15 [20,99]**

Tabl. 4.5 – Résultats du décodage de l'ensemble de dev issus du système Kaldi-Mix. Ceux de Nuance-NTE se trouvent entre crochets.

Kaldi-Mix [Nuance-NTE]					
Corpus test	**Mots reconnus**	**I**	**D**	**S**	**WER (%)**
dc1	2364/2981	88	213	316	20,70 [45,28]
dc2	5959/6797	195	182	461	12,33 [16,62]
exp2	4663/5388	151	262	312	13,46 [15,83]
Total	**12986/15166**	**WER total ponderé (%)**			**14,37 [22,93]**

Tabl. 4.6 – Résultats du décodage de l'ensemble de test issus du système Kaldi-Mix. Ceux de Nuance-NTE se trouvent entre crochets.

Dans les deux cas, nous pouvons constater que le taux d'erreur de Kaldi-Mix a été réduit par rapport au système probabiliste de Nuance-NTE. Si nous examinons plus en détail le corpus de développement, Kaldi parvient à réduire sensiblement le taux d'erreur, à l'exception du jeu de données *exp1Students*, où Nuance-NTE est plus performant. Cependant, le résultat global de Kaldi-Mix estime une réduction de 6,84% du taux d'erreur, avec un résultat tout-à-fait acceptable de 14,15% WER. En ce qui concerne l'ensemble des données de test, l'amélioration est particulièrement significative dans le sous-corpus *dc1*, mais elle est aussi présente dans *dc2* et *exp2*. L'amélioration du taux d'erreur de Kaldi-Mix atteint 8,87% par rapport à Nuance-NTE, ce qui donne un WER de 14,37%. Nous pouvons encore observer que Kaldi-Mix réduit de 4% le taux d'erreur par rapport à Kaldi-Ngram, qui aboutissait à 18,18%.

Par ailleurs, il est intéressant de constater que l'inclusion d'un modèle de langue généraliste n'a pas un impact très élevé sur la réduction du taux de WER dans le corpus de développement. Nous observons, en effet, qu'il n'y a pas une très grande différence entre Kaldi-Ngram, avec 15,51% de WER, et Kaldi-Mix, avec 14,15%. Cela peut être un

indice que le coefficient d'interpolation utilisé n'est pas le plus approprié au contexte de l'énonciation donné et que l'ajustement entre les deux modèles doit se faire utilisant une autre proportion. Or, il se peut tout simplement que les phrases prononcées par les locuteurs correspondent d'une manière assez proche au langage modélisé par la grammaire *G*, et que l'incorporation d'un modèle généraliste soit d'un intérêt plus marginal.

	WER (%)		
Corpus dev	**Kaldi-G**	**Kaldi-Ngram**	**Kaldi-Mix**
exp1GT	27,17	15,38	10,11
exp1Doctors	21,17	24,27	23,10
exp1Students	21,73	20,37	19,59
ellipsis	20,95	12,57	11,74
Total ponderé (%)	**21,82**	**15,51**	**14,15**

Tabl. 4.7 – Résultats du décodage de l'ensemble de dev issus de nos 3 systèmes Kaldi.

	WER (%)		
Corpus test	**Kaldi-G**	**Kaldi-Ngram**	**Kaldi-Mix**
dc1	38,85	27,34	20,70
dc2	21,54	15,65	12,33
exp2	21,51	16,30	13,46
Total ponderé (%)	**24,93**	**18,18**	**14,37**

Tabl. 4.8 – Résultats du décodage de l'ensemble de test issus de nos 3 systèmes Kaldi.

Nous observons au niveau global que les résultats en termes de WER s'améliorent avec la mise en œuvre de modèles stochastiques moins rigides que Kaldi-G, qui s'appuient en même temps sur le langage modélisé par la grammaire régulière. Tant dans le corpus de développement que dans le corpus de test, nous avons observé une amélioration des résultats qui atteint son meilleur taux d'erreur avec Kaldi-Mix : nous passons d'un WER de 23,4% en moyenne avec l'approche d'origine à un 14% de WER avec Kaldi-Mix. En outre, nous avons mis en évidence que nos systèmes *open source* dépassent d'une bonne marge les systèmes Nuance actuellement utilisés dans les HUG, ce qui signifie qu'ils seront bientôt prêts à être déployés en production.

4.5. Résultats issus des expériences menées au sein du TIM avec BERT

Vu que nous cherchons à favoriser que ce qui est énoncé par le locuteur soit correctement identifié par notre système de reconnaissance automatique de la parole, nous avons décidé d'intégrer une évaluation réalisée avec BERT. Il convient de noter que les expériences et résultats à cet effet ont été réalisés au sein du département de Traitement

de l'Information Multilingue (TIM) à l'Université de Genève, de sorte que nous ne disposons pas de tous les détails concernant la démarche suivie.

