Apport des modèles neuronaux de bout-en-bout pour la compréhension automatique de la parole dans l'habitat intelligent

Thierry Desot

Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE ALPES

Spécialité : **Informatique**

Arrêté ministériel : 25 mai 2016

Présentée par

Thierry DESOT

Thèse dirigée par **François PORTET**
et codirigée par **Michel VACHER**

préparée au sein **Laboratoire d'Informatique de Grenoble (LIG)**
et de **Ecole Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Apport des modèles neuronaux de bout-en-bout pour la compréhension automatique de la parole dans l'habitat intelligent

Thèse soutenue publiquement le **11 décembre 2020**,
devant le jury composé de :

M. Yannick ESTÈVE
Professeur des Universités, Université d'Avignon, LIA, Président
M. Jean-François BONASTRE
Professeur des Universités, Université d'Avignon, LIA, Rapporteur
M. Benoît FAVRE
Maître de conférences, Université Aix-Marseille, LIF, Rapporteur
Mme Solange ROSSATO
Maître de Conférences, Université Grenoble Alpes, LIG, Examinateur
M. François PORTET
Professeur des Universités, Université Grenoble Alpes, LIG, Directeur de thèse
M. Michel VACHER
Ingénieur de Recherche CNRS HDR, LIG, Co-Directeur de thèse

Remerciements

Je tiens à remercier tout d'abord mon directeur et co-directeur de thèse pour m'avoir guidé tout au long de ce parcours, de trois ans des travaux de la thèse, avec ses hauts et ses bas, pour leurs directions scientifiques, leur dynamisme et passion, et pour m'avoir intégré dans le projet VocADom et son contexte. Je les remercie également pour leurs efforts de relecture du manuscrit et pour leurs nombreux conseils pendant la rédaction de cette thèse, et les articles liés au contenu de cette thèse.

Je remercie mon directeur de thèse, François Portet. Ses nombreuses idées, conseils, son expérience et expertise, encouragement, énergie, optimisme, et engagement, étaient un vrai carburant et une motivation à faire avancer la thèse.

Je remercie mon co-directeur de thèse, Michel Vacher, pour avoir partagé son expérience et expertise au niveau du smart home, sa gestion du projet VocADom, pour son aide à démarrer la thèse, pour ses conseils, son soutien, notamment pendant l'étape initiale de la thèse qui demandait des efforts à m'adapter de nouveau à la vie d'étudiant.

Je les remercie également pour avoir pu bénéficier des opportunités de présenter des publications liées au contenu de la thèse à quelques conférences, qui étaient des expériences riches et inoubliables.

Je remercie également Laurent Besacier et tout le groupe de recherche GETALP de m'avoir accueilli dans leur équipe.

J'aimerais remercier Jean-François Bonastre et Benoît Favre pour avoir accepté d'être rapporteurs de ma thèse. Je remercie également Solange Rossato et Yannick Estève pour avoir accepté de participer au jury comme examinateurs.

Je remercie Jean-François Bonastre pour avoir transféré l'offre d'emploi pour cette thèse, et pour m'avoir invité à présenter le contenu de mon rapport mi-parcours de thèse pendant un séminaire au LIA.

J'exprime également mes remerciements à Solange Rossato pour ses conseils pendant l'étape de l'annotation semi-automatique multimodale d'une part du corpus VocADom@A4H, et sur l'analyse prosodique de cette thèse.

Un grand merci également aux doctorants, post-doctorants, stagiaires, et ingénieurs de l'équipe GETALP qui ont contribué à une bonne ambiance de travail et des moments de convivialité très agréables.

Je remercie finalement ma famille, notamment ma mère, Roland, mon frère, mon père, Rita et mes meilleurs amis, qui m'ont soutenu durant ma thèse, et m'ont aidé et encouragé pendant la période des premiers mois de la crise Covid-19.

Résumé

Les *enceintes intelligentes* offrent la possibilité d'interagir avec les systèmes informatiques de la maison. Elles permettent d'émettre un éventail de requêtes sur des sujets divers et représentent les premières interfaces vocales disponibles couramment dans les environnements domestiques. Très souvent, la *compréhension des commandes vocales* concerne des énoncés courts et ayant une syntaxe simple dans le domaine des *habitats intelligents* qui sont destinés à favoriser le maintien à domicile des personnes âgées qu'ils assistent dans leur vie quotidienne, améliorant ainsi leur qualité de vie, mais à qui ils peuvent aussi porter assistance en situations de détresse. La conception de tels habitats se concentre surtout sur les aspects de la *sécurité* et du *confort*, visant fréquemment la *détection de l'activité humaine*. L'aspect *communication* est moins pris en compte, c'est pourquoi il existe peu de corpus de parole spécifiques au domaine domotique, en particulier pour des langues autres que l'anglais, alors qu'ils sont essentiels pour développer les systèmes de communication entre l'habitat et ses habitants. La disponibilité de tels corpus, pourrait contribuer au développement d'une génération *d'enceintes intelligentes* qui soient capables d'extraire des commandes vocales plus complexes. Pour contourner une telle contrainte, une partie de notre travail consiste à développer un générateur de corpus produisant des commandes vocales spécifiques au domaine domotique, *automatiquement annotées* d'étiquettes d'intentions et de concepts.

Un système de *compréhension de la parole* (SLU - Spoken Language Understanding) est nécessaire afin d'extraire les intentions et les concepts des commandes vocales avant de les fournir au module de prise de décision en charge de l'exécution des commandes. De manière classique, un module de *compréhension du langage naturel* (NLU - Natural Language Understanding) est précédé par un module de *reconnaissance automatique de la parole* (RAP), convertissant automatiquement la parole en transcriptions. Comme plusieurs études l'ont montré, l'enchaînement entre RAP et NLU dans une approche séquentielle de SLU cumule les erreurs. Par conséquent, l'une des motivations principales de nos travaux est le développement d'un module de SLU de bout en bout (*End-to-End*) visant à extraire les concepts et les intentions directement de la parole. À cette fin, nous élaborons d'abord une approche SLU séquentielle comme approche de référence, dans laquelle une méthode classique de RAP génère des transcriptions qui sont transmises au module NLU, avant de poursuivre par le développement d'un module de SLU de bout en bout. Ces deux systèmes de SLU sont évalués sur un corpus enregistré spécifiquement au domaine de la domotique. Nous étudions si l'information *prosodique*, à laquelle la SLU de bout en bout a accès, contribue à augmenter les performances. Nous comparons aussi la *robustesse* des deux approches lorsqu'elles sont confrontées à un style de parole aux niveaux sémantiques et syntaxiques plus variés.

Cette étude est menée dans le cadre du projet VocADom financé par l'appel à projets génériques de l'ANR.

Mots-clés : Reconnaissance Automatique de la Parole (RAP), Compréhension du langage naturel (NLU), Compréhension de la parole (SLU), habitat intelligent

Abstract

Smart speakers offer the possibility of interacting with smart home systems, and make it possible to issue a range of requests about various subjects. They represent the first ambient voice interfaces that are frequently available in home environments. Very often they are only capable of inferring *voice commands* of a simple syntax in short utterances in the realm of *smart homes* that promote home care for senior adults. They support them during everyday situations by improving their quality of life, and also providing assistance in situations of distress. The design of these smart homes mainly focuses on the *safety* and *comfort* of its dwellers. As a result, these research projects frequently concentrate on *human activity detection*, resulting in a lack of attention for the *communicative* aspects in a smart home design. Consequently, there are insufficient speech corpora, specific to the home automation field, in particular for languages other than English. However the availability of these corpora are crucial for developing interactive communication systems between the smart home and its dwellers. Such corpora at one's disposal could also contribute to the development of a generation of *smart speakers* capable of extracting more complex voice commands. As a consequence, part of our work consisted in developing a corpus generator, producing home automation domain specific voice commands, *automatically annotated* with intent and concept labels.

The extraction of intents and concepts from these commands, by a *Spoken Language Understanding* (SLU) system is necessary to provide the decision-making module with the information, necessary for their execution. In order to react to speech, the *natural language understanding* (NLU) module is typically preceded by an *automatic speech recognition* (ASR) module, automatically converting speech into transcriptions. As several studies have shown, the interaction between ASR and NLU in a sequential SLU approach accumulates errors. Therefore, one of the main motivations of our work is the development of an *end-to-end* SLU module, extracting concepts and intents directly from speech. To achieve this goal, we first develop a sequential SLU approach as our baseline approach, in which a classic ASR method generates transcriptions that are passed to the NLU module, before continuing with the development of an End-to-end SLU module. These two SLU systems were evaluated on a corpus recorded in the home automation domain. We investigate whether the prosodic information that the end-to-end SLU system has access to, contributes to SLU performance. We position the two approaches also by comparing their *robustness*, facing speech with more semantic and syntactic variation.

The context of this thesis is the ANR VocADom project.

Keywords : Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), Spoken Language Understanding (SLU), smart-home

Table des matières

Table des figures

Liste des tableaux

CHAPITRE 1

Introduction

1.1 Motivation

Au cours de ces dernières années, les *enceintes intelligentes* ont connu une grande diffusion dans les foyers du monde entier, nous pouvons citer *Amazon Echo*, ou *Google Home* qui sont les plus connues. Elles offrent à leurs utilisateurs un moyen d'interagir par la voix avec les systèmes informatiques de la maison, sans qu'ils aient besoin de toucher un appareil. Elles permettent aux utilisateurs d'émettre un large éventail de requêtes concernant des sujets divers et représentent les premières interfaces vocales ambiantes qui sont facilement disponibles dans les environnements domestiques. Malgré leur grande diffusion, il y a une absence d'information sur la façon dont leurs architectures sont conçues. On doit se contenter de comparaisons entre les performances obtenues par les différents systèmes (López et coll., 2017), d'études sur les principaux types d'utilisateurs et sur les commandes les plus fréquemment utilisées etc. (Purington et coll., 2017). Il s'agit souvent de *compréhension de commandes vocales* concernant des thèmes courants concernant la météo, les itinéraires de voyage. Les énoncés ne dépassent pas en général une longueur moyenne de 5 mots (Google Home). En outre, seulement 25% de toutes les commandes contiendrait plus que deux mots et moins de 25% des commandes dépasserait une longueur de cinq mots (Bentley et coll., 2018).

Cependant, la compréhension des commandes vocales concerne aussi les dispositifs technologiques des *habitats intelligents* destinés à favoriser le maintien à domicile des personnes âgées. Le vieillissement rapide de la population des pays industrialisés, entraîne une augmentation importante du nombre de personnes dépendantes et va s'accroître durant les années prochaines. Il s'agirait de 1,2 millions en 2040 rien que pour la France (Duée et Rebillard, 2006). Par conséquent, le nombre de personnes en perte d'autonomie va considérablement augmenter et l'Assistance à l'Autonomie à Domicile (AAD ou AAL - *Ambient Assisted Living*), est vue comme une solution viable à ce problème (Chan et coll., 2008). L'objectif des maisons intelligentes est d'assister leurs habitants dans les situations du quotidien en améliorant leur qualité de vie (Vacher et coll., 2015), en apportant aussi de l'assistance en situations de détresse. Il n'est donc pas surprenant que la conception de ces maisons intelligentes (chapitre 2, section 2.1) se concentre surtout sur les aspects de la *sécurité*, du *confort* et de *l'ubiquité*. En conséquence, les projets de recherche les concernant se concentrent fréquemment sur la *détection de l'activité humaine*. Bien que l'aspect *communication* fasse éga-

lement partie de plusieurs définitions d'un habitat intelligent, il s'avère qu'il est moins pris en compte.

L'état de l'art des maisons et des enceintes intelligentes montre qu'il y a un manque de corpus de parole spécifiques au domaine de la domotique, en particulier pour d'autres langues que l'anglais (chapitre 2, section 2.7). Cependant ces ensembles de données sont essentiels pour développer des systèmes de communication entre une maison intelligente et ses habitants. De tels corpus open-source disponibles, pourraient aussi contribuer au développement d'une génération *d'enceintes intelligentes* qui pourraient extraire des commandes vocales plus complexes, moins fréquentes et sur un éventail de thèmes beaucoup plus large, que les sujets les plus répandus actuellement.

Il n'existe que quelques projets français qui se concentrent sur l'aspect de la *communication* et sur la reconnaissance *des commandes vocales* dans un contexte d'AVQ (activités de la vie quotidienne). Cependant, ces projets qui ont été mis en place dans les maisons intelligentes, appliquent des méthodes d'extraction d'intention basées sur des règles, ciblant un vocabulaire de petite taille ayant peu de variation linguistique (chapitre 2, section 2.5). Il s'agit par exemple du projet DESDHIS qui a conduit au développement du corpus *HIS* (Fleury et coll., 2010b) dans le HIS de TIMC-IMAG, et du projet *SWEET-HOME* qui a permis l'enregistrement du corpus *SWEET-HOME* (Vacher et coll., 2014) dans la maison intelligente DOMUS.

Cette pénurie de corpus spécifiques au domaine de la domotique, et le manque de systèmes de SLU pour la langue française capables d'interpréter des commandes vocales domotiques d'une variation linguistique riche, ont mis en évidence le défi que constitue le développement d'un système ciblant la reconnaissance de *commandes vocales* spécifiques au domaine de la *domotique*, et tenant compte de la *variation syntaxique et sémantique* de ses utilisateurs cibles.

1.2 Le projet VocADom

Le sujet de cette thèse relatif à la compréhension de commandes vocales dans un contexte domotique se déroule dans le cadre du projet ANR VocADom [1] (Vacher et coll., 2018). Ce projet vise en première étape à définir, en collaboration avec des utilisateurs finaux, les caractéristiques d'un système de commande vocale de la domotique. Dans une deuxième étape, le système retenu doit être capable de s'adapter à l'utilisateur dans des conditions réelles, en présence de bruit et de plusieurs personnes. Ce système sera ensuite mis en œuvre et évalué dans un habitat intelligent et au domicile de quelques personnes volontaires. Pour atteindre ses objectifs, il devra analyser le signal sonore capté par 16 microphones (4 antennes de 4 microphones), à raison de une antenne par pièce. L'organisation du système est de type séquentiel :

— Le *rehaussement de la parole* permet d'atténuer les signaux parasites tels que la radio ou la télévision ;

1. http://vocadom.imag.fr/

— La *détection du mot-clé* sur une antenne va initier les traitements nécessaires d'une commande vocale ;

— La *localisation du locuteur* sera utilisée pour limiter le traitement à l'antenne concernée et servira ensuite à déterminer le contexte dans lequel la commande vocale a été prononcée ;

— La *suppression de bruit* est réalisée après avoir déterminé la position du locuteur dans la pièce ;

— La *détection d'activité vocale* (VAD - Voice Activity Detection) permet de limiter l'analyse à la partie du signal contenant de la parole et de savoir si c'est l'utilisateur principal du logement qui parle ;

— **La reconnaissance automatique de la parole (RAP)** : permet de transcrire la parole prononcée sous forme de texte ;

— **La compréhension du langage naturel (NLU)** : permet d'identifier l'intention et les attributs de la commande ;

— La *prise de décision* en contexte va émettre des ordres au système domotique.

FIGURE 1.1 – VocADom - quelques exemples de commandes vocales de la domotique

Pour atteindre ces objectifs, ce projet rassemble des chercheurs et des ingénieurs du Laboratoire d'informatique de Grenoble (LIG) (spécialisé en traitement de la parole, compréhension du langage naturel, conception de maison intelligente et évaluation expérimentale), Inria Nancy (spécialisé dans le traitement de la parole et la suppression du bruit), le laboratoire GRePS (spécialisé en psychologie sociale) et une société, THEORIS (système temps réel, développement et intégration). VocADom disposera d'une maison intelligente à la *Maison de la Création et de l'Innovation*, (MACI [2]) de l'UGA afin de valider la preuve de concept expérimentalement.

2. https://www.univ-grenoble-alpes.fr/universite/ambition-et-valeurs/les-projets-emblematiques/la-maison-de-la-creation-et-de-l-innovation

1.3 Objectifs et contributions de ce travail de thèse

La cible de recherche de cette thèse est le développement d'un système de reconnaissance des *commandes vocales* spécifiques au domaine *domotique*, qui tienne compte de la *variation linguistique*, notamment au niveau *sémantique* et *syntaxique* de ses utilisateurs. À cette fin, nous avons choisi une *approche de compréhension de la parole de bout en bout* (SLU E2E – *End-to-end Spoken Language Understanding*). Nous pensons que nous pouvons exploiter l'accès d'une approche SLU E2E au niveau *acoustique* pour prédire les concepts et les intentions. L'étude de Serdyuk et coll. (2018) (chapitre 4, section 4.2.1) extrait les *intentions* directement des paramètres acoustiques MFCC. Cette recherche nous a incités à préférer une approche SLU de bout en bout à un modèle classique de SLU séquentielle qui constituera notre approche de référence.