BERT (*Bidirectional Encoder Representations from Transformers*) est une technique basée sur des réseaux de neurones permettant de pré-entraîner des représentations linguistiques (Devlin *et al*., 2019). Dans ce cas précis, il a été utilisé en tant qu'algorithme de classification basé sur la distance, de manière à déterminer si le sens de la phrase reconnue est équivalent à celui de la phrase canonique la plus proche. Pour ce faire, la procédure a consisté à prendre le modèle CamemBERT, qui est un *Transformer* pré-entraîné en français basé sur RoBERTa (une version optimisée de BERT), et d'en rajouter une couche de classification pour l'adapter aux données provenant des productions de la grammaire *G*. De cette façon, il a été possible de rechercher les phrases canoniques les plus proches pour chaque hypothèse.

Cela a permis d'obtenir une mesure en termes de *Sentence Error Rate* (SER), définie dans ce cas comme le pourcentage de phrases reconnues pour lesquelles la phrase canonique résultante n'est pas identique à la phrase canonique correcte (Mutal *et al*., 2020). Étant donné qu'il s'agit d'une évaluation expérimentale récente, les résultats ici présentés correspondent au corpus de développement. Les données relatives au sous-corpus *ellipsis* n'ont pas été prises en compte, car la technique de classification est différente. Les tableaux ci-dessous présentent les résultats obtenus pour les systèmes Nuance (Tabl. 4.9) et Kaldi (Tabl. 4.10). Notons que le SER global a été pondéré suivant le nombre de phrases présentes dans chaque sous-corpus.

	Nuance-GLM		**Nuance-NTE**	
Corpus dev	**Identiques**	**SER (%)**	**Identiques**	**SER (%)**
exp1GT	202/321	37,07	272/321	15,26
exp1Doctors	224/328	31,71	254/328	22,56
exp1Students	354/573	38,22	481/573	16,06
ellipsis	–		–	
Total pond.	**780/1222**	**36,17**	**1007/1222**	**17,59**

Tabl. 4.9 – Résultats de l'évaluation effectuée avec BERT en termes de SER pour les systèmes Nuance-GLM et Nuance-NTE.

	Kaldi-G		**Kaldi-Ngram**		**Kaldi-Mix**	
Corpus dev	**Identiques**	**SER (%)**	**Identiques**	**SER (%)**	**Identiques**	**SER (%)**
exp1GT	228/321	28,97	277/321	13,71	286/321	10,90
exp1Doctors	254/328	22,56	235/328	28,35	237/328	27,74
exp1Students	430/573	24,96	450/573	21,47	441/573	23,04
ellipsis	–		–		–	
Total pond.	**912/1222**	**25,37**	**962/1222**	**21,28**	**964/1222**	**21,11**

Tabl. 4.10 – Résultats de l'évaluation effectuée avec BERT en termes de SER pour les systèmes Kaldi-G, Kaldi-Ngram et Kaldi-Mix.

Dans les tableaux, nous observons de manière générale que les versions probabilistes de Nuance et Kaldi ont un taux d'erreur plus faible que les versions basées sur la grammaire contrainte compilée (Nuance-GLM et Kaldi-G, respectivement). En tout état de cause, il convient de noter que Kaldi-G réduit le SER de plus de 10% par rapport à son équivalent Nuance-GLM. Cela révèle un comportement similaire en ce qui concerne leurs résultats de WER (où il y avait également un écart de 10% WER), mais surtout une meilleure adaptation et correction sémantique des hypothèses lancées par Kaldi-G par rapport aux phrases canoniques.

Quant aux versions probabilistes, Nuance-NTE aboutit au taux d'erreur le plus faible, avec 17,59% SER. Il faut noter pourtant que ses analogues Kaldi obtiennent aussi des résultats de qualité. Kaldi-Ngram, contenant un modèle *n-gram* basé sur les productions de *G*, obtient 21,28% SER, et Kaldi-Mix, qui mélange lesdits *n-grams* avec un modèle général, obtient un 21,11% SER. Le faible écart entre les deux semble suggérer que l'inclusion d'un modèle linguistique générique n'a pas un impact important sur la correction sémantique des transcriptions renvoyées par le système de reconnaissance vocale. Toutefois, il est probable qu'une configuration différente du coefficient d'interpolation de Kaldi-Mix ait favorisé un SER plus bas. Les prochaines recherches se concentreront notamment sur cette question.

5. Conclusion et perspectives

Dans ce mémoire, nous avons montré qu'il a effectivement été possible de créer un système *open source* de reconnaissance automatique de la parole adapté au discours médical dans un contexte d'urgence. À la lumière des résultats obtenus, nous pouvons constater que les systèmes construits à partir de la boîte à outils Kaldi montrent une amélioration considérable en termes de performance par rapport aux systèmes fournis par Nuance. Nous pouvons donc conclure que les objectifs initialement prévus ont été largement atteints dans le cadre du projet BabelDr.

Afin de créer notre version de reconnaissance vocale basée sur des outils libres, nous avons essentiellement modifié et ajusté deux éléments de son architecture :

- D'une part, nous avons mené à bien un entraînement de modèles acoustiques hybrides (HMM-DNN) avec un corpus d'apprentissage de plus de 350h, regroupant les collections ESLO2, ESTER1 et ESTER2. L'objectif était de regrouper un ensemble de données suffisamment représentatif de la diversité géolectale et sociolinguistique du français, pour ainsi favoriser une robustesse du système face à la variabilité inter-locuteur.
- D'autre part, nous avons implémenté une modélisation linguistique visant à représenter le discours médical à partir d'un ensemble de grammaires régulières élaborées par des traducteurs professionnels. De cette manière, nous avons cherché à encore favoriser la robustesse du système face au contexte dans lequel il doit être utilisé. Sur cette base, nous avons construit trois modèles de langue différents, de plus contraignant du point de vue lexico-syntaxique jusqu'au moins contraignant : une grammaire formelle compilée, une version *n-gram* basée sur le langage engendré par ladite grammaire et un modèle mixte, qui regroupe le modèle *n-gram* avec un modèle générique.

La combinaison de nos modèles acoustiques entraînés avec le corpus sonore indiqué et les différents modèles de langue créés nous a conduit à la conception de trois systèmes de transcription différents : Kaldi-G, qui comprend la grammaire contrainte comme modèle de langue, Kaldi-Ngram, qui correspond à la version *n-gram*, et Kaldi-Mix, qui

contient un modèle de langue interpolée. Cela nous a permis de les comparer avec les systèmes de reconnaissance vocale actuellement utilisés dans le cadre du projet BabelDr : Nuance-GLM (également basé sur une grammaire formelle) et Nuance-NTE (qui correspond à une version probabiliste basée sur des modèles mixtes).

Nous avons voulu tester leur fonctionnement dans un contexte réaliste. C'est pourquoi nous les avons évalués sur la base d'un corpus sonore qui comporte des données issues d'expériences effectuées au sein des HUG. Celles-ci étaient destinées à évaluer les systèmes Nuance, et se caractérisent par une notoire variabilité inter-locuteur (reflétée dans le sexe, l'âge ou la variété des accents des participants) et par des conditions d'enregistrement (affectant l'acoustique de la pièce). Ainsi, il a été possible de réaliser une évaluation comparative non seulement entre nos systèmes Kaldi, mais aussi par rapport à ceux actuellement utilisés par BabelDr.

Les résultats en termes de WER révèlent une meilleure performance de nos systèmes Kaldi par rapport aux technologies boîte noire de Nuance. Kaldi-G réduit considérablement le taux d'erreur de plus de 20% par rapport à son analogue Nuance-GLM, qui atteint 45,31% WER pour le corpus de test. Certes, Kaldi-G obtient un taux d'erreur tout-à-fait acceptable, mais nous avons montré que la qualité des transcriptions s'améliore fortement avec la mise en place de modèles linguistiques moins restrictifs sur le plan lexico-syntaxique. Nous avons observé, en effet, que parfois la transcription de phrases qui ne font pas partie du langage généré par la grammaire peut entraîner des effets inattendus sur le plan sémantique de la phrase reconnue. Cette limitation est notablement surmontée par nos systèmes probabilistes, qui réduisent par ailleurs le WER obtenu par Nuance-NTE. Kaldi-Ngram atteint le 18,18% WER et Kaldi-Mix aboutit aux meilleurs résultats avec 14,37% WER.

De plus, nous avons procédé à une évaluation en termes de SER pour vérifier de manière alternative la correction sémantique des transcriptions renvoyées par nos systèmes[26]. À l'aide de CamemBERT, nous avons observé une amélioration de Kaldi-G, qui a obtenu 25,37% SER, sur Nuance-GLM, avec 36,17% SER. Kaldi-Ngram et Kaldi-Mix améliorent ces résultats avec un 21% SER en moyenne, mais ne réduisent pas le taux d'erreur de 17,59% renvoyé par Nuance-NTE. Cela révèle qu'il est encore possible de perfectionner nos systèmes de reconnaissance automatique de la parole appuyés sur des outils libres.

En bref, nous pouvons constater globalement un meilleur fonctionnement de nos systèmes Kaldi par rapport à ceux fournis par Nuance, ce qui est notamment issu d'une bonne capacité d'adaptation à la diversité de données évaluées. Cependant, cette robustesse face à la variabilité (dans toutes ses dimensions : inter-locuteur, signal ou type de discours modélisé, entre autres) est encore susceptible d'être explorée et peaufinée. L'une des pistes possibles consisterait à modifier et enrichir les modèles de prononciation. Il y a de potentiels utilisateurs de BabelDr qui ne sont pas de langue maternelle française ;

[26] Il convient de noter, en tout cas, qu'il s'agit de résultats expérimentaux réalisés sur le corpus de développement.

il serait donc intéressant d'incorporer des variantes de prononciation permettant de couvrir une diversité d'accents étrangers et ainsi accroître la robustesse et l'accessibilité de notre système pour des locuteurs non natifs. Par ailleurs, le réentraînement de nos modèles acoustiques pourrait également contribuer à cet effet. L'incorporation d'un corpus oral tel que *Common Voice*, qui rassemble 173 heures de parole transcrite en français et 3000 locuteurs différents (Ardila *et al*., 2020), s'avère un objectif souhaitable et atteignable dans le contexte donné.

La modification des modèles interpolés existants (au moyen d'autres coefficients) ou l'incorporation de nouveaux modèles pourraient également être des pistes à explorer. À cet égard, la jonction de la grammaire contrainte avec un modèle de langue générique constituerait un objectif à court terme.