En extrayant les intentions et les concepts du signal directement, nous essaierons d'éviter la *cascade d'erreurs* introduites par l'interaction entre les modèles de RAP et de NLU introduite par une méthode séquentielle. L'écueil principal de cette dernière approche est souvent de trouver une méthodologie permettant de réduire l'écart entre les composants RAP et NLU comme nous le montre l'état de l'art (chapitre 4). En outre l'étude de Ghannay et coll. (2018) met en évidence que des transcriptions de RAP parfaites *ne* sont *pas* nécessaires pour prédire les intentions et les concepts.

Vu qu'il y a un *manque* de corpus français spécifiques au domaine de la domotique, ou d'ensembles de données d'une taille suffisamment large, une part importante de ce travail portera sur la constitution de corpus dont nous aurons besoin pour créer nos modèles de SLU :

— L'enregistrement d'un ensemble de données de test, la transcription des énoncés et l'annotation au niveau sémantique de concepts et d'intentions.

— La génération de suffisamment de données artificielles pour entraîner des modèles de NLU et sa conversion en *parole synthétique* pour entraîner des modèles de SLU.

Pour faire face à la *distance* entre des données d'entraînement artificielles d'une part, et des données de test de parole réelle d'autre part, nous utiliserons une approche *d'apprentissage par transfert* pour mieux exploiter un Modèle Acoustique (MA) appris sur des données de parole réelle, augmentée de parole artificielle. Les études de Caubrière et coll. (2019) et Lugosch et coll. (2019) ont en effet montré des résultats prometteurs en effectuant un apprentissage par transfert. Les résultats de Li et coll. (2018) et les travaux de Lugosch et coll. (2020) confortent également cette stratégie d'augmentation de données d'apprentissage de parole réelle par des données de parole artificielle.

Un objectif essentiel est le *positionnement* de l'approche cible SLU E2E de bout en bout par rapport à l'approche SLU séquentielle. Nous vérifions si l'information prosodique à laquelle la SLU E2E a accès constitue un atout et contribue aux performances de la SLU. En outre nous comparerons la *robustesse* des deux approches dans une situation où le décalage entre le style de parole des utilisateurs cibles et celui des modèles entraînés s'accroît.

Enfin, les modules de SLU, intégrés dans la chaîne complète du système de commande vocale VocADom, seront testés en milieu réaliste. Des tests avec des volontaires pour ces expérimentations à domicile et dans l'appartement intelligent MACI seront effectués.

Suite à cette introduction, ce manuscrit comporte 9 autres chapitres dont trois consacrés à l'état de l'art.

Un état de l'art des *habitats intelligents* et des *corpus* correspondants sera présenté lors du **chapitre 2**. Dans ce chapitre nous allons donner un aperçu des définitions les plus communément admises de l'habitat intelligent, afin d'en établir une qui soit suffisamment générale pour englober les maisons intelligentes et les *enceintes intelligentes* apparues plus récemment. Cette définition sera suivie d'un aperçu d'habitats intelligents académiques et industriels représentatifs, ainsi que par un aperçu des enceintes intelligentes les plus répandues. Nous présentons ensuite deux *projets* ayant mis en œuvre l'interaction par *commande vocale* dans un habitat intelligent, ces deux projets ont été une source d'inspiration importante pour la réalisation de cette thèse. L'étude de l'état de l'art des *corpus* spécifiques au domaine *domotique* en français et en langues autre que le français, nous permettra de tirer des enseignements utiles pour construire à la fois notre ensemble de test réaliste et l'ensemble d'apprentissage artificiel.

Le **chapitre 3** sera consacré à l'état de l'art de la *RAP* et de la *NLU*, car notre approche *SLU séquentielle* de référence utilise un module de RAP et un module de NLU. Nous comparerons des approches de RAP classiques statistiques avec des approches de réseaux de neurones profonds DNN pour lesquels nous nous concentrerons sur les approches de bout en bout. L'état de l'art consacré à la NLU détaillera l'évolution des premières approches basées sur règles jusqu'aux modèles de RNN multi-tâches les plus récents.

Au cours du **chapitre 4**, nous vérifierons, au travers de l'état de l'art des approches SLU, comment les approches *séquentielles* traitent la *dépendance* entre le module de NLU et les transcriptions issues du module de RAP. Nous essaierons également de comprendre les avantages et désavantages d'une approche *E2E* de bout en bout qui tente de combiner la RAP et la NLU en une *tache unique*, en ayant accès au deux niveaux acoustique et prosodique.

Le **chapitre 5** présente les questions de recherche et la méthode suivie durant cette thèse pour tenter d'y répondre. Il donnera les motivations qui ont conduit à la collecte du *corpus VocADom@A4H* de parole *réelle* et à la création d'un générateur de données de parole *artificielle*. Les méthodes pour construire notre approche SLU séquentielle de référence et l'approche de bout en bout seront également exposées. Nous présenterons les méthodes et les outils *d'évaluation* et de comparaison des performances des approches SLU construites. Nous illustrerons également comment nous avons pu mesurer la distance entre la parole réelle et la parole artificielle.

Le **chapitre 6** décrira la collecte du *corpus* de parole *réelle* qui servira de test. Il énumérera aussi les étapes de la génération de l'ensemble de données d'apprentissage de commandes vocales artificielles au format de texte et sous forme de signal acoustique, avec une annota-

tion automatique d'intentions et de concepts. Nous décrirons également la technique d'enrichissement des transcriptions d'étiquettes de symboles de concepts et d'intentions. La collecte d'un corpus d'apprentissage d'énoncés *sans* commandes vocales est également décrit.

Le **chapitre 7** est consacré à l'approche SLU séquentielle de référence. Nous décrirons la méthode d'apprentissage des Modèles Acoustiques (MA), la génération du Modèle de Langage (ML) et le lexique du module de RAP qui précède le module de NLU. Notre stratégie pour essayer d'améliorer le passage entre les modules de RAP et de NLU consistera à utiliser un modèle appris sur des transcriptions non-alignées car il est plus flexible.

Dans l'approche SLU de bout en bout cible, les processus de RAP et de NLU sont intimement associés, aussi le **chapitre 8** illustrera leur optimisation *conjointe* en appliquant l'approche SLU de bout en bout. En tant que phase préparatoire de validation de notre approche, des modèles de RAP de bout en bout ont été entraînés *sans* et *avec* la parole de synthèse vocale du corpus artificiel VocADom@ARTIF. Nous appliquerons aussi un apprentissage par *transfert* et nous positionnerons notre approche cible par rapport à l'approche SLU séquentielle de référence.

Contrairement à l'approche séquentielle, la SLU de bout en bout a accès aux deux niveaux *acoustique* et *prosodique* pour prédire les concepts et les intentions. Au cours du **chapitre 9** nous vérifierons l'impact de la prosodie sur les performances de la SLU de bout en bout. Nous chercherons aussi à savoir si le modèle de SLU de bout en bout peut être efficace sans nécessairement être très performant au niveau de la RAP. Comme nos utilisateurs cibles peuvent s'écarter d'une grammaire figée de commandes vocales, nous comparerons la *robustesse* de l'approche SLU de référence et l'approche SLU cible, en augmentant le taux de *mots hors vocabulaire* (OOV - *out of vocabulary*) et la *variation syntaxique* de l'ensemble de test.

Pour finir, nous présenterons les *conclusions* de ce travail de thèse et nos perspectives dans le **chapitre 10**.

CHAPITRE 2

État de l'art de l'habitat intelligent et des corpus correspondants

Dans ce chapitre nous allons donner en section 2.1 un aperçu des définitions les plus communément admises de l'*habitat intelligent*, afin d'en établir une qui soit suffisamment générale pour englober non seulement les maisons intelligentes *académiques* et *industrielles* mais aussi les *enceintes intelligentes* apparues plus récemment. Nous allons souligner de ce que ces habitats ont en commun, mais aussi les caractéristiques spécifiques qui les différencient. Nous remarquerons ensuite, en section 2.2 et 2.3, que souvent la conception de ces maisons intelligentes se concentre essentiellement sur les aspects *sécurité*, *confort* et *ubiquité*. En conséquence, les recherches se concentrent fréquemment sur la *détection de l'activité humaine*.

Bien que l'aspect *communication* fasse également partie de la définition, il s'avère qu'il attire moins l'attention. Néanmoins nous illustrerons cet aspect en section 2.5 en présentant deux projets pour lesquels la *communication* est la caractéristique principale. Il s'agit d'un projet de détection d'appels de détresse (section 2.5.1) et un deuxième projet de reconnaissance de commandes vocales appliqué à la domotique (section 2.5.2). Ces 2 projets nous ont beaucoup inspiré pour la réalisation de ce projet de thèse.

L'intérêt croissant porté aux maisons intelligentes s'est traduit par la mise en œuvre de différents projets qui ont permis la construction de plusieurs corpus adaptés à ce domaine. Ce processus de développement comprend l'enregistrement et l'annotation d'ensembles de données spécifiques au domaine. En effet, le développement de technologies pour la maison intelligente nécessite des ensembles de données obtenus à partir de plateformes expérimentales fournissant les conditions nécessaires pour représenter des situations réelles. Ces situations domotiques *réalistes* doivent être reflétées par les ensembles de données spécifiques au domaine. Cependant la collecte de ces données est très exigeante en termes d'efforts, de ressources, de temps, et demande un bon savoir-faire afin d'obtenir des données d'entraînement de bonne qualité. Étant donné le coût d'une telle acquisition, ces ressources sont souvent collectées dans le cadre de projets de recherche qui biaisent trop les données en fonction des objectifs d'un projet spécifique. En outre, ils ne sont souvent que disponibles pour une petite partie de la communauté de recherche. Par conséquent, cela ralentit les progrès de la recherche dans le domaine.

Pour surmonter cette situation, un certain nombre de corpus ont été collectés et mis à

disposition de la communauté, notamment pour la langue anglaise, chacun avec ses propres avantages et limites. L'étude de l'état de l'art de ces corpus domotiques nous permettra de tirer des enseignements utiles pour développer à la fois notre ensemble de test réaliste *VocADom@A4H* mais aussi notre ensemble de données d'apprentissage artificielles spécifiques au domaine. La section 2.6 est un aperçu des corpus spécifiques au domaine domotique en d'autres langues que le français. Dans la section 2.7 nous nous concentrons sur des corpus similaires, mais en français.

2.1 L'habitat intelligent : quelle définition ?

Le dictionnaire *Larousse* nous indique que le mot *domotique* a été construit à partir du mot latin "Domus" désignant une demeure patricienne auquel est ajouté le suffixe adjectival "tique". Ce même dictionnaire définit la domotique comme étant "l'Ensemble des *techniques* visant à intégrer à *l'habitat* tous les automatismes en matière de *sécurité*, de *gestion*, de l'énergie et de *communication*, etc."

La définition de ce mot par l'encyclopédie en ligne *Wikipédia* fait également ressortir dans sa définition les mots-clés suivants : (1) *technique*, (2) *habitat*, (3) *sécurité*, (4) *gestion* et (5) *communication* : "La domotique est l'ensemble des *techniques* de l'électronique, de la physique du bâtiment, des automatismes, de l'informatique et des *télécommunications* utilisées dans les *bâtiments*, plus ou moins interopérables et permettant de *centraliser le contrôle* des différents systèmes et sous-systèmes de la maison et de l'entreprise (chauffage, volets roulants, porte de garage, portail d'entrée, prises électriques, etc.). La domotique vise à apporter des solutions techniques pour répondre aux *besoins* de *confort* (gestion d'énergie, optimisation de l'éclairage et du chauffage), de *sécurité* (alarme) et de *communication* (commandes à distance, signaux visuels ou sonores, etc.) que l'on peut retrouver dans les maisons, les hôtels, les lieux publics." Aux 5 notions-clé de la définition du dictionnaire *Larousse*, la définition de *Wikipedia* ajoute donc (6) *le confort* et (7) les *besoins* (humains).

Sakamura (1996) définit par contre ce qui empêche un habitat d'être considéré comme un habitat intelligent à savoir l'absence de moyens de *communication* intérieure et extérieure et de *confort* : "Une maison sera disqualifiée au regard du classement dans la catégorie des maisons intelligentes si *l'information ne peut pas circuler* librement de l'intérieur de la maison vers le monde extérieur, et vice-versa... si elle est équipée avec des fonctions sophistiquées *difficiles à utiliser*." De cette manière, il renforce l'importance de l'aspect de la *communication*.

La définition de Cancellieri (1992) reprend ces caractéristiques mais en y ajoutant la notion (8) *d'adaptation* : "... immeuble qui réalise l'intégration de confort, sécurité, productivité et économie grâce aux ressources les plus récentes de la technologie... assurant un ensemble de services grâce à des systèmes réalisant plusieurs fonctions, et pouvant être connectés entre eux et à des réseaux internes et externes de communication... qui répond aux besoins actuels et *à venir*..."

Cook et Das (2004) reprennent cette caractéristique *d'adaptation* comme la notion la

plus importante permettant de qualifier ce qu'est l'intelligence d'un habitat. Ils étendent la définition de l'habitat intelligent en ajoutant une nouvelle caractéristique : (9) *l'acquisition de connaissance* sur ses habitants : "Un environnement intelligent est capable *d'acquérir* et d'appliquer de la *connaissance* d'un environnement et aussi de *s'adapter* à ses habitants afin d'améliorer leur expérience dans cet environnement". D'après leur étude, il s'agit non seulement du confort des habitants mais aussi de leur sécurité : "Ils peuvent souhaiter que l'environnement assure la sécurité de ses habitants".

Dans les définitions précédentes, la maison intelligente est plutôt considérée comme une maison, un bâtiment ou une entité *physique* spécifique, alors qu'elle peut également être considérée comme une entité invisible *omniprésente* dans la maison. C'est l'étape que franchit Weiser (2002) en introduisant (10) *l'ubiquité*. L'habitat intelligent est plutôt considéré comme "...du matériel et des logiciels, connectés par des fils, des ondes radio et infrarouges...si omniprésents que personne ne remarquera leur présence...".

En reprenant les dix caractéristiques essentielles citées précédemment, nous pouvons définir un habitat intelligent comme étant « un ensemble de (1) *techniques*, dans un (2) *bâtiment* ou une entité invisible mais (10) *ubiquitaire*, qui savent (5) *communiquer* et acquérir de la (9) *connaissance* sur ses habitants, s'y (8) *adapter*, en visant à intégrer les automatismes nécessaires pour satisfaire leurs (7) *besoins* en matière de (3) *sécurité*, (6) *confort*, *communication* et (4) *gestion* de l'habitat. »

2.2 Quelques exemples d'habitats intelligents académiques

Cette section 2.2 trace un aperçu des maisons intelligentes dans un contexte de recherche universitaire ou d'applications industrielles. Cela montrera que, bien souvent, les fonctionnalités citées dans la section précédente n'ont pas toutes été prises en compte pour la conception d'un habit intelligent spécifique, ou que, selon la finalité recherchée, certaines fonctionnalités sont mises en avant.

2.2.1 Aware Home

FIGURE 2.1 – Aware Home - (a) plan de l'habitat et (b) aspect extérieur (Abowd et coll., 2002)

Le projet *Aware Home* se concentre surtout sur le développement d'un habitat qui répond aux besoins de *sécurité* de ses habitants, notamment des personnes âgées, ce qui nécessite un environnement capable d'acquérir une connaissance sur leur présence et leurs

FIGURE 2.2 – Aware Home - (a) antenne RFID cachée sous le tapis (b) vue de l'antenne RFID (Abowd et coll., 2002)

activités dans les différentes pièces de la maison. Cet habitat intelligent fut mis en service en 2000. Il était constitué d'une maison à deux étages ayant une surface totale de 468m^2 et comportant deux appartements identiques ayant chacun deux chambres, une cave et un grenier (figure 2.1) (Abowd et coll., 2002). Un programme de recherche, l'Aware Home Research Initiaive (AHRI), a été lancé pour surmonter les défis multidisciplinaires de l'habitat intelligent et le défi de la détection de l'activité humaine dans un cadre domestique. Ce projet se concentre particulièrement sur le soutien aux personnes âgées ayant des capacités cognitives et sensorielles en déclin. Cela nécessite de détecter les situations de crise ou d'urgence. Un environnement *conscient* devait leur offrir de l'aide lors de l'apparition de troubles cognitifs empêchant la prise de médicaments, la préparation des repas ou l'utilisation d'appareils ménagers. Par exemple, lorsque l'habitant souffre d'une perte de mémoire à court terme, il pourrait être aidé par l'affichage d'une image de la pièce où il était précédemment montrant ce qu'il y faisait.

Les capteurs utilisés sont des antennes RFID (Radio-Frequency Identification). Les utilisateurs en portent à leur genou, d'autres sont placés sous les tapis comme le montre la figure 2.2). D'autres capteurs disponibles sont des microphones et des caméras placées au premier étage. Les données recueillies permettent de localiser chaque habitant à chaque instant en gardant en mémoire le type de capteur à l'origine de la localisation.

L'assistance à la personne nécessite une *acquisition de connaissance* de son activité, mais celle-ci est soumise à une incertitude : la personne assise sur le canapé tourne-t-elle les pages d'un journal ou actionne-t-elle la télécommande de la télévision? Certaines activités ne sont pas liées à un emplacement spécifique mais plutôt déterminée par l'interaction entre l'occupant et certains objets dans certaines pièces de l'habitat. Par exemple, le déplacement d'objets du réfrigérateur au comptoir, du comptoir à la table, etc. La reconnaissance de ces activités nécessite une représentation des relations temporelles et spatiales entre les éléments constitutifs des activités. Les auteurs ont utilisé des grammaires statistiques sans contexte (SCFG - Stochastic Context Free Grammar). Ces grammaires peuvent par exemple définir la préparation des repas comme nécessitant une activité près d'un réfrigérateur avant de s'asseoir à table avec une assiette de nourriture. La nature probabiliste des grammaires SCFG permet selon les auteurs l'utilisation des données incertaines provenant des capteurs.

Cette étude ne mentionne aucun résultat concernant l'utilisation de ces grammaires.

2.2.2 MavHome

Le projet *MavHome* (Managing an Adaptive Versatile Home) était un projet de recherche multidisciplinaire conduit à l'Université d'Arlington au Texas. Il avait pour objet la création d'un environnement domestique intelligent et polyvalent. Comme son nom l'indique, la caractéristique recherchée était de rendre cette maison *adaptable* pendant l'interaction avec les habitants. Le système MavHome est *ubiquitaire*, contrôlant et gérant la maison pour répondre aux *besoins de confort* des habitants plutôt qu'à leurs besoins de sécurité. Pour cela il trace les actions et la localisation des habitants dans les différentes pièces de la maison, qui comprend entre autre un laboratoire et une cuisine.

Dans le cadre de ce projet, l'habitat était considéré comme un agent relationnel, percevant l'état de la maison à l'aide des capteurs et agissant sur l'environnement par des actionneurs. Par ailleurs, cet agent maximise le confort des habitants tout en minimisant la consommation de ressources (par exemple, électricité, eau, gaz naturel) (Cook et Das, 2004).

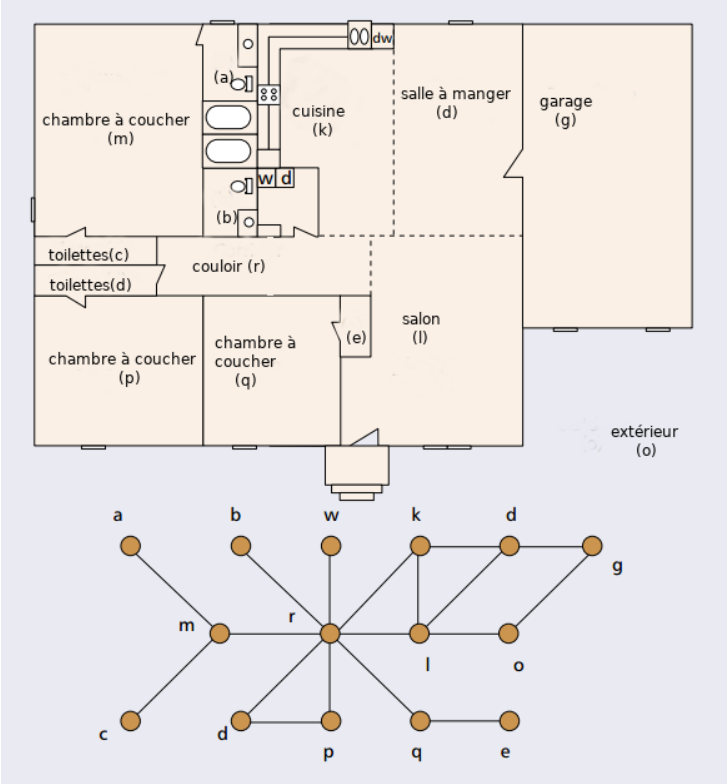


FIGURE 2.3 – MavHome - plan de l'appartement divisé en 15 zones et son graphe (Das et coll., 2002)

Cela impose que l'agent soit en mesure de prévoir la mobilité et les activités de ses habitants. Par conséquent, des activités qui normalement seraient initiées par les habitants sont prises en charge par le système. Par exemple, MavHome augmente le chauffage 15 minutes avant le réveil de ses habitants. La sonnerie pour réveiller les habitants se déclenche 15 mi-

nutes plus tard, ce qui entraîne l'allumage de la lumière de la chambre ainsi que la mise en route de la cafetière dans la cuisine.

Le système MavHome s'appuie sur 3 fonctions connectées dans une architecture modulaire : la collecte de données, la prédiction d'activité et la communication entre plusieurs agents coopérants. Cette ubiquité s'appuie sur l'utilisation de l'historique des mouvements pour prédire les emplacements futurs probables, pour déterminer quels épisodes de l'histoire d'activités d'un habitant sont les plus significatives. D'une telle façon, l'habitat peut prédire la prochaine interaction la plus probable entre les habitants et la maison.

Cette maison intelligente traque ses habitants à l'intérieur et tout autour de l'habitat, c'est l'étendue de la gestion de la localisation qui est divisée en plusieurs zones. Lorsque MavHome doit contacter un habitant, le système lance une recherche en interrogeant toutes les zones où l'habitant est susceptible de se trouver. La figure 2.3 montre un plan de la maison comprenant 15 zones et sa représentation graphique (Cook et Das, 2004; Das et coll., 2002).

Il s'agit donc d'un projet qui vise à ôter à l'habitant toute initiative, la maison devenant un automatisme en capacité de se gérer lui-même de manière autonome. Ceci ne va pas sans poser un problème éthique grave, l'habitant va être conduit à modifier son comportement pour s'adapter à celui que l'habitat intelligent à prévu pour lui.

2.2.3 House_n

Contrairement aux maisons intelligentes décrites dans les sections précédentes, l'approche de *House_n* vise à une utilisation éducative de la technologie. Le groupe MIT House_n envisage la technologie comme un moyen utile à l'individu pour maintenir une vie mentalement et physiquement stimulante au fur et à mesure du vieillissement. Cette démarche se situe alors aux *antipodes* de celles qui envisagent de créer une technologie gérant de manière *ubiquitaire* et proactive l'habitat.

Des capteurs sont utilisés pour déterminer quand et comment présenter des informations aux personnes au moment et à l'endroit où elles en ont besoin, à des "moments propices à l'apprentissage". L'objectif est d'aider les gens à prendre des décisions *sans* les priver de leur sentiment de contrôle sur leur environnement. Dans ce but, des environnements domestiques sont nécessaires pour permettre aux chercheurs, non seulement d'analyser le fonctionnement de la technologie domestique, mais également d'observer si les gens appliquent ce qu'ils apprennent, s'ils sont réceptifs aux informations présentées par la technologie et si ces informations sont intégrées dans les activités de la vie quotidienne (Intille, 2006).

2.2.4 GERHOME

L'habitat *GERHOME* cible la satisfaction aux besoins de *sécurité* et de *confort* des personnes âgées en suivant leur présence et leurs activités. Ce laboratoire expérimental combine l'utilisation de caméras et de capteurs *intégrés*. Il a été construit au CSTB (Centre Scien-

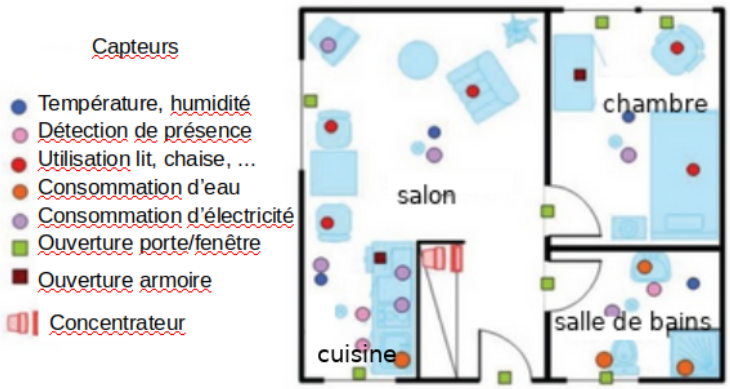


FIGURE 2.4 – GERHOME - plan montrant le positionnement des capteurs (Zouba et coll., 2009)

FIGURE 2.5 – GERHOME - instrumentation de l'habitat : capteurs, caméra et concentrateur (Zouba et coll., 2009)

tifique et Technique du bâtiment) de Sophia Antipolis en France [1]. Ce habitat simule un appartement destiné à une personne âgée, il a une surface de 41m^2 et se compose d'un hall d'entrée, d'un salon, d'une chambre à coucher, d'une salle de bains, et d'une cuisine. Cet habitat intelligent est équipé de différents capteurs destinés à évaluer les scénarios d'Activité de la Vie Quotidienne (AVQ ou ADL - Activities of Daily Living) prédéfinis par des gérontologues. Quatre caméras vidéo y sont installées : une caméra dans la cuisine, deux dans le salon et une dans la chambre à coucher (figure 2.4).

L'ensemble de ces caméras et capteurs intégrés sont utilisés pour collecter des données (figure 2.5). Ces capteurs environnementaux sont robustes et précis mais les coûts sont élevés en raison du nombre de capteurs requis. Les caméras sont moins précises bien qu'une seule caméra soit suffisante pour chaque pièce.

Le système d'évaluation consiste en :

— Un composant d'analyse vidéo qui *détecte* et suit les *habitants observés*, reconnaît leur posture et un ensemble d'événements détectés par une caméra vidéo.

— Un composant d'analyse de données de *capteur environnemental* qui collecte des données sur les interactions interpersonnelles et les objets contextuels. Ce composant reconnaît également un ensemble d'événements environnementaux simples (par

1. https://www.inria.fr/fr/centre-inria-sophia-antipolis-mediterranee

exemple, la porte du réfrigérateur qui est ouverte).

— Un composant de reconnaissance d'activité *multimodale* qui combine les événements vidéo et environnementaux pour reconnaître des activités complexes (par exemple, un habitant qui prépare un repas).

La sortie du système est un ensemble d'événements reconnus représentés en fichiers XML, ou par une visualisation 3D (Zouba et coll., 2009).

Plusieurs projets de détection d'activité humaine ont utilisé l'appartement GERHOME. Pomponio et coll. (2012) décrivent une approche de détection d'activité humaine où ils proposent un cadre théorique général pour définir et identifier les activités des habitants. À l'instar d'autres projets de détection d'activité humaine, ils considèrent la reconnaissance de l'activité comme s'appuyant sur un processus d'abstractions successifs : événements discrets, séquence d'événements discrets et puis classification taxonomique au plus haut niveau. Cependant, ce travail le combine avec une théorie d'observation chronométrée. Dans (Vandewiele et Motamed, 2011), les informations provenant des caméras et des capteurs sont traitées dans le cadre d'une détection d'activité humaine non supervisée.

2.2.5 CASAS

La caractéristique principale de l'habitat *CASAS* est son *adaptabilité* qui lui permet d'acquérir une connaissance suffisante des activités de ses habitants, qui sont des personnes âgées, pour ensuite s'adapter aux changements dans leur comportement. Ce système utilise des techniques d'apprentissage automatique pour découvrir des *motifs* dans les activités quotidiennes des résidents. Il peut également s'adapter aux changements des motifs découverts en fonction des commentaires implicites et explicites des résidents. Par conséquent, il peut automatiquement mettre à jour son modèle pour refléter ces changements. L'adaptabilité est importante dans ce contexte pour aider les personnes ayant des limitations cognitives et physiques. En découvrant des séquences répétitives, en modélisant leurs contraintes temporelles, les tâches quotidiennes répétitives peuvent être automatisées de manière intelligente. Dans ce travail, les études entreprises se sont limitées aux cas où un seul résident est présent (Rashidi et Cook, 2009).

Cette adaptabilité est réalisée en combinant un outil de recherche de correspondance de motifs fréquents (FPAM, Frequent Pattern Activity Mining), un modèle d'activité hiérarchique (HAM, hierarchical activity model) et un explorateur de motifs d'adaptation (PAM, Pattern Adaption Miner). Les données d'entrée collectées à partir de capteurs sont extraites par l'algorithme FPAM pour découvrir des motifs d'activités fréquentes et d'intérêt pour l'automatisation. Ces modèles sont ensuite classés par le modèle d'activité hiérarchique (HAM). HAM capture les relations temporelles entre les événements d'une activité en représentant explicitement les ordres de séquence dans une structure arborescente. L'algorithme Pattern Adaptation Miner (PAM) s'adapte à toutes les modifications de ces motifs. Il analyse les données d'événements récents et recherche les changements dans le motif, tels que l'heure de début du motif, les durées, les périodes ou la structure du motif.

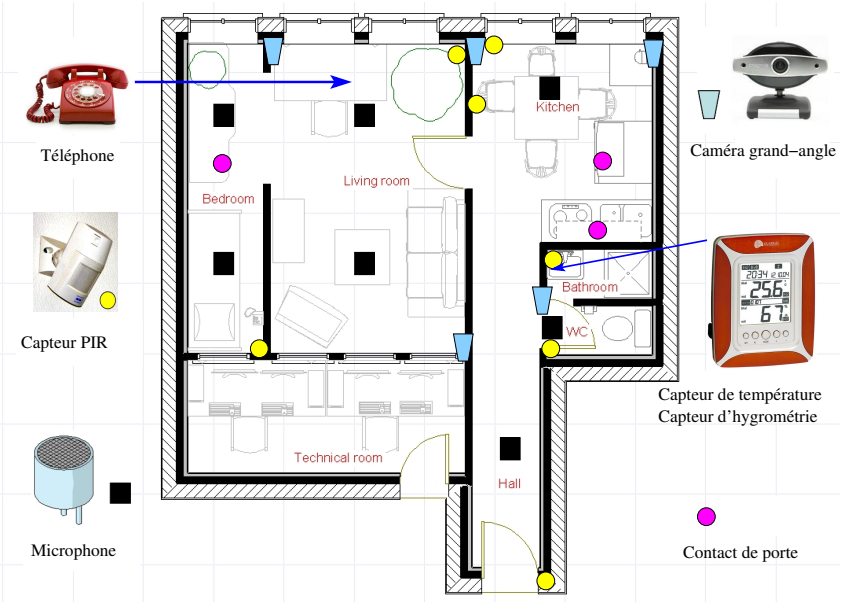


FIGURE 2.6 – HIS TIMC-IMAG - plan montrant l'emplacement des principaux capteurs et des caméras, la régie est en bas à gauche (Fleury et coll., 2010b)

2.2.6 Appartement HIS du laboratoire TIMC-IMAG

Contrairement aux habitats décrits dans les sections précédentes, l'aspect *communication* a cette fois été abordé ici. L'habitat *HIS* (Health Smart Home of Grenoble) a été construit en 1999 dans un but médical pour des études consacrées aux populations fragiles, essentiellement les personnes âgées, afin d'étudier les moyens de les aider à rester vivre à leur domicile. Le but recherché dans cet habitat est d'assurer la *sécurité* des personnes âgées. C'est en effet à cette époque que des études sur l'Assistance à domicile (Ambient Assisted Living) ont été initiées (Rialle et coll., 2001). Cet appartement a été implanté par l'équipe AFIRM du laboratoire TIMC-IMAG (Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications) à l'intérieur de la Faculté de Médecine de Grenoble. Cet appartement de 47m^2 était composé d'une chambre, d'un salon, d'un hall, d'une cuisine et d'une salle de bain. Un local technique supplémentaire contient tout l'équipement nécessaire à la gestion du réseau informatique reliant les capteurs et à l'enregistrement des données (Fleury et coll., 2010b).

Cet appartement était équipé au départ de capteurs de détection de présence infra-rouge (PIR), d'un thermomètre, de contacts d'ouverture de porte sur les portes des placards, du réfrigérateur et sur les tiroirs ainsi que de capteurs médicaux (pèse personne, tensiomètre, etc.). L'occupant du HIS pouvait aussi porter des capteurs placés directement sur son corps (par exemple un actimètre). Cet équipement a permis de réaliser différentes études dont l'une consacrée à la détection de l'évolution du rythme circadien pouvant être signe d'une affection clinique (Virone et coll., 2002).

Cet habitat intelligent a ensuite été équipé au cours de l'année 2000 de microphones placés au plafond et dirigés verticalement vers le sol. Le but était de détecter et reconnaître des sons caractéristiques d'un danger pour la personne, par exemple des cris ou un bruit de chute (Istrate et coll., 2006). Le développement du logiciel temps-réel *AuditHIS* a ensuite

permis de reconnaître des appels à l'aide de la personne en cas de danger ou lorsqu'elles estiment qu'elles ont besoin d'aide. Une expérimentation décrite dans Fleury et coll. (2010a) a impliqué 12 personnes volontaires qui ont joué des scénarios permettant de réaliser plusieurs activités de la vie quotidienne dans le HIS incluant aussi des appels à l'aide. Outre les résultats obtenus sur la reconnaissance des AVQ, cela a aussi permis de construire le premier corpus multimodal comprenant des paroles enregistrées par des locuteurs français lors d'interactions avec une maison intelligente (Fleury et coll., 2010b) (ce corpus est décrit en détail dans la section 2.7 de ce chapitre). Ce corpus a permis d'approfondir les études sur l'AVQ mais aussi de mettre en évidence les défis posés par la reconnaissance automatique de la parole dans un milieu sonore difficile et non contraint qui est celui de l'habitat (Vacher et coll., 2011)

Dans les deux sections qui suivent, nous examinerons le cas de deux habitats intelligents, *DOMUS* et *Amiqual4Home*, pour lesquels la mise en œuvre de l'aspect de *communication* langagière pourra apporter une nouvelle dimension, l'interaction, grâce à la domotique qu'il sera possible de commander par la voix. Deux corpus très utiles pour nos recherches ont été enregistrés dans ces habitats : SWEET-HOME (section 2.7.5) et VocADom@A4H (chapitre 6, section 6.2).