Bien que nous ayons fait une étude qualitative sur les transcriptions renvoyées par les systèmes de Kaldi, nous n'avons pu en faire autant avec les systèmes de Nuance. Une analyse comparative sur les hypothèses des deux systèmes serait également une source intéressante pour découvrir d'autres éléments à améliorer et pour examiner de plus près les différences entre les deux technologies.

Dans une perspective à long terme, nous souhaiterions nous attarder sur le processus d'association entre la phrase reconnue et sa phrase canonique correspondante, qui constitue la phase suivante au processus de transcription. Certes, les résultats obtenus avec CamemBERT au sein du TIM à l'Université de Genève sont un bon préambule ; cependant, nous comptons également générer un modèle associatif entre la phrase transcrite et la phrase canonique au moyen d'un système de règles. En outre, nous envisageons de créer une approche alternative à cet effet basé sur l'entraînement des modèles via FlauBERT. Nous présupposons qu'un tel système serait susceptible de donner de bons résultats étant donnés les 4500 millions de phrases qui font partie du langage généré par notre grammaire régulière compilée.

Bibliographie

Adda-Decker, M., & Lamel, L. (2000). The use of lexica in Automatic Speech Recognition. In F. Van Eynde & D. Gibbon (Éds.), *Lexicon Development for Speech and Language Processing* (p. 235-266). Springer Netherlands. http://link.springer.com/10.1007/978-94-010-9458-0_8

Ahmed, F., Bouillon, P., Gerlach, J., Spechbach, H., & Destefano, C. (2017). *A robust medical Speech-to-speech/Speech-to-sign phraselator*. Interspeech. https://archive-ouverte.unige.ch/unige:96955

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). *Common Voice : A Massively-Multilingual Speech Corpus*. http://arxiv.org/abs/1912.06670

Bartkova, K., & Jouvet, D. (2004). Multiple models for improved speech recognition for non-native speakers. *Proceedings - SPECOM 2004, 9th International Conference Speech and Computer*, 7.

Besacier, L. (2018). *Automatic speech recognition : Introduction, current trends and open problems*. http://lig-membres.imag.fr/blanchon/SitesEns/NLSP/resources/ASR2018.pdf

Besacier, L., Gelas, H., & Pellegrino, F. (2012). Développement de ressources en swahili pour un système de reconnaissance automatique de la parole. *Actes de la conférence conjointe JEP-TALN-RECITAL, 1*, 633-640.

Bouillon, P., Flores, G., Starlander, M., Chatzichrisafis, N., Santaholma, M., Tsourakis, N., Rayner, M., & Hockey, B. A. (2007). A bidirectional grammar-based medical speech translator. *Proceedings of the Workshop on Grammar-Based Approaches to Spoken Language Processing*, 42-48. http://portal.acm.org/citation.cfm?doid=1626333.1626341

Bouillon, P., Gerlach, J., Spechbach, H., Tsourakis, N., & Halimi, S. (2017). BabelDr vs Google Translate : A user study at Geneva University Hospitals (HUG). *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*. https://archive-ouverte.unige.ch/unige:94511/

Bouillon, P., Spechbach, H., Durieux-Paillard, S., Gerlach, J., Halimi, S., Hudelson, P., Rayner, M., Strasly, I., & Tsourakis, N. (2016). BabelDr : A Web Platform for Rapid Construction of Phrasebook-Style Medical Speech Translation Applications. *19th Annual Conference of the European Association for Machine Translation*.

Boujon, V., Bouillon, P., Spechbach, H., Gerlach, J., & Strasly, I. (2018). Can speech-enabled phraselators improve healthcare accessibility ? A case study comparing BabelDr with MediBabble for anamnesis in emergency settings. *Proceedings of the 1st Swiss Conference on Barrier-Free Communication*, 50-65. https://archive-ouverte.unige.ch/unige:105852/

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, *2*(3-4), 133-142.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 113-121.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-End Factor Analysis For Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

Dew, K. N., Turner, A. M., Choi, Y. K., Bosold, A., & Kirchhoff, K. (2018). Development of machine translation technology for assisting health communication : A systematic review. *Journal of Biomedical Informatics*, *85*, 56-67.

Dubois, J., Marcellesi, J.-B., Giacomo, M., & Mével, J.-P. (1994). *Dictionnaire de linguistique et des sciences du langage*. Larousse.

Ehsani, F., Kimzey, J., Zuber, E., Master, D., & Sudre, K. (2008). *Speech to speech translation for nurse patient interaction*. 54-59.

Elloumi, Z. (2019). *Prédiction de performances des systèmes de Reconnaissance Automatique de la Parole*. Laboratoire d'Informatique de Grenoble.

Eshkol-Taravella, I., Baude, O., Maurel, D., Hriba, L., Dugua, C., & Tellier, I. (2011). Un grand corpus oral disponible : Le corpus d'Orléans 1 1968-2012. *Ressources Linguistiques Libres*, *53*(2), 17-46.

Estève, Y. (2002). *Intégration de sources de connaissances pour la modélisation stochastique du langage applique à la parole continue dans un contexte de dialogue oral homme-machine*. Université d'Avignon.