2.2.7 DOMUS

DOMUS était un appartement intelligent d'une trentaine de mètres carrés comprenant une salle de bain, une cuisine, une chambre à coucher et un bureau situé au Centre des Technologies Logicielles (CTL) de l'Université de Grenoble. Toutes les pièces étaient équipées de capteurs et d'actionneurs reliés à un système domotique. L'appartement était entièrement fonctionnel et habitable (figure 2.7).

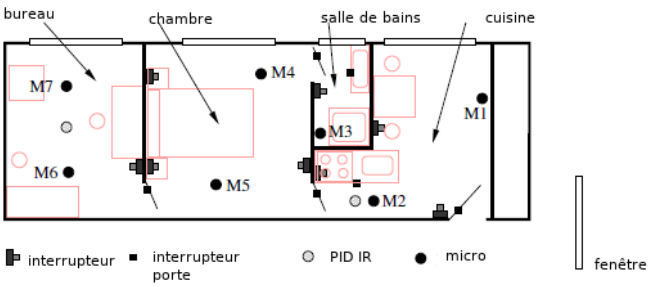


FIGURE 2.7 – DOMUS - plan de l'appartement montrant position des 7 microphones et des capteurs domotiques (Vacher et coll., 2015)

L'objectif recherché, lors de la construction de la plate-forme DOMUS, était de disposer d'un appartement intelligent intégré dans un environnement propice à l'expérimentation orientée usage. Il a ainsi été conçu pour une utilisation suivant une approche "centrée utilisateur", en considérant qu'un habitat est un système interactif enveloppant l'individu (Gallissot et coll., 2013). Ce habitat intelligent faisait partie de la plateforme d'expérimentation du laboratoire LIG et était dédié aux projets de recherche. DOMUS était équipé :

— de capteurs, tels de simples interrupteurs, ou des appareils de mesure de la consommation d'énergie et d'eau, du niveau d'humidité, de la température.

— et d'actionneurs capables de contrôler l'éclairage, les stores, la diffusion multimédia (voix de synthèse, TV, radio, musique).

Les capteurs et actionneurs étaient répartis dans la cuisine, la chambre, le bureau et la salle de bain. Une couche logicielle d'interfaçage a été conçue pour envoyer des commandes aux différents actionneurs, et de recevoir les modifications des valeurs des capteurs. Une instrumentation d'observation, avec caméras, microphones et systèmes de suivi d'activité, permettait de contrôler et de superviser les expérimentations depuis une salle de contrôle, la régie, connectée à l'environnement domotique de DOMUS (figure 2.7). Aperçu des technologies disponibles dans l'appartement DOMUS :

— KNX pour les réseaux domotique (éclairage, volets roulants, capteurs température, luminosité)

— X2D pour des détecteurs d'ouverture sans fil (portes, fenêtres et placards)

— UPnP/DLNA pour les loisirs numériques (téléviseur, diffusion sonore type multizones)

— RFID pour les interactions tangibles

— DMX512 pour l'éclairage d'ambiance

— ZigBee "Green Power" pour des mesures énergétiques localisées.

L'appartement était équipé au départ de 7 microphones radio SENNHEISER ME2 cachés dans le plafond (2 par chambre sauf un pour la salle de bain) qui peuvent enregistrer en temps réel grâce à un logiciel dédié capable d'enregistrer simultanément les canaux audio couplés à une carte multi-canal National Instrument PCI-6220E. Un 8ème canal pouvait être utilisé pour enregistrer un microphone placé devant une source de bruit (par exemple une radio ou un aspirateur). Cet appartement a été notamment utilisé pour des expérimentations de commande vocale de la domotique, soit en magicien d'Oz, soit en interagissant avec un système automatique temps-réel (Vacher et coll., 2015). Il a aussi servi pour une expérimentation de reconnaissance d'appels des personnes âgées (Vacher et coll., 2019). Plus récemment, chaque pièce a été équipée d'une antenne de 4 microphones LC 97 TWS Lavalier, placés chacun à l'angle d'un carré de 10cm de côté. Les antennes étaient fixées au plafond et dirigées vers le sol.

L'appartement intelligent DOMUS dans sa version initiale a été démonté en 2019, une version plus moderne est en cours d'installation à la Maison de la Création et de l'Innovation (MACI) de l'UGA. C'est dans ce nouvel habitat intelligent que notre approche SLU séquentielle de référence et notre approche SLU cible de bout en bout seront implémentées et évaluées.

2.2.8 Amiqual4Home

Amiqual4Home[2] est une plate-forme d'expérimentation pour la recherche et l'innova-

2. https://amiqual4home.inria.fr

tion située à Montbonnot. Elle est composée de plusieurs équipements : des ateliers de prototypage (FabLab), des espaces d'expérimentation (dont un habitat intelligent), et des outils mobiles permettant l'observation d'activité humaine. Amiqual4Home est un équipement de la communauté Université Grenoble Alpes géré par Inria. Nous conviendrons de désigner son habitat intelligent lui-même par "Amiqual4Home".

Cet appartement de 87m^2 comprend deux étages équipés de systèmes domotiques, d'appareils multimédias et de réseaux de microphones (figure 2.8). Environ 150 capteurs et actionneurs ont été installés pour acquérir de la parole, contrôler les lumières, régler le chauffage, etc. Des expérimentateurs dans une chambre de contrôle, cachés aux locuteurs, peuvent réagir aux commandes vocales des participants en suivant une stratégie de *magicien d'Oz* pour rendre l'interaction entre les locuteurs et la maison intelligente aussi réaliste que possible (Voir figures 2.9 et 2.10).

FIGURE 2.8 – Amiqual4Home - cuisine équipée (https://amiqual4home.inria.fr)

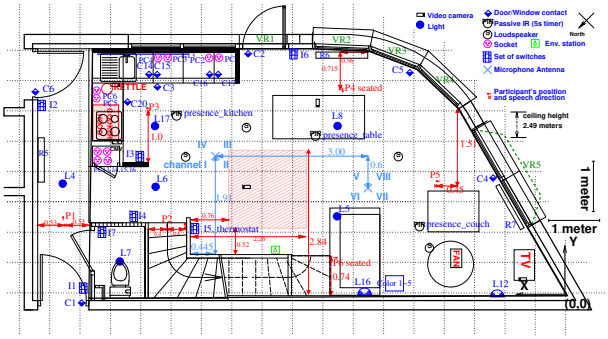


FIGURE 2.9 – Amiqual4Home - rez-de-chaussée : cuisine et salon (https://amiqual4home.inria.fr)

Les capteurs et actionneurs domotiques, par exemple, l'éclairage, les volets, les systèmes de sécurité, la gestion de l'énergie, le chauffage, etc., sont connectés par un bus KNX[3] (norme ISO/IEC 14543). Outre KNX, plusieurs bus de terrain coexistent, comme UPnP (Universal Plug and Play) pour la distribution multimédia, X2D pour la détection des contacts (portes, fenêtres et armoires), RFID pour l'interaction avec les objets tangibles. La gestion du réseau domotique, l'envoi de commandes aux différents actionneurs et la réception des modifications des valeurs des capteurs, est effectuée via openHAB[4]. Cette couche garantit

3. https://www.knx.org/
4. https://www.openhab.org/

l'interopérabilité des données provenant des différents bus de terrain et permet la communication entre eux et vers des applications virtuelles, comme le suivi d'activité. Grâce à cette passerelle, tous les appareils, y compris les éléments multimédias, peuvent être contrôlés à distance. De plus, 9 caméras sont installées au plafond des pièces. Ce réseau domotique de capteurs openHAB génère des fichiers de trace contenant des étiquettes d'activités des dispositifs de l'habitat avec leur horodatage ainsi que la pièce où le dispositif est activé ou désactivé 2.11.

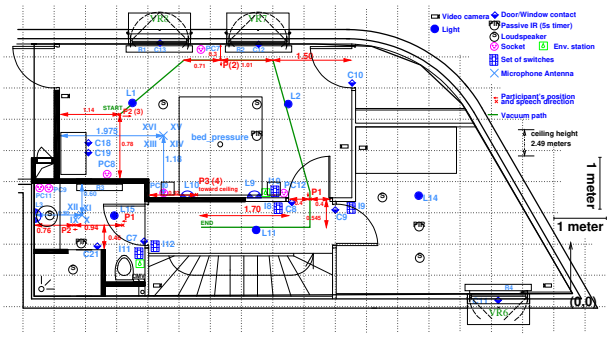


FIGURE 2.10 – Amiqual4Home - premier étage : chambre à coucher, salle de bains, toilettes (https://amiqual4home.inria.fr)

```
2017-06-13 16:52:15.171000;livingroom_music_playfile;no_keyword
2017-06-13 16:52:38.538000;livingroom_music_playfile;not_understood
2017-06-13 16:53:09.027000;VR1,VR2,VR3,VR4,VR5;LOWER
2017-06-13 16:53:09.946000;VR1;LOWER
2017-06-13 16:53:10.715000;VR2,VR3,VR4,VR5;LOWER
2017-06-13 16:53:29.139000;VR1,VR2,VR3,VR4,VR5;RAISE
2017-06-13 16:53:29.994000;VR1;RAISE
2017-06-13 16:53:30.978000;VR2,VR3,VR4,VR5;RAISE
2017-06-13 16:53:50.259000;livingroom_music_playfile;no_keyword
2017-06-13 16:54:01.099000;livingroom_music_playfile;not_understood
```

FIGURE 2.11 – Amiqual4Home - extrait d'un fichier de trace openhab.log enregistrant l'activation des dispositifs (https://amiqual4home.inria.fr)

Quatre antennes de 4 microphones chacunes, comme illustré sur la figure 2.12, sont installées dans le plafond de la cuisine, du salon, de la salle de bain et de la chambre à coucher, dirigées vers le sol. Chaque antenne est composé de 4 microphones LC 97 TWS Lavalier. En plus, le participant peut porter un microphone Sennheiser HSP 4 devant la bouche pour faciliter la transcription de la parole. Grâce à cet appartement, nous avons pu collecter les enregistrements en *micro-porté* et en *micro-distant* du corpus VocADom@A4H. Ce corpus qui représente notre ensemble de test réaliste sera décrit en détail au chapitre 6 section 6.2.

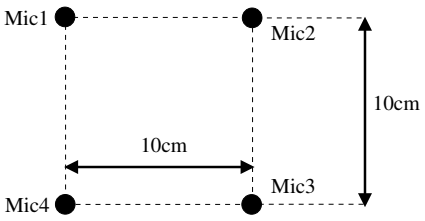


FIGURE 2.12 – Amiqual4Home - réseau de microphones (https://amiqual4home.inria.fr)

2.3 Exemples d'habitats intelligents industriels

2.3.1 Microsoft Easy Living

Le projet *Microsoft Easy Living* met l'accent sur *l'ubiquité* des installations informatiques omniprésentes d'un système de maison intelligente. Il doit permettre à l'utilisateur de passer d'une pièce à l'autre tout en conservant une session interactive avec l'ordinateur, l'interface utilisateur suivant l'utilisateur (Brumitt et Shafer, 2001; Shafer et coll., 1998).

À la différence des maisons intelligentes décrites dans les exemples précédents, il a été développé comme un ensemble à installer dans la maison de l'utilisateur. Cependant, il devrait offrir les services suivants : détecter et localiser ses utilisateurs, comprendre leur relation physique et fonctionnelle avec les appareils, répondre aux commandes vocales et gestuelles et être facilement étendu. Le suivi des utilisateurs est basé sur un module de vision composé de quatre caméras, avec contrôle et traitement effectués sur un PC (Shafer et coll., 1998).

En plus du module de vision, EasyLiving a essayé d'améliorer la détection et la localisation des utilisateurs en intégrant son modèle dit de *géométrie*. Pour communiquer avec quelqu'un, qui est par exemple au travail, le moyen le plus efficace de communiquer consiste à utiliser l'appareil qui lui est le plus proche dans son environnement de travail (écrans, téléphone fixe, téléphone portable, haut-parleur d'ordinateur, etc.), et non d'utiliser un canal de communication fixe comme par exemple le téléphone portable de l'utilisateur. Une telle approche nécessite une compréhension de l'emplacement de la personne, de sa relation physique avec les appareils qui l'entourent et de son environnement. Dans une situation à domicile, *l'acquisition de connaissance*, la compréhension de la géométrie permettent à une session de suivre l'utilisateur d'une pièce à une autre, par exemple de l'ordinateur au bureau vers le salon, en la projetant automatiquement sur un grand écran mural dans le salon (figure 2.13) (Brumitt et coll., 2000).

Le noyau de la compréhension géométrique d'EasyLiving est le module géométrique EasyLiving (EZLGM - Easy Living Geometric Model). Il se concentre sur l'informatique omniprésente à la maison ou au bureau, avec des dispositifs de perception et informatiques observant plusieurs utilisateurs. EZLGM et le module de vision EasyLiving sont combinés avec la compréhension de la parole et l'interprétation des mouvements humains tels que

FIGURE 2.13 – Easy Living - session projetée automatiquement sur un grand écran mural (Brumitt et coll., 2000)

les gestes. Les expériences ont montré que la parole était préférée pour décrire des objets et donner des commandes. Les gestes, comme le pointage, ont été préférés pour transmettre des emplacements (Shafer et coll., 1998).

Ce type de projet se cantonne au cas d'une personne qui souhaite pouvoir retrouver chez elle son cadre de travail, il ne s'intéresse pas à la vie quotidienne de l'habitant.

2.3.2 IBM Pervasive Computing Lab

Pervasive Computing Lab développé en 2003 par IBM Research à Austin, Texas, a permis d'effectuer des travaux de recherche grâce à un salon multimédia avancé, une cuisine connectée à un réseau avec des appareils intégrés. L'appartement occupe environ la moitié du laboratoire et comprend trois pièces : un salon, une cuisine et un garage. Chaque appareil de ces pièces est connecté à une passerelle qui permet aux appareils périphériques d'échanger des données, de se connecter à un serveur ou de naviguer sur le Web. Parmi les objets intelligents, il y avait un fer à repasser, une cuisinière, un réfrigérateur, une télévision, des stores, des lumières.

L'ordinateur passerelle et les appareils domestiques sauf deux exécutent des programmes Java sur Linux. Les deux autres exécutent des programmes Java sur un RTOS propriétaire de QNX. QNX est un système d'exploitation compatible POSIX qui intègre un micro-noyau, conçu principalement pour le marché des systèmes embarqués tels que les voitures mais aussi pour les industries et les services médicaux des hôpitaux (Cook et Das, 2004).

2.4 Enceintes intelligentes

Les enceintes intelligents, telles qu'*Amazon Echo* ou *Google Home*, sont devenues des accessoires répandus dans les foyers du monde entier depuis les cinq dernières années. Ces appareils offrent à leurs utilisateurs un moyen d'interagir avec les systèmes informatiques de la maison, par la voix, sans avoir besoin de toucher un appareil. Ils permettent aux utilisateurs d'émettre un large éventail de requêtes concernant des sujets divers (la météo, la bourse, les plans de voyage, les heures d'ouverture des magasins, etc.).

Notons que ces systèmes enregistrent en général les paroles prononcées et les utilisent par ailleurs sans que leurs utilisateurs en aient pleinement conscience. Ceci pose des problèmes d'éthique surtout lorsque ces systèmes fonctionnent mal et enregistrent sans qu'il y ait eu prononciation du mot-clé censé les mettre en route.

Les constructeurs de ces appareils affirment offrir également une intégration intelligente de la maison pour faciliter l'utilisation de la domotique (contrôle des lumières, des systèmes de chauffage, des accès,etc.). Contrairement aux "maisons intelligentes traditionnelles", les utilisateurs établissent une connexion émotionnelle avec le haut-parleur intelligent, ou plutôt avec la voix de l'enceinte intelligente (Purington et coll., 2017).

Cette nouvelle gamme d'enceintes intelligentes représente les premières interfaces vo-

cales ambiantes qui sont facilement disponibles dans les environnements domestiques, pour des millions de personnes, d'autant plus que leurs prix baissent.

Malgré la grande diffusion de ces produits commerciaux, il y a malheureusement une absence totale d'information sur la façon dont leurs architectures sont conçues. On doit se contenter des comparaisons et des évaluations de performances réalisées pour les produits les plus répandus tels que Google et *Alexa*. Des études sur les principaux types d'utilisateurs, les commandes les plus fréquemment utilisées, la longueur de ces commandes, sont disponibles (Bentley et coll., 2018; Purington et coll., 2017). L'étude de Bentley et coll. (2018) mentionne par exemple que la longueur moyenne des énoncés des commandes extraites par Google Home comptait 4 mots. 25% de toutes les commandes ne contenait que deux mots et moins de 25% des commandes dépassait une longueur de cinq mots.

Les études comparatives sur les performances disponibles ne révèlent pas l'approche utilisée et ne permettent pas de localiser ou d'analyser des failles dans leurs architectures. Un exemple d'une telle évaluation est présenté dans l'étude de López et coll. (2017) qui compare les systèmes Google home à Amazon Alexa, Apple Siri et Microsoft Cortana.

Ces systèmes sont comparés sur la base des mêmes aspects et services vocaux. L'évaluation a été effectuée par 92 étudiants. Chaque participant a évalué les quatre assistants personnels sur le *taux de naturel* de la réponse d'une part et *le taux de l'exactitude* de la réponse d'autre part. *Exact* signifie, sans erreur, et conformément aux faits ou à la vérité. Plusieurs scénarios ont été développés pour évaluer les enceintes intelligentes. Tous les participants ont répondu en utilisant une échelle de 5 points pour évaluer le naturel et l'exactitude de chaque réponse. Cette échelle comprend cinq niveaux : (1) très médiocre, (2) faible, (3) moyen, (4) supérieur et (5) excellent. La figure 2.14 présente les résultats pour le *naturel* des réponses. Alexa et Google montrent les meilleures performances, ce dernier devançant Alexa d'environ 12%. Le *taux de l'exactitude* des réponses est montrée dans la figure 2.15. Les performances de Google et d'Alexa sont également supérieures. Siri par contre, bien que l'un des assistants vocaux les plus répandus du marché, montre des performances faibles par rapport aux trois autres systèmes. La section *résultats et discussion* de cet article ne contient aucune analyse, explication ou même hypothèse pour expliquer ces résultats étant donné les secrets industriels dont s'entourent les fabricants.

2.5 La commande vocale dans l'habitat intelligent

Au cours des sections précédentes, nous avons pu mettre en évidence que les études décrites sur la plupart des habitats intelligents sont plutôt consacrées à l'évaluation des systèmes de l'habitat intelligent, notamment la détection de présence et des activités des habitants. Par contre, les études sur les enceintes intelligentes se concentrent sur l'interaction vocale, mais sans aborder les spécificités linguistiques de cette interaction. Nous allons maintenant présenter deux projets d'interaction par commande vocale dans l'habitat intelligent pour lesquels l'aspect *communication* est la caractéristique principale mais qui s'intéressent également aux spécificités *linguistiques* de l'interaction entre la maison intelligente et ses

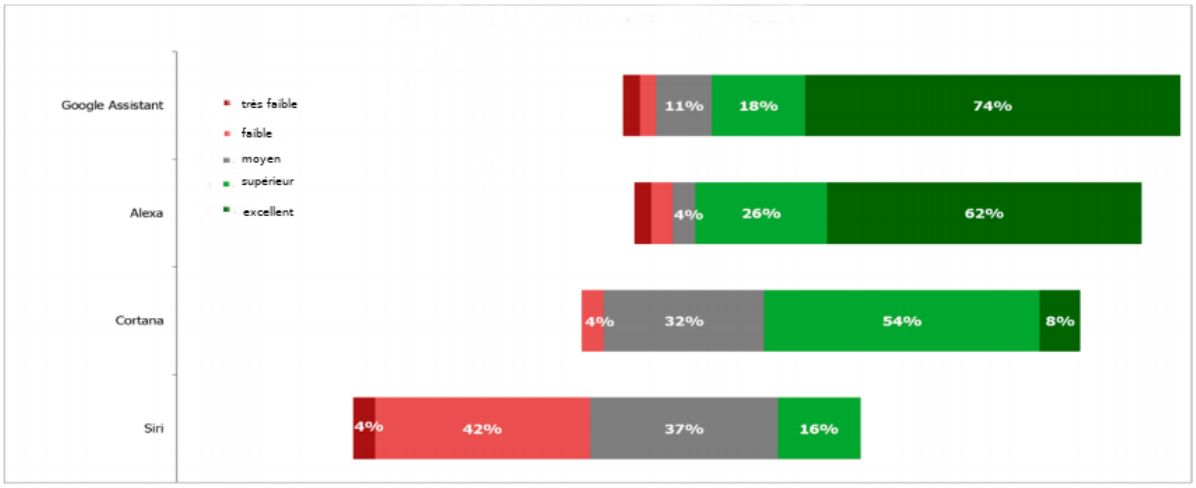


FIGURE 2.14 – Enceintes intelligentes - comparaison du naturel de la réponse (López et coll., 2017)

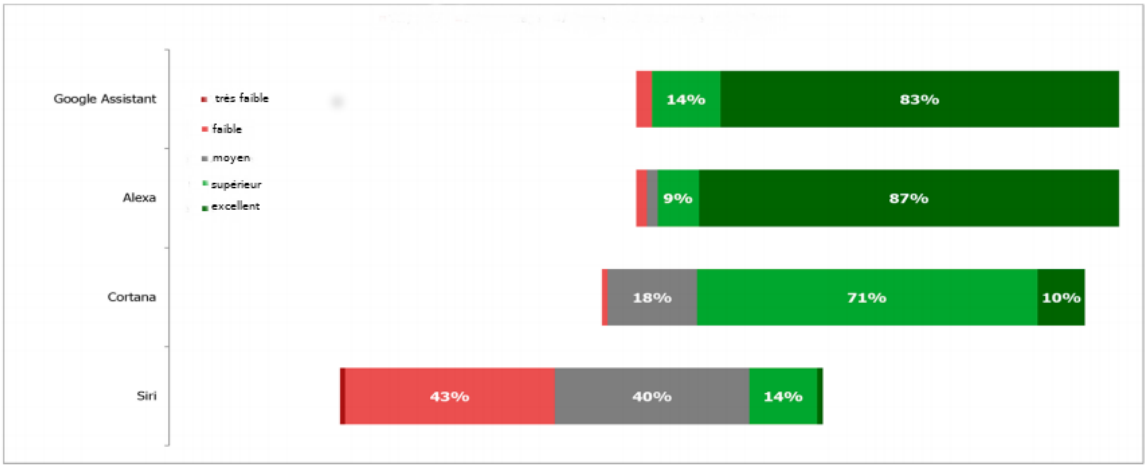


FIGURE 2.15 – Enceintes intelligentes - comparaison de l'exactitude de la réponse (López et coll., 2017)

habitants.

2.5.1 DESDHIS : Détection des Signaux de Détresse dans l'Habitat Intelligent pour la Santé

Le projet DESDHIS de l'équipe GEOD du laboratoire CLIPS financé par l'ACI Santé du ministère a été mené en collaboration avec le laboratoire TIMC-IMAG. Les expérimentations ont été menée dans le HIS du laboratoire TIMC-IMAG qui est décrit à la section 2.2.6 (Fleury et coll., 2010b). En ce qui concerne la parole, il s'agissait de reconnaître des appels à l'aide. Les enregistrements réalisés lors de ces expérimentations ont permis de constituer le corpus *HIS* (section 2.7) qui est le premier corpus en français incluant des enregistrements des commandes vocales dans une maison intelligente en condition de parole distante.

Les expérimentations ont utilisé le logiciel temps-réel d'analyse sonore AuditHIS qui analysait les sons recueillis par les 8 microphones placés dans l'appartement en conservant à chaque fois l'évènement sonore de plus fort RSB (Rapport Signal sur Bruit, SNR - Signal

TABLE 2.1 – Résultats du projet DESDHIS (Vacher et coll., 2011)

	MAR (%)	FAR (%)	GER (%)
Taux d'erreurs	29.5	4	15.6

to Noise Ratio). Un classificateur GMM classait chaque événement audio comme son non langagier ou comme parole.

Un système de RAP basé sur le système *JANUS* était ensuite en charge de la reconnaissance de la parole des évènements classés comme parole. Son modèle acoustique (MA) était entraîné sur les corpus de parole BRAF100 (Vaufreydaz et coll., 2000), BREF80 et BREF120 (Gauvain et coll., 1990). Son petit modèle de langage (ML) se composait de 299 unigrammes, 729 bigrammes et 862 trigrammes. Il était obtenu à partir du corpus de français conversationnel *ANODIN DETRESSE* (AD) (section 2.7.2). L'objectif du système était la détection correcte d'un appel de détresse grâce à la détection de mots clés, sans vraiment comprendre la conversation du locuteur.

Pour valider le système, un scénario a été défini et joué par 10 locuteurs (3 femmes et 7 hommes) ont dû prononcer 45 phrases (20 phrases de détresse, 10 phrases normales et 15 phrases de conversation téléphonique en effectuant des activités de la vie quotidienne prédéfinies (AVQ). Les enregistrements contiennent :

— 39 ordres domotiques (exemple, "Allume la lumière")

— 93 phrases de détresse (exemple, "Une infirmière vite")

— 93 phrases de conversation quotidienne (exemple, "Il fait froid")

Le temps de parole recueillie est très court, 7,8 minutes, comparé à la durée totale des scénarios.

La détection des appels de détresse détectés était évaluée en définissant les alarmes manquées (MA, missed alarm) et les fausses alarmes (FA, false alarm) :

— Il s'agit d'une MA si la phrase prononcée est une phrase de détresse sans mot-clé.

— Il s'agit d'une FA si la phrase prononcée est une phrase sans détresse, avec ou sans ordre domotique, mais contenant un mot-clé de détresse.

Les auteurs ont défini le taux d'alarmes manquées (MAR, missing alarm rate), le taux de fausses alarmes (FAR, false alarm rate) et le taux d'erreurs globales (GER, global error rate) en équations (2.1) et (2.2) où n fait référence au « nombre de », DS (distress sentence) aux phrases de détresse et NS (normal sentence) aux phrases normales. Le tableau 2.1 montre les taux d'erreurs obtenus.

$$MAR = \frac{nMA}{nDS}, FAR = \frac{nFA}{nNS} \tag{2.1}$$

$$GER = \frac{nMA + nFA}{nDS + nNS} \tag{2.2}$$

2.5.2 Sweet Home : Système domotique d'assistance au domicile

TABLE 2.2 – Sweet-Home - grammaire des commandes vocales

Commande basicCmd = key initiateCommand object | key stopCommand [object] | key emergencyCommand
Mot-clé key = "Nestor" | "maison"
Type de commande initiateCommand = "ouvre" | "ferme" | "baisse" | "éteins" | "monte" |"allume" | "descend" | "appelle" stopCommand = "stop" | "arrête" emergencyCommand = "au secours" | "à l'aide"
Objet de la commande object = [determiner] (device | person | organisation)
Type d'objet determiner = "mon" | "ma" | "l'" | "le" | "la" | "les" | "un" | "des" |"du" device = "lumière" | "store" | "rideau" | "télé" | "télévision" |"radio" person = "fille" | "fils" | "femme" | "mari" | "infirmière" |"médecin" | "docteur" organisation = "samu" | "secours" | "pompiers" | "supérette" |"supermarché"

Le projet Sweet-Home a été financé par l'ANR en vue d'aider les personnes âgées ou malvoyantes grâce à la commande vocale de la domotique (Vacher et coll., 2015). Les expérimentations ont eu lieu dans l'habitat intelligent DOMUS (section 2.2.7).

Les commandes vocales ont été définies à base d'une grammaire simple (tableau 2.2). Chaque commande appartient à l'une des trois catégories de commandes : *initiation* (initiateCommand), *arrêt* (stopCommand) et *appel d'urgence* (emergencyCommand). À l'exception de l'appel d'urgence, chaque commande commence par un mot-clé unique qui permet de savoir si la personne parle à la maison intelligente ou non. Le tableau 2.3 montre des exemples où "Nestor" est utilisé comme mot-clé. Chaque commande, sauf les appels d'urgence, contient un objet (*object*).

C'est l'outil *PATSH* qui était en charge d'effectuer la reconnaissance de commandes vocales en temps réel et d'interagir avec le système domotique. Le modèle acoustique (MA) du système de RAP était appris sur 80 heures de parole annotée. Le modèle de langage (ML) trigramme final est le résultat d'une interpolation d'un ML générique (poids = 10%) et spécifique au domaine (poids = 90%) à base de 1000M mots. Le lexique contenait 10K mots. Lorsque les hypothèses de RAP des énoncés des commandes vocales ne respectaient pas la grammaire, l'énoncé était rejeté. La reconnaissance de la commande prononcée était effectuée grâce à la recherche des mots clefs de la grammaire dans la meilleure hypothèse issue du système de RAP.

Plusieurs expérimentations ont été réalisées, soit avec des personnes jeunes, soit avec des utilisateurs potentiels du système (6 personnes âgées et 5 personnes malvoyantes). Les personnes jeunes ont joué des scénarios incluant 7 AVQ. Pour les personnes âgées ou malvoyantes, les scénarios étaient simplifiés.

TABLE 2.3 – Sweet-Home - exemples de commandes vocales(Vacher et coll., 2014)

Phrase	**Type de commande**	**Description**
Nestor ferme fenêtre	clé initiation objet	activer un *actionneur*
Nestor arrête	clé arrêt [objet]	arrêter un actionneur
Au secours	appel d'urgence	appeler un service de secours

L'ensemble des enregistrements a été réuni à chaque fois dans un corpus dédié, mis à la disposition de la communauté de la recherche (Vacher et coll., 2014). Le corpus multi-modal *SWEET-HOME*, contenant des commandes vocales, enregistré par des personnes jeunes jouant des scénarios comprenant 7 AVQ, est décrit en section 2.7.5.

2.6 Corpus pour d'autres langues que le français

L'attention croissante portée aux maisons intelligentes dans la communauté de recherche a permis la mise à disposition de plusieurs corpus de parole adaptés à ce domaine par la communauté scientifique. Les sections suivantes fournissent un aperçu de corpus spécifiques au domaine domotique pour d'autres langues que le français.

2.6.1 CHiME-1, CHiME-2 et CHiME-5 corpus domotiques

Le corpus anglais *CHiME-1*, comprend 1000 commandes domotiques différentes qui sont générées suivant une grammaire fixe (Fig. 2.16). *CHiME-2* par contre a étendu CHiME-1 à des locuteurs ayant une position variable dans l'espace et utilisant un vocabulaire de 5000 mots tirés du Wall Street Journal (WSJ) (Barker et coll., 2013, 2017). Les corpus CHiME pour la langue anglaise visaient à faire progresser le développement de la reconnaissance vocale dans un contexte quotidien. Ils contiennent des enregistrements effectués en conditions distantes en présence de bruit. Les sources principales de bruit sont typiques d'un environnement domestique : la télévision, l'ordinateur, la conversation entre plusieurs locuteurs, certains bruits de la rue provenant de la circulation à l'extérieur et les bruits des pièces adjacentes, y compris le bruit général de la cuisine. Les deux corpus contiennent des enregistrements de 34 locuteurs (18 hommes et 16 femmes)

À la différence des corpus CHiME-1 et CHiME-2 qui contiennent des commandes basées sur des règles, le corpus *CHiME-5* contient de la parole spontanée. Les participants étaient libres de discuter entre eux sur les sujets de leur choix. Le corpus est enregistré dans un contexte de dîner. 20 dîners ont été enregistrés avec chaque fois deux hôtes accompagnés de deux invités. Chaque session de dîner dure environ deux heures et était aussi naturelle que possible. Un dîner se compose de trois phases, dont chaque phase se déroule dans une pièce différente (Barker et coll., 2018).

2.6.2 Le corpus DIRHA anglais

Le corpus anglais multi-microphone *DIRHA* (Ravanelli et coll., 2015) est également enregistré dans un environnement domestique, avec des microphones distants. Il complète l'ensemble des corpus précédemment collectés dans le cadre du projet DIRHA en quatre autres langues (allemand autrichien, grec, italien et portugais) (Cristoforetti et coll., 2014). 50% des données sont basées sur des simulations et l'autre moitié sur des enregistrements réels. Les énoncés contiennent des mots-clés pour activer le système domotique. Le dialogue qui est ainsi déclenché donne à l'utilisateur un accès aux appareils et services, par exemple, ouvrir/fermer les portes et les fenêtres, allumer/éteindre les lumières, contrôler la température, jouer de la musique, etc.

Le corpus total comprend onze heures de commandes domotiques lues, de paroles extraites du Wall Street Journal Corpus (WSJ) (Paul et Baker, 1992) ou de paroles spontanées. Les enregistrements ont été faits par douze locuteurs natifs du Royaume-Uni et des États-Unis (6 hommes et 6 femmes).