Eurostat Statistics Explained. (2019). *Asylum Statistics*. Eurostat. https://ec.europa.eu/eurostat/statistics-explained/index.php/Asylum_statistics

Galibert, O., Camelin, N., Deléglise, P., & Rosset, S. (2016). Estimation de la qualité d'un système de reconnaissance de la parole pour une tâche de compréhension. *Actes de la conférence conjointe JEP-TALN-RECITAL*, 274-282.

Galliano, S., Geoffrois, É., Gravier, G., Mostefa, D., & Choukri, K. (2006). *Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News*. 139-142.

Graves, A., Fernandez, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist Temporal Classification : Labelling unsegmented sequence data with Recurrent Neural Networks. *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*, 369-376.

Gravier, G., Chaubard, L., Geoffrois, É., & Choukri, K. (2008). *Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophoniques. Plan d'évaluation ESTER 2 Phase 1*.

Hacker, K., Anies, M. E., Folb, B., & Zallman, L. (2015). Barriers to health care for undocumented immigrants : A literature review. *Risk Management and Healthcare Policy*, 175-183.

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable Modified Kneser-Ney Language Model Estimation. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, *2*, 690–696.

Horndasch, A., Kaufhold, C., & Nöth, E. (2016). How to add word classes to the Kaldi Speech Recognition Toolkit. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Éds.), *Text, Speech, and Dialogue* (Vol. 9924, p. 486-494). Springer International Publishing. http://link.springer.com/10.1007/978-3-319-45510-5_56

Huang, X., Baker, J., & Reddy, R. (2014). A historical perspective of speech recognition. *Communications of the ACM*, *57*(1), 94-103.

Huang, X., & Deng, L. (2010). An overview of modern speech recognition. *Handbook of Natural Language Processing*, 339-366.

Hudelson, P. (2019). *Communiquer avec les patients allophones* (Aides linguistiques - Service de médecine de premier recours). Hôpitaux Universitaires de Genève. https://www.hug.ch/sites/interhug/files/structures/medecine_de_premier_recours/Strategies/aides_linguistiques_2019.pdf

Janakiram, A. A., Bouillon, P., Gerlach, J., & Hudelson, P. (2019). Mon patient a de la peine à -communiquer aux urgences. Quels outils sont disponibles ? *Revue Médicale Suisse*.

Jelinek, F. (1998). *Statistical methods for speech recognition*. The MIT Press.

Jyothi, P. (2017). *Automatic Speech Recognition – An Overview*. Microsoft Research India Summer Workshop on Artificial Social Intelligence, Bangalore.

Jyothi, P. (2019). *Introduction to Statistical Speech Recognition*.

Karwacka, W. (2014). *Quality assurance in medical translation*. *21*, 19-34.

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). *Audio Augmentation for Speech Recognition*.

Ku, L., & Flores, G. (2005). Pay now or pay later : Providing interpreter services in health care. *Health Affairs*, *24*(2), 435-444.

Le Blouch, O. (2009). *Décodage acoustico-phonétique et applications à l'indexation audio automatique*. Université de Toulouse III.

Livescu, K., Jyothi, P., & Fosler-Lussier, E. (2016). Articulatory feature-based pronunciation modeling. *Computer Speech & Language*, *36*, 212-232.

Llisterri, J. (2003). Las tecnologías del habla : Entre la ingeniería y la lingüística. *Actas del congreso internacional « La ciencia ante el público. Cultura humanística y desarrollo científico y tecnológico »*, 44-67.

Maas, A. L., Qi, P., Xie, Z., Hannun, A. Y., Lengerich, C. T., Jurafsky, D., & Ng, A. Y. (2017). Building DNN acoustic models for large vocabulary speech recognition. *Computer Speech & Language*, *41*, 195-213.

Miao, Y., Jiang, L., Zhang, H., & Metze, F. (2014). Improvements to speaker adaptive training of deep neural networks. *2014 IEEE Spoken Language Technology Workshop (SLT)*, 165-170. http://ieeexplore.ieee.org/document/7078568/

Mohri, M., Pereira, F., & Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, *16*(1), 69-88.

Moreno Cabrera, J. C. (2000). *La dignidad e igualdad de las lenguas*. Alianza.

Moreno Fernández, F. (1998). *Principios de sociolingüística y sociología del lenguaje*. Ariel.

Moses, D. A., Mesgarani, N., Leonard, M. K., & Chang, E. F. (2016). Neural speech recognition : Continuous phoneme decoding using spatiotemporal representations of human cortical activity. *Journal of Neural Engineering*, *13*(5). https://iopscience.iop.org/article/10.1088/1741-2560/13/5/056004

Muñoz-Miquel, A. (2016). La traducción médica como especialidad académica : Algunos rasgos definitorios. *Hermeneus*, *18*, 235-267.

Mutal, J., Bouillon, P., Gerlach, J., Estrella, P., & Spechbach, H. (2019). Monolingual backtranslation in a medical speech translation system for diagnostic interviews—A NMT approach. *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, *2*, 169-203. https://www.aclweb.org/anthology/W19-6734.pdf

Mutal, J., Gerlach, J., Bouillon, P., & Spechbach, H. (2020). Ellipsis Translation for a Medical Speech to Speech Translation System. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 281–290.