2.6.3 Le corpus DICIT

Le corpus *DICIT* (Brutti et coll., 2008) a été créé dans le contexte du projet européen DICIT1 (*Distant-talking Interfaces for Control of Interactive TV*). Chaque session d'enregistrement a été divisée en deux phases. Dans une première phase, des phrases phonétiquement riches ont été lues. Au cours de la deuxième phase, chaque personne a interagi avec le système d'une manière spontanée en essayant d'accomplir une liste de tâches prédéfinies. Les enregistrements en anglais, allemand et italien ont été effectués avec un groupe de quatre personnes pour simuler un scénario domestique typique, comme les membres d'une famille qui regardent la télé ensemble. Il s'agit d'une interface homme-machine qui permet à l'utilisateur d'interagir avec la télévision, les appareils numériques et services d'infodivertissement. Pendant une expérience Magicien d'Oz, un sujet est invité à effectuer des tâches spécifiques. Chaque interaction avec l'utilisateur a duré environ 10 minutes, pour un total de 360 minutes d'enregistrements. Les annotations ont été effectuées à l'aide de l'outil

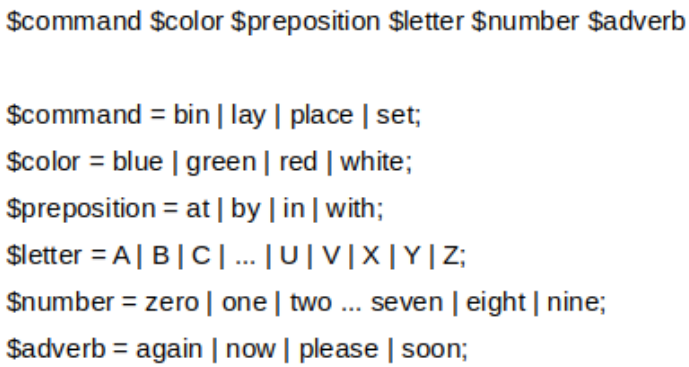


FIGURE 2.16 – Corpus CHIME - commandes à base d'une grammaire fixe de six mots (Barker et coll., 2013)

d'annotation *Transcriber*[5] qui permet une vue multicanal. Ces annotations comprennent le nom (ID) du locuteur, la transcription des phrases énoncées, des bruits de fond, d'après le protocole d'annotation. Sept classes de bruits ont été identifiées et annotées entre crochets. Par exemple, `[pap]` signifie bruissement du papier (figure 2.17).

Etiquette	Evénement acoustique	
[sla]	claquement de porte	(door slamming)
[cha]	chaise en mouvement	(chair moving)
[pho]	sonnerie du téléphone	(phone ringing)
[cou]	tousser	(coughing)
[lau]	rire	(laughing)
[fal]	objets tombants	(objects falling)
[pap]	bruissement du papier	(paper rustling)
[spk]	bruits du locuteur	(noises from speaker)
[unk]	autres bruits inconnus	(other unknown noises)

FIGURE 2.17 – DICIT corpus - annotations bruit de fond - Transcriber (Brutti et coll., 2008)

2.6.4 Le corpus ITAAL

Le corpus *ITAAL* pour la langue italienne a été créé dans le but de reconnaître les appels de détresse et les commandes vocales domotiques dans une maison intelligente (Principi et coll., 2013). Les appels de détresse (par exemple "Appelez un médecin!") sont reconnus dans le but d'aider les personnes à domicile : lorsqu'ils sont détectés, un appel téléphonique est automatiquement fait avec un contact dans un carnet d'adresses et par conséquent la personne en détresse peut demander de l'aide. Le corpus contient également des phrases lues, phonétiquement riches. Ces dernières phrases ont été extraites du corpus *APASCI* (Angelini et coll., 1993) et couvrent tous les phonèmes de la langue italienne. Ce corpus contient environ 20 heures d'enregistrements par 20 locuteurs natifs italiens (moité hommes et moitié femmes). L'âge moyen des locuteurs est de 41,70 ans. Les énoncés ont été prononcés à une distance de 3 mètres du réseau de microphones. Les commandes et appels ont été enregistrés dans des conditions normales et criées.

Comme dans le cas des corpus CHiME-1 et CHiME-2 (section 2.6.1), le système est basé sur des règles pour détecter les commandes et les appels. Le tableau 2.4 fournit des exemples de règles de grammaire pour les commandes et les appels de détresse.

TABLE 2.4 – Corpus ITAAL - grammaire de détection de commandes et détresse (Principi et coll., 2013)

Intention	**Exemple (Italien)**	**traduction française**
`<commande>`	*(accendi | spegni) la luce)*	*(allume | éteins) la lumière*
`<appel de détresse>`	*aiuto | ambulanza*	*au secours | ambulance*

5. http://trans.sourceforge.net

TABLE 2.5 – Le corpus Fluent Speech Commands - aperçu (Lugosch et coll., 2019)

Ensemble	**# de locuteurs**	**# d'énoncés**	**# d'heures**
Entraînement	77	23,123	14.7
Développement	10	3,118	1.9
Test	10	3,793	2.4
Total	97	30,043	19.0

2.6.5 Fluent Speech Commands dataset

La plupart des ensembles de données SLU disponibles sont soit non *open source*, soit trop petits. C'est pourquoi le corpus *Fluent Speech Commands*[6] a été créé par Lugosch et coll. (2019) et collecté par des locuteurs en *crowdsourcing*. Plus précisément, ce corpus anglais peut être utilisé pour entraîner et tester un système capable de reconnaître un ensemble de commandes vocales pour interagir avec un assistant vocal dans un scénario de maison intelligente. Le corpus contient 30043 énoncés de 97 locuteurs et est enregistré sous forme de fichiers WAV monocanaux enregistré à 16 kHz. Chaque fichier contient un seul énoncé comme par exemple, "mettez de la musique" ou "augmentez la température dans la cuisine". Chaque audio est étiqueté avec des concepts d'action, d'objet ou d'emplacement et avec des valeurs. Par exemple, le concept "emplacement" peut avoir comme valeur "aucun", "cuisine", "chambre" ou "salle de bain". L'intention est définie comme une combinaison spécifique de valeurs d'attributs (*slots*). Les énoncés "allumer les lumières", "est-ce que tu peux allumer les lumières", correspondent à l'intention action : "activer", l'objet : "lumières", l'emplacement : "aucun".

Les informations démographiques sur ces locuteurs anonymes (âge, sexe, etc.) sont incluses dans le corpus. Les énoncés sont divisés de manière aléatoire en ensembles d'entraînement, de développement et de test. En outre ils sont partagés de manière à ce qu'aucun locuteur n'apparaisse dans plus d'un ensemble. Le tableau 2.5 donne plus d'informations sur cet ensemble de données.

2.6.6 Speech Commands Dataset for Limited-Vocabulary Speech Recognition

L'ensemble de données *Speech Commands* est un ensemble de données de mots prononcés en anglais également collecté par des locuteurs en *crowdsourcing*. Différent des corpus des sections précédentes, il a été conçu pour entraîner et évaluer les systèmes de repérage des mots-clés (Warden, 2018). Son objectif principal est de fournir un moyen de construire et de tester de petits modèles qui détectent un seul mot, prononcé à partir d'un ensemble de dix mots cibles ou moins.

Un tel système doit aussi être capable de distinguer le mot-clé du bruit de fond ou d'autres conversations. Cela signifie que les enregistrements n'ont pas été effectués dans un studio car il manquerait du bruit de fond. En outre de tels enregistrements seraient capturés

6. fluent.ai/research/fluent-speech-commands/

avec des microphones de haute qualité. Les modèles de détection de mot-clé performants, doivent faire face à des environnements bruyants, à un équipement d'enregistrement de mauvaise qualité et à des personnes qui produisent de la parole spontanée. Pour refléter tout cela, tous les énoncés ont été capturés à l'aide de microphones de téléphone ou d'ordinateur portable, à n'importe quel endroit où les utilisateurs se trouvaient. La durée de chaque énoncé était limitée à une durée standard d'une seconde. Le vocabulaire limité résultant de mots courts se compose d'environ 20 mots communs.

Afin de respecter la vie privée, il a été évité d'enregistrer toute information personnelle des locuteurs. Cela signifie que les informations sur le sexe ou l'origine ethnique ou toute autre information pouvant être liée à des données personnelles ne sont pas disponibles.

Pour mesurer la qualité, les critères d'acceptation ou de rejet des soumissions étaient la bonne compréhension de la parole enregistrée par un auditeur humain. Les échantillons contenant des mots qui semblaient incorrects ou qui n'étaient pas compréhensibles ont finalement été rejetés. Cela a été effectué par des travailleurs commerciaux en *crowdsourcing*. L'ensemble de données final, avec des enregistrements de 2618 locuteurs, comprend 105829 énoncés de 35 mots.

2.6.7 Domotica dataset

			Speaker ID																		
Frame	slot	Slot-values	28	29	30	31	32	33	34	35	37	17	11	40	41	42	43	44	45	46	47	48
Triple commands	object	headrest	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	0	1	1	0
	object	standing light	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	0	1	1	0
	action	1	1	0	1	1	1	1	0	0	0	1	1	1	1	1	0	1	0	1	1	0
	action	2	1	1	1	0	1	1	0	1	0	1	1	1	1	1	0	1	0	1	1	0
	action	3	1	0	1	0	1	1	0	1	0	1	1	1	1	1	0	1	0	1	1	0
On/off	object	bath room light	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	object	bed room light	1	0	1	1	1	0	0	0	1	1	1	1	1	1	1	1	0	1	0	0
	object	living room kitchen light	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	object	kitchen light	0	0	0	0	0	1	0	0	1	1	1	0	1	1	0	1	0	1	1	0
	object	Kitchen stove light	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	object	Kitchen table lamp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	object	reading light	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	object	living room light	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	1	1	1
	object	All lights	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	action	on	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	action	off	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1
Open/close	object	bath room door	1	0	0	1	1	0	0	0	1	1	1	1	1	1	0	1	0	1	1	0
	object	living room door shutter	1	0	0	0	0	0	0	1	1	1	1	1	1	1	0	1	0	1	1	0
	object	bed room shutter	1	0	1	0	1	0	0	0	1	1	1	1	1	0	0	1	0	1	1	0
	object	Living room window shutter	1	0	0	1	1	0	0	1	1	1	1	1	1	1	0	1	0	1	1	0
	object	bed room door	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0	1	0	1	1	0
	object	Living room door	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
	object	front door	1	0	1	0	1	0	0	1	1	1	1	1	1	1	0	1	0	1	1	0
	object	Sleeping room shutter	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	action	open	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	action	close	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Increase Heating	{}	{}	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

FIGURE 2.18 – Corpus Domotica - concepts et valeurs par locuteur (Tessema et coll., 2013)

Pour la langue néerlandaise (belge flamand), nous mentionnons l'ensemble de données *Domotica*[7] (Tessema et coll., 2013), un corpus basé sur un système domotique qui cible les locuteurs pathologiques. Le corpus Domotica a été collecté en deux phases. Pendant une première phase, neuf locuteurs pathologiques ont été invités à contrôler 31 appareils dans

7. https://www.esat.kuleuven.be/psi/spraak/downloads

un environnement 3D simulé. L'expérience a été réalisée dans une configuration magicien d'Oz, dans laquelle l'utilisateur pouvait visiter l'environnement 3D.

Dans la deuxième phase, 21 locuteurs pathologiques ont été enregistrés. Huit locuteurs sur 21 ont été diagnostiqués avec une sclérose en plaques, une maladie évolutive qui entraîne une dégénérescence de la voix. La base de données qui en résulte, contient 2049 énoncés. Ils ont tous été transcrits et annotés manuellement de classes d'intentions et de concepts. Les 25 concepts ont trait aux objets actionnables (lumières, portes, etc.) et toutes les actions qu'on peut effectuer dans cet environnement 3D. La figure 2.18 montre les concepts par locuteur.

2.7 Les corpus français

2.7.1 Le corpus HIS

Le corpus *HIS* (Fleury et coll., 2010b) est le premier corpus français à inclure des enregistrements de la parole et de la vie quotidienne dans une maison intelligente, ici le HIS (Health Smart Home of Grenoble), dans un contexte de santé. Différent des corpus mentionnés précédemment, il contient aussi les traces des capteurs domotiques disponibles.

Chacun des 15 participants était libre de choisir l'ordre dans lequel il voulait effectuer les activités de la vie quotidienne afin d' éviter les schémas répétitifs : (1) Dormir; (2) Se reposer : regarder la télévision, écouter la radio, lire un magazine, etc.; (3) S'habiller et se déshabiller; (4) Alimentation : réaliser et prendre un repas; (5) Activité hygiénique : lavage des mains, des dents, etc.; et (6) Communiquer : utiliser le téléphone. Selon le participant, chaque expérimentation a duré entre 23 minutes et 1 h 35 minutes. Pendant l'expérience, 1886 sons non langagiers et 669 énoncés de parole ont été recueillis. Les paroles sont soit des paroles anodines de la vie quotidienne soit des appels à l'aide.

2.7.2 Le corpus ANODIN-DETRESSE (AD)

De la même manière que le corpus ITAAL de la section 2.6.4 de ce chapitre, le contexte du corpus *ANODIN-DETRESSE* (*AD*) a trait à la détection de situations de détresse. Ce corpus a été enregistré en studio par l'équipe GEOD du CLIPS lors du projet DESDHIS (Vacher et coll., 2006), dans le cadre de détection de situations de détresse utilisant la détection de mots-clés et la discrimination entre son et parole. 21 personnes âgées (11 femmes et 10 hommes, entre 22 et 64 ans) ont été enregistrées dans un contexte domestique, prononçant un total de 2646 énoncés annotés pour une durée totale de 38 minutes de parole (Vacher et coll., 2008). Les 126 phrases prononcés, des énoncés courts (1-3 secondes), sont les mêmes pour tous les locuteurs et se composent de 66 phrases *anodines* qui sont typiques d'une conversation normale, comme "Bonjour", "il fait beau", etc. Les autres 60 phrases expriment une situation de *détresse*, comme par exemple "Au secours", "Un médecin vite".

2.7.3 Le corpus ERES38

Le corpus *ERES38* (Entretiens RESidences 38) est un corpus de parole spontanée enregistré dans le cadre du projet CIRDO dans le lieu de vie de personnes âgées volontaires, qui étaient résidentes de structures spécifiques pour personnes âgées, comme par exemple des maisons de retraite. Au cours des entretiens, chaque locuteur a été invité à lire un texte sur des activités de jardinage, mais il leur a également été demandé de parler librement de leur vie. Il a été acquis auprès de 22 personnes âgées (14 femmes et 8 hommes) entre 68 et 98 ans. Le corpus comprenait 48 minutes de lecture et 17 heures d'entretiens. Le but de l'enregistrement de ce corpus était l'étude des caractéristiques de la voix des personnes âgées ainsi que l'adaptation des modèles acoustiques à la voix des personnes âgées (Aman et coll., 2013).

2.7.4 Le corpus CIRDO

Le corpus *CIRDO* est un corpus enregistré dans des conditions réalistes à l'intérieur de l'habitat intelligent DOMUS au cours du projet CIRDO financé par l'appel « Technologies pour la Santé » de l'ANR. Des participants ont joué dans cet appartement des scénarios comprenant de vraies chutes sur un tapis ainsi que des appels à l'aide (Vacher et coll., 2016). Le corpus a été utilisé pour des études liées à la détection de détresse en temps réel, à l'analyse de la parole lors de ces conditions particulières, et à la détection de chute par analyse vidéo. Un dispositif appelé e-lio était utilisé préférentiellement dans le cadre du projet pour acheminer les appels à l'aide identifiés.

Quatre scénarios étaient liés à une chute, chacun comprenant des phrases appelant à l'aide. Pour rendre les scénarios aussi réels que possible, un simulateur de vieillesse réduisait la mobilité, la vision et l'ouïe des participants non âgés. Ces scénarios ont été élaborés suite à une étude de terrain (Bobillier-Chaumon et coll., 2012).

Les phrases des scénarios ont été choisies au cours de l'étude sur le terrain et contiennent des phrases de détresse avec ou sans mot-clé, comme par exemple "Je peux pas me relever", "e-lio appelle du secours". 17 participants ont été recrutés (9 hommes et 8 femmes). Parmi eux, 13 personnes avaient moins de 60 ans et ont utilisé le simulateur. La durée d'acquisition était longue, 2 heures et 30 minutes par personne.

Le corpus est divisé en trois parties : le corpus audio pour l'analyse du son et de la parole, le corpus vidéo et une troisième partie avec les annotations. La parole a été annotée en utilisant l'outil Transcriber. La parole bruitée par le son produit par la chute, ainsi que les cris, les interjections, les soupirs etc., a été ignorée.

2.7.5 Le corpus SWEET-HOME

SWEET-HOME est un corpus multimodal français de commande vocale enregistré dans des conditions réalistes dans la maison intelligente DOMUS. Ce habitat intelligent et le traitement de commandes vocales sont respectivement décrits dans les sections 2.2.7 et 2.5.2 de ce chapitre. Il s'agit du premier corpus français d'énoncés de commandes vocales dans

une maison enregistré en conditions distantes accompagné des traces domotiques (Vacher et coll., 2014).