Narbona, A. (1996). Sintaxis y pragmática en el español coloquial. *El español hablado y la cultura oral en España e Hispanoamérica*, 223-246.

Ormaechea Grijalba, L., Gerlach, J., Schwab, D., Bouillon, P., & Lecouteux, B. (2020). *Building and enhancement of an ASR system for emergency medical settings : Towards a better accessibility for allophone and disabled patients*.

Phraselator. (n.d.). [Wikipedia: The Free Encyclopedia]. Phraselator. https://en.wikipedia.org/wiki/Phraselator

Pieraccini, R. (2012). *From Audrey to Siri : Is speech recognition a solved problem?* Mobile Voice Conference, San Francisco. http://www.icsi.berkeley.edu/pubs/speech/audreytosiri12.pdf

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., & Khudanpur, S. (2018). Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. *Interspeech 2018*, 3743-3747.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlıcek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). *The Kaldi Speech Recognition Toolkit*. IEEE 2011.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., & Khudanpur, S. (2016). *Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI*. 2751-2755.

Priebe, S., Sandhu, S., Dias, S., Gaddini, A., Greacen, T., Ioannidis, E., Kluge, U., Krasnik, A., Lamkaddem, M., Lorant, V., Riera, R. P., Sarvary, A., Soares, J. J., Stankunas, M., Straßmayr, C., Wahlbeck, K., Welbel, M., & Bogic, M. (2011). Good practice in health care for migrants : Views and experiences of care professionals in 16 European countries. *BMC Public Health*, *11*(1). https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-11-187

Pullum, G. K., & Gazdar, G. (1982). Natural Languages and Context-Free Languages. *Linguistics and Philosophy*, *4*(4), 471-504.

Rabiner, L., & Juang, B. H. (1986). An introduction to Hidden Markov Models. *IEEE ASSP Magazine*, *3*(1), 4-16.

Rabiner, L., & Juang, B. H. (1993). *Fundamentals of Speech Recognition*. Pearson.

Rabiner, L., & Juang, B. H. (2004). Automatic speech recognition – A brief history of the technology development. *Elsevier Encyclopedia of Language and Linguistics*.

Rayner, M., Armando, A., Bouillon, P., Ebling, S., Gerlach, J., Halimi, S., Strasly, I., & Tsourakis, N. (2016). Helping domain experts build phrasal speech translation systems. In J. F. Quesada, F.-J. Martín Mateos, & T. Lopez-Soto (Éds.), *Future and Emergent Trends in Language Technology* (Vol. 9577, p. 41-52). Springer International Publishing. https://doi.org/10.1007/978-3-319-33500-1_4

Rayner, M., Bouillon, P., Tsourakis, N., Spechbach, H., & Gerlach, J. (2018). Handling Ellipsis in a Spoken Medical Phraselator. In *Statistical Language and Speech Processing* (Vol. 11171, p. 140-152). Springer International Publishing. https://doi.org/10.1007/978-3-030-00810-9_13

Rouleau, M. (2007). La langue médicale : Une langue de spécialité à emprunter le temps d'une traduction. *TTR : traduction, terminologie, rédaction*, *8*(2), 29-49.

Ruiz Costa-jussà, M. R., Farrús, M., José A., F., & José B., M. (2010). Automatic and human evaluation study of a rule-based and a statistical Catalan-Spanish machine translation systems. *Seventh Conference on International Language Resources and Evaluation*, 1706-1711.

Serpollet, N., Bergounioux, G., Chesneau, A., & Walter, R. (2007). *A large reference corpus for spoken French : ESLO 1 and 2 and its variations*.

Soubrier, J. (2011). Enseignement de la traduction médicale : Entre considérations théoriques et modalités pratiques. *Équivalences*, *38*(1), 135-163.

Spechbach, H., Gerlach, J., Mazouri Karker, S., Tsourakis, N., Combescure, C., & Bouillon, P. (2019). A speech-enabled fixed-phrase translator for emergency settings : Crossover study. *JMIR Medical Informatics*, *7*(2). http://medinform.jmir.org/2019/2/e13167/

Stolcke, A. (2004). SRILM – An extensible language modeling toolkit. *Proceedings of the 7th International Conference on Spoken Language Processing*. ICSLP.

Swiss Refugee Council. (2020). *Country Report : Switzerland*. Asylum Information Database. https://www.asylumineurope.org/reports/country/switzerland

Tatman, R. (2017). *Modeling the perceptual learning of novel dialect features*. University of Washington.

Turner, A. M., Choi, Y. K., Dew, K., Tsai, M.-T., Bosold, A. L., Wu, S., Smith, D., & Meischke, H. (2019). Evaluating the Usefulness of Translation Technologies for Emergency Response Communication : A Scenario-Based Study. *JMIR Public Health and Surveillance*, *5*(1). http://publichealth.jmir.org/2019/1/e11171/

Vaissière, J. (2006). *La phonétique*. Presses Universitaires de France.