Le premier jeu du corpus est un corpus multimodal d'interaction. Les participants ont effectué des AVQ comme (1) dormir, (2) se reposer, (3) s'habiller/se déshabiller, (4) préparer/prendre le repas, (5) avoir des activités d'hygiène et (6) communiquer. 21 participants (14 hommes et 7 femmes) ont été invités à effectuer ces activités sans aucune condition sur le temps passé ou sur la manière de faire des activités.

Le deuxième jeu du corpus est un corpus de commandes vocales en conditions distantes. Les commandes vocales ont cette fois été enregistrées par 23 locuteurs (14 hommes et 9 femmes) avec un âge moyen de 35 ans. Les commandes sont définies en utilisant une grammaire simple et sont détaillées dans la section 2.5.2 sur *le système domotique d'assistance au domicile SWEET-HOME,* tableau 2.2. Chaque commande vocale commence par un mot-clé. Afin d'être plus réaliste, deux types de bruit de fond ont été considérés pendant que les locuteurs parlaient : la radio et de la musique classique. Chaque participant a prononcé 30 phrases dans différentes pièces suivant 3 conditions différentes : condition (1) sans bruit de fond, condition (2) avec la radio allumée et condition (3) avec de la musique classique comme bruit de fond. Les phrases ont été annotées manuellement sur le meilleur canal RSB à l'aide de Transcriber. Le corpus est disponible pour la communauté de recherche [8]. Il en résulte un riche ensemble de données multimodales, collectées en associant de la parole avec des données domotiques.

2.7.6 Les corpus VoiceHome et VoiceHome-2

Nous mentionnons les corpus *VoiceHome* et *VoiceHome-2* pour le traitement vocal multicanal dans des maisons réelles (Bertin et coll., 2019). L'objectif du projet voiceHome est la commande micro-distant des appareils multimédias et domestiques intelligents à base de dialogue naturel. VoiceHome-2 est une extension du corpus VoiceHome (Bertin et coll., 2016). Par rapport à VoiceHome-2 plus de données ont été collectées et annotées, y compris dans un plus grand nombre de maisons, de pièces et de locuteurs. Comme corpus complémentaires, VoiceHome et VoiceHome-2 sont un ensemble de données complet et autonome permettant l'apprentissage automatique, le développement et le test de techniques de traitement de la parole. La différence principale entre les deux corpus est que ce dernier corpus contient de la parole spontanée alors que seulement un ensemble de phrases courtes ont été enregistrées pour construire le corpus VoiceHome. Le tableau 2.6 compare les 2 corpus.

Tous les énoncés sont entièrement annotés, les transcriptions de parole sont accompagnées de l' emplacement dans l'une des 12 pièces. Cependant, le corpus ne contient aucune trace de capteur domotique. Les énoncés ont été générés à base d'une grammaire de seulement 3 intentions possibles : une question, un souhait et une commande. Elles commencent toutes par le mot-clé "OK Vesta". Les intentions sont générées à base d'un vocabulaire de 345 mots comprenant quelques concepts (principalement des noms de chaînes de télévi-

8. http://sweet-home-data.imag.fr

TABLE 2.6 – Les corpus VoiceHome et VoiceHome-2 - aperçu (Bertin et coll., 2019)

Corpus caractéristiques	**VoiceHome**	**VoiceHome-2**
Maisons (parole bruitée)	1	4
Pièces (parole bruitée)	1	12
Locuteurs (parole bruitée)	3	12
Durée totale (parole bruitée)	2.5 heures	5 heures
Énoncés	360	1560
Parole spontanée	non	72 min.
Room impulse response	188 (12 pièces)	non

sion françaises). Le corpus contient uniquement des énoncés générés par la grammaire et aucun énoncé sans intention. Malgré la génération d'énoncés avec 3 intentions possibles basées sur une grammaire, les étiquettes d'intention n'ont pas été générées et n'ont pas non plus été annotées manuellement.

2.7.7 SNIPS spoken-language-understanding-research-datasets

Similaire aux corpus Fluent Speech Commands (section 2.6.5) et Speech Commands Dataset for Limited-Vocabulary Speech Recognition (section 2.6.6), le corpus *SNIPS*[9] a été collecté par *crowdsourcing* pour la compréhension de la parole en anglais et français. Il est publiquement accessible et contient quelques milliers de requêtes textuelles avec leurs intentions et leurs attributs (*slots*) (Saade et coll., 2018). Pour chaque requête de texte dans l'ensemble de données, un énoncé est collecté. SNIPS contient aussi des données enregistrées à distance qui sont créées en les enregistrant à l'aide d'un réseau de microphones positionnés à une distance de 2 mètres. L'ensemble de données français couvre 8 intentions permettant de contrôler une enceinte intelligente et de jouer de la musique à partir de grandes bibliothèques d'artistes et d'albums. La taille du vocabulaire est de plus de 70000 mots en français. Un aperçu des 1992 intentions et de leur fréquence est présenté dans le tableau 2.7. Elles sont fournies en version micro-porté et micro-distant. Chaque énoncé est annoté en format .json par intention, avec les concepts et leur valeurs. Pour la phrase "je veux écouter", l'intention est "PlayMusic" (Jouer musique), et le concept "musicArtist" (artiste musicien) :

```
"intents": {
  "PlayMusic": {
    "utterances": [
      {
        "data": [
          {
            "text": "je veux écouter "
          },
          {
            "entity": "snips/musicArtist",
            "slot_name": "artist_name",
            "text": "Sade"
```

9. https://github.com/snipsco/spoken-language-understanding-research-datasets

TABLE 2.7 – Le corpus SNIPS - spoken-language-understanding-research-dataset French

Intention	**Fréquence**
NextSong (Chanson prochaine)	126
PreviousSong (Chanson précédente)	62
SpeakerInterrupt (Interruption de locuteur)	421
ResumeMusic (Continuer la Musique)	107
VolumeShift (Changer le volume)	437
VolumeSet (Régler le volume)	229
GetInfos (Demande d'info)	62
PlayMusic (Jouer musique)	548
Total	1992


```
      }
    ]
  },
  ...
```


2.8 Conclusion

À partir de plusieurs définitions d'habitats intelligents et de l'état de l'art, nous avons proposé une définition permettant de caractériser aussi bien les habitats intelligents académiques, industriels que les enceintes intelligentes. Les caractéristiques principales concernent la *sécurité*, le *confort* et l'*ubiquité*. Malgré la grande diffusion des enceintes intelligentes ces dernières années, il y a une pénurie d'informations sur la façon dont leurs architectures sont conçues ce qui ne leur permet pas de contribuer à l'enrichissement de la connaissance au sein de la communauté scientifique. Il n'existe que quelques projets français qui se concentrent sur l'aspect de la *communication* et sur la reconnaissance des commandes vocales dans un contexte de l'assistance à domicile. Cependant, ces projets qui ont été mis en place dans les maisons intelligentes HIS du laboratoire TIMC-IMAG et DOMUS appliquent des méthodes d'extraction d'intention basées sur des règles, ciblant un vocabulaire de petite taille avec peu de variation linguistique.

Il s'avère que, concernant les corpus domotiques, qu'ils soient en français ou dans une langue autre que le français, la plupart ont été développés dans un contexte visant la reconnaissance de la parole distante et bruitée, et non la compréhension de la parole. Par conséquent ces ensembles de données ne sont pas annotés au niveau sémantique. En outre la taille du vocabulaire et le nombre d'intentions s'avérant plutôt réduit, ces corpus sont bien adaptés à des systèmes de reconnaissance d'intentions à base de règles. Notons que lorsque ces corpus contiennent de la parole spontanée, il s'agit la plupart du temps d'énoncés sans intention précise.

Comme nos utilisateurs cibles sont des adultes âgés qui s'écartent facilement d'une grammaire figée de commandes vocales (Möller et coll., 2008; Takahashi et coll., 2003; Vacher et coll., 2015), les modèles NLU doivent être entraînés sur des corpus de commandes

vocales avec beaucoup plus de variations syntaxique et sémantique, et contenant des énoncés de longueur plus variable. Le corpus SWEET-HOME (2.7.5), enregistré dans DOMUS est le corpus qui est malgré tout le plus proche de nos besoins sans vraiment les satisfaire. En effet, le contexte de l'interaction à base de commandes vocales entre les utilisateurs et l'habitat intelligent exige de traiter plus d'intentions et de concepts que celles présentes dans ce corpus. C'est pourquoi nous verrons qu'un autre corpus plus adapté a été construit dans le cadre du projet VocADom, c'est le corpus VocADom@A4H qui sera présenté ultérieurement au cours du chapitre 6.

CHAPITRE 3

État de l'art de la reconnaissance automatique de la parole et de la compréhension du langage naturel

Dans ce chapitre, nous décrivons l'état de l'art de la reconnaissance automatique de la parole (RAP) et de la compréhension du langage naturel (NLU - Natural Language Understanding). En effet, la compréhension du langage naturel peut s'effectuer, soit à partir de l'hypothèse de reconnaissance issue d'un système de RAP, il s'agit alors d'une méthode de compréhension de la parole (SLU - Spoken Language Understanding) *séquentielle*, soit directement par analyse du signal sonore dans le cas des systèmes SLU dits « *de bout en bout* » (E2E - *End-to-End*).

3.1 Reconnaissance Automatique de la Parole

Les systèmes de Reconnaissance Automatique de la Parole (RAP) à l'état de l'art utilisent des méthodes statistiques. Nous présentons tout d'abord les principes de base des méthodes les plus classiques ainsi que les composants sur lesquels s'appuie dans ce cas le décodeur, à savoir le Modèle Acoustique (MA), le lexique phonétique, et le Modèle de Langage (ML).

Ensuite nous décrivons les approches plus récentes utilisant les réseaux de neurones profonds (*DNN - Deep Neural Networks*) qui permettent une reconnaissance automatique de la parole de bout en bout. Nous concluons la section sur la RAP en présentant de manière détaillée deux outils de RAP : *Deep Speech* et *ESPnet*.

3.1.1 Systèmes de RAP classiques

Les premiers systèmes modernes de RAP statistiques de Jelinek (1976) estiment la séquence de mots la plus probable $\hat{W}$ contenue dans un signal de parole après une étape de pré-traitement nécessaire pour découper le signal en trames et extraire une séquence de paramètres acoustiques $X = x_1, x_2, \ldots, x_y$. La séquence de mots $\hat{W}$ est obtenue en appliquant la formule de Bayes afin de maximiser $P(X|W)P(W)$, comme le montre l'équation 3.1.

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(W|X) = \underset{W}{\operatorname{argmax}}\, \frac{P(X|W)P(W)}{P(X)} = \underset{W}{\operatorname{argmax}}\, P(X|W)P(W) \tag{3.1}$$

Le décodeur s'appuie donc sur un premier étage d'extraction de paramètres acoustiques, un Modèle Acoustique (MA) probabiliste $P(X|W)$, un lexique de prononciation et un Modèle de

Langage ML $P(W)$ combiné pour extraire la séquence de mots la plus probable. La figure 3.1 montre l'architecture d'un tel système de RAP statistique.

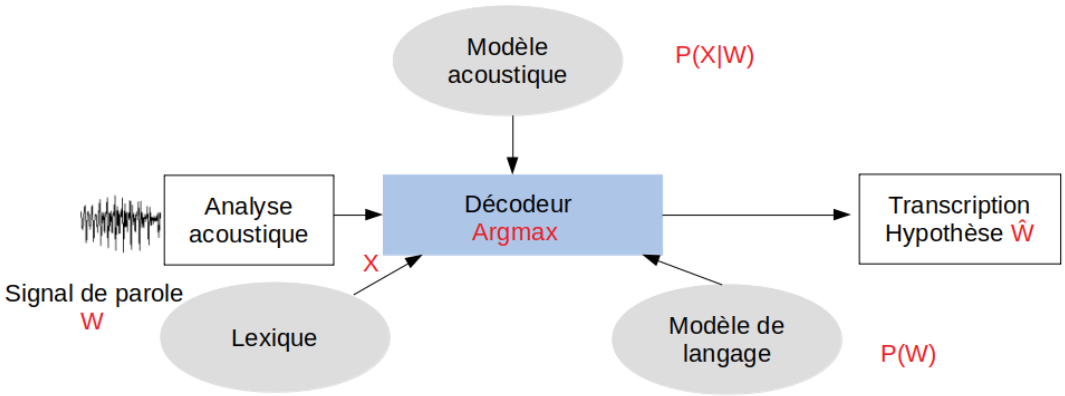


FIGURE 3.1 – RAP - architecture d'un système de RAP statistique

3.1.1.1 Pré-traitement acoustique : paramètres MFCC et bancs de filtres

Avant d'être traité par le système de RAP, le signal acoustique est converti en vecteurs de paramètres. L'extraction de ces paramètres consiste à produire une représentation vectorielle du signal de parole en utilisant des fenêtres glissantes (trames) dont la largeur est généralement 20 ms avec un chevauchement fixé en général à la moitié de la largeur de trame. Un filtrage est nécessaire sur chaque trame pour éviter un repliement de spectre, le filtre de *Hamming* est le plus généralement utilisé (Istrate, 2003). Les représentations vectorielles peuvent être les *Mel-Frequency Cepstral Coefficients* (MFCC) (Davis et Mermelstein, 1980), les *Perceptual Linear Prediction* (PLP) (Hermansky et Cox Jr, 1991) et les *Linear Prediction Ceptral Coefficients* (LPCC) (Markel et Gray, 2013). En outre, elles peuvent être enrichies de leurs dérivées premières Δ et secondes $\Delta\Delta$.

Les paragraphes suivants présentent une comparaison entre les paramètres MFCC et les bancs de filtres (fbank - *filter bank*). Le calcul des MFCC utilise une échelle fréquentielle non-linéaire appelée Mel qui tient compte de la sensibilité de l'oreille humaine (L'oreille humaine est moins sensible aux composantes à haute fréquence du son qu'à celles à basse fréquence (Davis et Mermelstein, 1980)). L'échelle de fréquence Mel est définie par la formule suivante :

$$B(f) = 2595 log\left(1 + \frac{f}{700}\right) \tag{3.2}$$

où f représente la fréquence exprimée en Hz et $B(f)$ la fréquence suivant l'échelle de fréquence Mel. Le processus complet de calcul des coefficients MFCC est présenté sur la figure 3.2.

Après le fenêtrage, une transformée de Fourier discrète (FFT - *Fast Fourier Transform* est appliquée aux échantillons de la fenêtre analysée. Le spectre du signal résultant est filtré par des filtres triangulaires (voir figure 3.3) dont les bandes passantes sont de même largeur

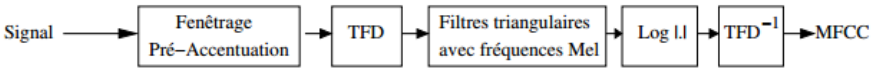


FIGURE 3.2 – Calcul des MFCC (Istrate, 2003)

selon l'échelle Mel. Les points de frontières B_m des filtres en échelle de fréquence Mel se calculent comme suit :

$$B_m = B_1 + m\frac{B_h - B_b}{M+1} \qquad 0 \leq m \leq M+1 \tag{3.3}$$

où M désigne le nombre de filtres, B_h la fréquence la plus haute et B_b la fréquence la plus basse du signal.

Dans le domaine fréquentiel, les points f_m discrets correspondants sont calculés d'après

$$f_m = \left(\frac{N}{F_s}\right) B^{-1}\left(B_b + m\frac{B_h - B_b}{M+1}\right) \tag{3.4}$$

où $B^{-1}(i)$ désigne la fréquence correspondante à la fréquence i de l'échelle Mel, $B_i^{-1} = 700\left(10^{\frac{i}{2595}} - 1\right)$

On calcule ensuite l'énergie totale en sortie de chacun des filtres triangulaires avant d'en prendre le logarithme. L'ensemble de ces résultats forme un vecteur fbank contenant le log de l'énergie pour chaque filtre. Une transformée en cosinus discrète (DCT - *Discrete Cosine Transform*) inverse, produit alors les coefficients MFCC.

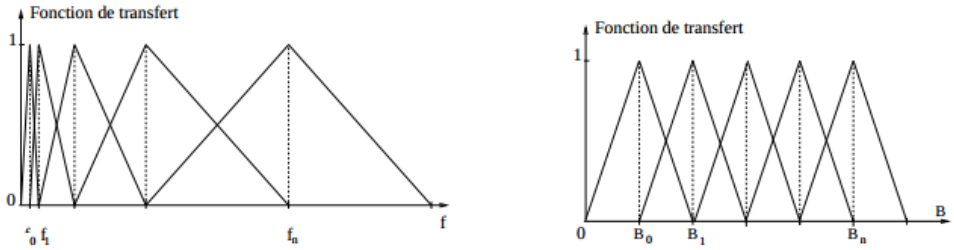


FIGURE 3.3 – Filtres triangulaires passe-bande selon une échelle mel B_f (sur la gauche) ou linéaire f (sur la droite) (Istrate, 2003)

Il en résulte une représentation compressée dont on ne conserve en général que les coefficients cepstraux 2-13, tandis que les autres coefficients sont rejetés.