Vasilescu, I., Hernandez, N., Vieru, B., & Lamel, L. (2018). Exploring Temporal Reduction in Dialectal Spanish : A Large-scale Study of Lenition of Voiced Stops and Coda-s. *Interspeech 2018*, 2728-2732. https://doi.org/10.21437/Interspeech.2018-1256

Vieru-Dimulescu, B., Adda-Decker, M., & de Mareüil, P. B. (2011). Characterisation and identification of non-native French accents. *Speech Communication*, *53*(3), 292-210.

Vieru-Dimulescu, B., de Mareüil, P. B., Adda-Decker, M., & Woehrling, C. (2008). Accents étrangers et régionaux en français. *TAL*, *49*(3), 135-163.

Wang, D., Wang, X., & Lv, S. (2019). An overview of End-to-End Automatic Speech Recognition. *Symmetry*, *11*(8). https://www.mdpi.com/2073-8994/11/8/1018

Young, S., Jansen, J., Odell, J. J., & Woodland, P. C. (1995). *The HTK book*. Cambridge University Engineering Department.

Liste d'abréviations

ANR	*Agence Nationale de Recherche*
ARPA	*Advanced Research Projects Agency*
ASCII	*American Standard Code for Information Interchange*
BERT	*Bidirectional Encoder Representations from Transformers*
CER	*Character Error Rate*
CORAL	*Centre Orléanais de Recherche en Anthropologie et Linguistique*
CPC	*Connectionist Temporal Classification*
DNN	*Deep Neural Networks*
ELDA	*Evaluations and Language resources Distribution Agency*
ESLO	*Enquête Sociolinguistique à Orléans*
ESTER	*Évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques*
fMLLR	*Feature Space Maximum Likelihood Linear Regression*
FST	*Finite State Transducers*
GETALP	*Groupe d'Études en Traitement Automatique de la Langue et la Parole*
GLM	*Grammar-based Language Model (système de reconnaissance vocale Nuance-GLM)*
GMM	*Gaussian Mixture Models*
HMM	*Hidden Markov Models*
HUG	*Hôpitaux Universitaires de Genève*
IPA	*International Phonetic Alphabet*
L1	*Langue maternelle*
L2	*Langue seconde*
LEP	*Limited English Proficiency*
LIG	*Laboratoire d'Informatique de Grenoble*
LPC	*Linear Predictive Coding*
LVCSR	*Large Vocabulary Continuous Speech Recognition*

MFCC	*Mel Frequency Cepstral Coefficients*
MMI	*Maximum Mutual Information*
NIST	*National Institute of Standards and Technology*
NTE	*Nuance Transcription System (système de reconnaissance vocale Nuance-NTE)*
OOV	*Out-Of-Vocabulary*
PER	*Phone (ou Phoneme) Error Rate*
SAMPA	*Speech Assessment Methods Phonetic Alphabet*
SAT	*Speaker Adaptive Training*
SER	*Sentence Error Rate*
S-MINDS	*Speaking Multilingual Interactive Natural Dialog System*
SUR	*Speech Understanding Research*
TDNN	*Time Delay Neural Network*
TIM	*Traitement de l'Information Multilingue*
UGA	*Université Grenoble-Alpes*
UNIGE	*Université de Genève*
WER	*Word Error Rate*
WFST	*Weighted Finite State Transducers*

Index des tableaux

Index des figures

Glossaire

Anamnèse	Entretien effectué par le médecin afin de recueillir des informations sur les antécédents médicaux du patient.
Automates à états finis (ou *Finite State Automata*)	Modèle mathématique qui est représenté avec un certain nombre d'états (ou nœuds) et d'arcs étiquetés où il existe *a minima* un état initial et un état final.
***Character Error Rate* (CER)**	Mesure d'évaluation qui représente le taux de caractères incorrectement reconnus par rapport à un texte de référence.
Décodeur	Dernier élément constituant la structure d'un système de reconnaissance automatique vocale. Il reçoit en entrée les paramètres acoustiques, ainsi que les modèles acoustiques, de prononciation et de langue, et produit en sortie la chaîne de mots la plus probable correspondant au signal de parole donnée en entrée.
Dysfluences	Hésitations, répétitions ou tics de langage qui sont connaturels à la production du discours parlé.
Éléments extralinguistiques	Éléments extrinsèques à la structure de la langue qui influencent son usage.
Éléments géographiques	Éléments relatifs à l'influx de la variation dialectale sur les réalisations linguistiques.
Éléments paralinguistiques	Éléments non verbaux associés à l'élocution (tels que le volume, l'intonation ou le rythme).
Éléments sociolinguistiques	L'influence des aspects sociaux tels que le sexe, l'âge, le niveau d'éducation ou l'origine ethnique sur l'usage de la langue.
Éléments stylistiques	Éléments liés au registre de langue ainsi qu'à la variation qui résulte du contexte situationnel dans lequel le message est exprimé.
***Gaussian Mixture Models* (GMM)**	Technique permettant de modéliser la distribution de probabilités des paramètres acoustiques pour les phones.

Hypothèse	Dans le cadre de la reconnaissance automatique de la parole, transcription effectuée par le système sur un enregistrement sonore.
Idiolecte	Façon de parler caractéristique à un individu.
Internet des objets (ou *Internet-of-Things*)	Ensemble des outils technologiques basés sur l'usage d'assistants intelligents reliés à Internet, dont dispose, au quotidien, un utilisateur humain.
Langue de spécialité	D'après (Dubois *et al*., 1994) : « un sous-système linguistique tel qu'il rassemble les spécificités linguistiques d'un domaine particulier ».
***Mel Frequency Cepstral Coefficients* (MFCC)**	Paramètres acoustiques les plus répandus dans le domaine de la reconnaissance de la parole. Ils cherchent à imiter le comportement auditif humain.
Modèles (ou chaînes) de Markov cachés (HMM)	Modèle qui, suivant un calcul mathématique, permet de déterminer les états cachés (phones) à partir des observations (paramètres acoustiques).
Modèles à base de connaissance	Modèles de langue représentés par des grammaires formelles. Ils requièrent des experts en linguistique pour les constituer manuellement.
Modèles acoustiques	Ils visent à transformer les vecteurs de paramètres acoustiques en une séquence de phones qui soit cohérente avec le message sonore donné en entrée.
Modèles de langue (ou grammaires)	Ils visent à modéliser les régularités d'une langue naturelle de façon à prédire la séquence la plus probable de mots lors du décodage.
Modèles de prononciation (ou dictionnaires)	Ils établissent l'association entre un nombre fini de mots, représentés sous forme de graphèmes, et leur équivalent, ou équivalents phonétiques.
Modèles mixtes	Modèles de langue qui résultent de la fusion ou interpolation de plusieurs modèles (souvent un modèle spécifique et un modelé généraliste).
Modèles probabilistes (ou stochastiques)	Modèles de langues qui visent à modéliser les contraintes linguistiques à partir des événements observés dans un corpus d'apprentissage. Les modèles appelés *n-grams* sont les plus répandus.
Néosémie	Création de nouveaux sens pour des mots déjà existants.

Paramètres acoustiques (ou *acoustic features*)	Éléments qui résultent d'un processus d'échantillonnage du signal de parole. Ils permettent de fournir une représentation discrétisée de la parole.
Phone	Unité phonétique minimale.
Phonème	Unité phonologique minimale.
***Phone* (ou *Phoneme*) *Error Rate* (PER)**	Mesure d'évaluation qui représente le taux de phones (ou phonèmes) incorrectement reconnus par rapport à un texte de référence.
Phonétique	Étude portant sur la réalisation concrète de la parole.
Phonologie	Étude théorique portant sur le système de sons d'une langue.
Phrase canonique (ou *core sentence*)	Dans le cas de BabelDr, résultat de la simplification lexicale, syntaxique et sémantique de la phrase reconnue par le système de reconnaissance automatique de la parole.
Phraselator	Dispositif permettant de fournir une traduction préétablie à un ensemble fini d'énoncés.
Population allophone	Personnes dont leur langue maternelle est une langue étrangère dans la communauté où elles résident.
Post-édition	Révision manuelle d'un texte traduit par un système de traduction automatique.
Prétraitement du signal acoustique	Dans un système de reconnaissance automatique vocale, il réfère à la conversion de l'entrée sonore continue en une représentation discrète, notamment sous la forme d'une séquence de vecteurs des paramètres acoustiques.
Référence	Dans la cadre de la reconnaissance automatique de la parole, transcription manuelle fournie pour un enregistrement sonore.
Rétro-traduction monolingue (ou *backtranslation*)	Dans le domaine de BabelDr, traduction ou conversion de la phrase reconnue vers sa phrase canonique correspondante.
***Sentence Error Rate* (SER)**	Mesure d'évaluation qui représente le pourcentage des phrases qui contiennent au moins une erreur.

Système bout-en-bout (ou *End-to-End system*)	Système de reconnaissance automatique de la parole où l'objectif est d'associer directement les paramètres acoustiques aux caractères (graphèmes).
Système de reconnaissance automatique de la parole (ou *speech-to-text system*)	Technologie permettant d'effectuer une transcription d'un message sonore en mots graphiques.
Système de synthèse vocale (ou *text-to-speech system*)	Technologie permettant de créer de la parole artificielle à partir d'un texte.
Théorie source-filtre	Théorie phonético-acoustique qui part du principe que la production des sons est divisée en deux phases : *source* (transformation du courant d'air en voix par les organes de phonation) et *filtre* (transformation de la voix en parole par les articulateurs).
Traduction automatique (ou *machine translation*)	Traduction d'un texte réalisé par un ordinateur, sans aucune intervention d'un traducteur humain.
Traduction *speech-to-speech* (ou traduction vocale quasi instantanée)	Processus permettant de traduire à haute voix (dans une langue cible) une phrase prononcée par un locuteur.
Transducteur à états finis (ou *Finite State Transducers*)	Modèle mathématique similaire aux automates à états finis, où les arcs étiquetés produisent des symboles en sortie.
***Word Error Rate* (WER)**	Mesure d'évaluation qui représente le taux de mots incorrectement reconnus par rapport à un texte de référence.