Les paramètres MFCC ont longtemps été les plus employés, par exemple dans les approches de RAP GMM-HMM (sections 3.1.1.2.1 et 3.1.1.2.2). Cependant, plus récemment, les paramètres de fbank sont de plus en plus utilisés par les systèmes de RAP utilisant réseaux de neurones profonds. En effet, la différence des paramètres MFCC, dans le cas des filtres fbank on n'applique pas de DCT inverse et par conséquent les paramètres ne sont pas décorrelés. Or, avec l'avènement des réseaux de neurones profonds, le choix de paramètres MFCC n'est plus nécessaire car ces approches de réseaux sont peu sensibles aux entrées hautement corrélées (Abdel-Hamid et coll., 2012; Fayek, 2016).

3.1.1.2 Modèles acoustiques

Avant l'avènement des modèles réseaux de neurones profonds, la combinaison HMM et GMM étaient le modèle acoustique le plus fréquemment utilisé pour la reconnaissance automatique de la parole. Plus récemment sont apparu des systèmes hybrides qui combinent une approche HMM avec une approche de réseaux de neurones profonds. Le MA modélise une séquence de vecteurs de caractéristiques étant donné une séquence de *phonèmes* (et non de *mots*) bien que nous utilisions la notation $P(X|W)$ pour le MA. Les modèles de Markov cachés (HMM - Hidden Markov Model) (Jelinek, 1976; Rabiner, 1989) ont permis de faire des progrès considérables lorsqu'ils ont été mis en œuvre dans les systèmes de RAP.

3.1.1.2.1 Modèles de Markov Cachés (HMM)

Les HMM sont des automates probabilistes à états finis utilisés pour le calcul de la probabilité d'émission d'une séquence d'observations. Ces observations sont des paramètres acoustiques extraits d'un signal de parole. Comme la figure 3.4 le montre, un HMM se compose,

— d'un nombre d'états, $S = \{S_0, S_1, \ldots, S_N\}$, avec des états de début S_I et de fin S_E

— de probabilités de transition, $P(q_t = S_i | q_{t-1} = S_j) = a_{ji}$

— de distributions d'émission, à l'état j pour symbole x, $P(y_t = O_x | q_t = S_j) = b_j(x)$

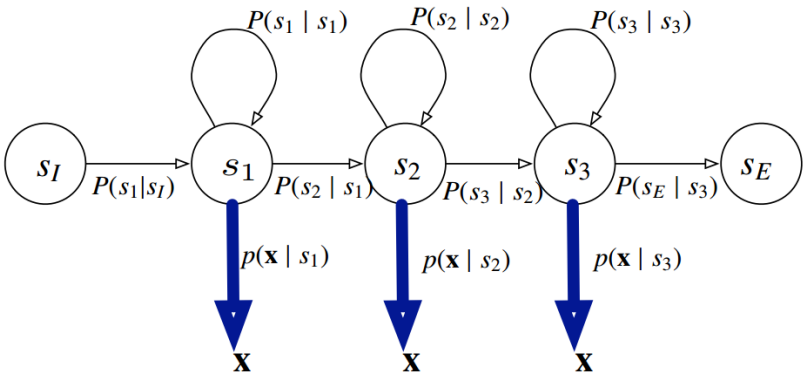


FIGURE 3.4 – HMM à 3 états (Renals et Hain, 2010)

Afin d'améliorer les performances, les unités de sous-mots en contexte peuvent être représentées comme des HMM triphone. La figure 3.5 montre un aperçu des triphones dans le cas du mot anglais "*rock*" (Virtanen et coll., 2012).

3.1.1.2.2 Les modèles de Markov cachés avec densité de probabilité à mélange de gaussiennes (HMM-GMM)

Un modèle à mélange de gaussiennes (GMM - Gaussian Mixture Model) est un modèle statistique, dont l'objectif est de représenter la densité de probabilité acoustique d'un état HMM comme une somme pondérée de K gaussiennes, $k \in 1, \ldots, K$. Chaque gaussienne d'index i est affectée d'un poids de mélange w_i, la somme des poids étant étant égale à 1 :

$$\sum_{i=1}^{K} w_i = 1 \tag{3.5}$$

La fonction de densité de chaque distribution gaussienne s'exprime comme

$$\mathcal{N}(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right) \tag{3.6}$$

où x représente le vecteur de données, D est le nombre de dimensions des vecteurs de données, μ_i et Σ_i sont respectivement le vecteur moyenne et la matrice de covariance de la distribution gaussienne.

Les modèles de mélange de gaussiennes peuvent donner une meilleure représentation des données. À gauche de la figure 3.6, une distribution gaussienne unique a été ajustée aux données en utilisant le maximum de vraisemblance (*Maximum Likelihood*). Cette distribution ne parvient pas à englober les deux clusters dans les données et place une grande partie de sa masse de probabilité dans la région centrale là où les données sont relativement clairsemées. Une distribution gaussienne simple est donc incapable de représenter cette structure. Par contre une superposition linéaire de deux gaussiennes, à droite sur la figure 3.6, permet une meilleure caractérisation de l'ensemble de données. De telles superpositions peuvent être formulées comme des modèles probabilistes appelés mélange de gaussiennes (Bishop, 2006). Les GMM sont donc souvent utilisés en combinaison avec des HMM comme densité de probabilité pour modéliser les phonèmes.

HTK (Young et coll., 2002) est un exemple d'un outil classique basé sur la méthode HMM-GMM. Julius (Lee et coll., 2001) est un outil similaire *Open Source* fonctionnant en temps réel. Il a permis une précision égale à 95% sur une tâche de dictée de 20K mots en japonais. L'outil de RAP HMM *Sphinx* (Walker et coll., 2004) comporte une architecture modulaire permettant le choix entre un ML n-gramme, mais aussi un ML CFG (*Context Free Grammar*). Pour une tâche avec vocabulaire de 60K mots, il obtient un WER de 3.8%. Pour un vocabulaire de la même taille, l'outil de RAP HMM *RWTH* (Rybach et coll., 2009) obtient un WER de 15%.

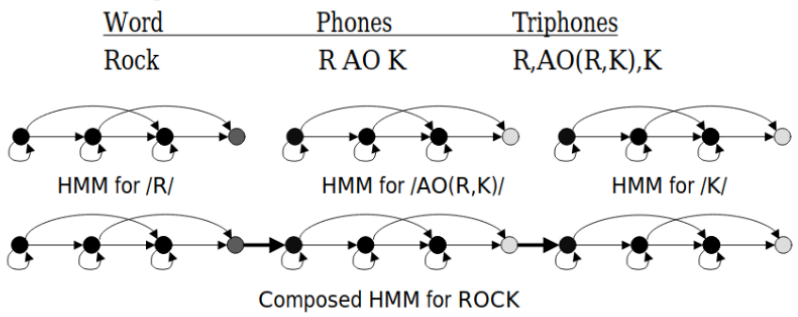


FIGURE 3.5 – HMM triphone : exemple du mot anglais "*rock*" (Virtanen et coll., 2012)

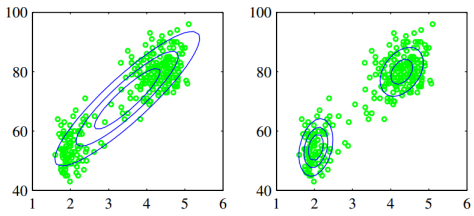


FIGURE 3.6 – Distributions gaussiennes : exemple de représentation d'un jeu de données par une ou 2 gaussiennes (Bishop, 2006)

3.1.1.3 Modèle de langage (ML)

Le ML permet au système de RAP de mieux décoder la séquence de mots la plus pertinente en estimant la probabilité $P(W)$. Les approches de ML les plus fréquemment utilisées sont les n-grammes et les réseaux de neurones. Un modèle à n-grammes est fondé sur l'hypothèse de Markov, et évalue la probabilité d'une phrase de longueur N comme

$$P(W) \approx \prod_{i=1}^{|W|+1} P(w_i|w_0,\ldots,w_{i-1}) \tag{3.7}$$

.

Un ML unigramme ignore le contexte, un modèle bigramme par contre, ajoute un mot de contexte, et un modèle de trigramme ajoute 2 mots de contexte,

$$P(w_i|w_0 \ldots wi-1) \approx P(w_i|w_{i-2}wi-1) \tag{3.8}$$

.

Pendant le décodage, il est bien possible que des séquences de mots non observées dans les données d'apprentissage du ML soient présentées au système de RAP. Plusieurs méthodes de *smoothing*, ou d'interpolation linéaire traitent ce problème (Good, 1953; Witten et Bell, 1991; Kneser et Ney, 1995). On peut comparer les ML générés en calculant leur perplexité. (Jelinek et coll., 1977). La perplexité est un indicateur de la capacité du modèle à prédire la bonne séquence de mots, elle est calculée comme suit

$$H = \frac{1}{n}\sum_{i=1}^{n} logP(s_i) \tag{3.9}$$

H est l'entropie pour un ensemble de données S de n phrases ($S = s_1, s_2, \ldots, s_n$), et la perplexité ($PPL$) est définie comme

$$PPL = 2^H \tag{3.10}$$

Si la valeur de la perplexité est trop élevée, cela signifie probablement que le ML est incertain. si cette valeur est par contre trop basse, le ML est probablement trop contraint.

3.1.1.4 Lexique phonétique

Le MA et ML sont souvent combinés avec un lexique dont la cible est d'associer à chaque mot du vocabulaire, les variantes de prononciation possibles, au niveau de sous-mots, sous

forme d'unités sonores telles que des syllabes, des graphèmes ou des phonèmes. La qualité du dictionnaire a un impact sur les performances du système de RAP. Si la phonétisation d'un mot est erronée ou s'il s'agit d'un mot hors vocabulaire, cette erreur peut se propager aux mots voisins. Pour la langue française, *BDLEX* (Perennou, 1986) est un dictionnaire de prononciation souvent utilisé pour l'apprentissage des modèles acoustiques. Il contient environ 440k formes fléchies, générées à partir de 50k mots.

3.1.2 Réseaux de neurones artificiels

Au cours des dix dernières années, plusieurs projets ont été orientés vers l'utilisation des réseaux de neurones artificiels (ANN - *Artificial Neural Networks*). Les ANN ont des structures de traitement massivement parallèle, ce qui les rend appropriées pour des implémentations de haute performance (Haton, 1999; Haton et coll., 2006). Le modèle ANN s'est inspiré des données neurobiologiques sur le cortex humain qui se compose d'un grand nombre de neurones. Un neurone est une cellule biologique capable de traiter des informations en raison des interconnexions complexes avec d'autres neurones. McCulloch et Pitts (1943) ont proposé un modèle formel d'un neurone sous la forme d'une unité de seuil binaire. Ce modèle calcule une somme pondérée de ses entrées qui est injectée en entrée à une fonction d'activation. Le neurone délivre une sortie binaire égale à 1 si la sortie de cette fonction dépasse un seuil prédéterminé. Les ANN utilisés actuellement sont basés sur ce modèle initial, ils sont organisés en couches et reliés entre eux. On distingue deux grandes catégories :

— des réseaux *Feedforward*, dans lesquels aucune boucle n'existe dans les connexions entre les neurones,

— des réseaux récurrents (RNN), qui contiennent des boucles entre les neurones correspondant aux connexions de *Feedback*. La section 3.2.6 explique ce dernier modèle en détail.

La figure 3.7 montre un ANN à cinq couches composé d'une couche d'entrée, de trois couches cachées et d'une couche de sortie. Nous désignons la couche d'entrée comme couche 0 et la couche de sortie comme couche L. Cette dernière couche permet d'estimer les probabilités correspondant à une observation acoustique.

L'équation suivante définit la sortie de chaque couche $l \in \{1,\ldots,L\}$ comme

$$Z_l = \varphi\left(W_l \cdot X + b_l\right) \tag{3.11}$$

où φ est une fonction d'activation permettant de transférer les sorties d'une couche à la suivante (de la première couche jusqu'à la couche $L-1$), b_l un vecteur de biais, W_l une matrice de poids et X le vecteur d'entrée (Yu et Deng, 2016). La dernière couche délivre la probabilité *softmax* pour chaque état j qui est définie par l'équation suivante :

$$P\left(j|x_t\right) = \frac{e^{Z_L}}{\sum_{j=1}^{C} e^{Z_L}}; j \in \{1,\ldots,N\} \tag{3.12}$$

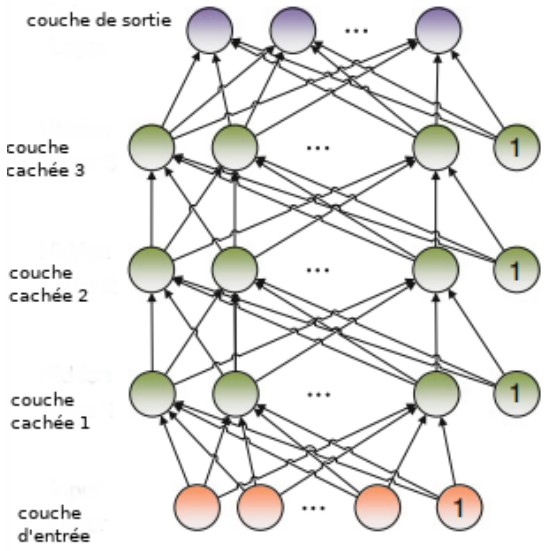


FIGURE 3.7 – Exemple de réseau ANN (Yu et Deng, 2016)

Les fonctions d'activation les plus utilisées sont *sigmoid()*, *tanh()* et *ReLU()* (figure 3.8). La fonction sigmoid est en forme de S. Comme sa sortie est comprise entre 0 et 1, elle est utilisée pour les probabilités :

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{3.13}$$

La fonction tangente hyperbolique est aussi en forme de S, mais sa sortie est comprise entre -1 et 1.

$$tanh(z) = \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \tag{3.14}$$

La fonction d'activation *Rectified Linear Unit* (ReLU) est la plus utilisée. Sa sortie est zéro lorsque z est inférieur à zéro et égale à z lorsque z est supérieur ou égal à zéro.

$$ReLU(z) = max(0,z) \tag{3.15}$$

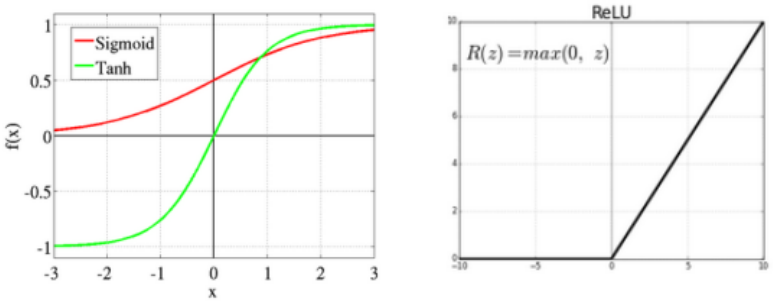


FIGURE 3.8 – Fonctions d'activation sigmoid(), tanh() et ReLU() d'un neurone artificiel

3.1.3 Kaldi, combinaison de RAP HMM et ANN

Kaldi (Povey et coll., 2011a) est une boite à outils de RAP *Open Source* développé en C++. Kaldi est basé sur des transducteurs à états finis (FST - *Finite State Transducer*), il contient le code informatique nécessaire pour entraîner et décoder les différents MA statistiques à l'état de l'art, qu'ils soient de type GMM ou DNN (réseaux de neurones profonds - *Deep Neural Network*) (nnet2 et nnet3). Les paramètres acoustiques utilisés par défaut sont MFCC et PLP. Kaldi permet également d'utiliser plusieurs méthodes d'adaptation des modèles acoustiques au locuteur, comme la *Maximum Likelihood Linear Regression* (MLLR) (Leggetter et Woodland, 1995), la *Constrained Maximum Likelihood Linear Regression* (fMLLR) (Digalakis et Neumeyer, 1996) et le *Speaker Adaptive Training* (SAT) (Anastasakos et coll., 1996). En utilisant la librairie OpenFST on peut exploiter les FST en choisissant parmi les différents modèles acoustiques possibles, en définissant le ML et le modèle de prononciation pour construire un système de RAP.

Kaldi peut aussi combiner des ANN avec des HMM. En utilisant la version Kaldi DNN (*Deep Neural Network*) nnet2, les ANN estiment les probabilités postérieures d'un état HMM, les sorties des ANN sont alors utilisées comme paramètres d'entrée d'un modèle HMM-GMM (figure 3.9). L'avantage de cette approche hybride DNN est qu'elle généralise mieux et offre la possibilité d'apprendre un modèle en utilisant des processeurs graphiques (GPU - *Graphics Processing Unit* (Povey et coll., 2015; Kipyatkova et Karpov, 2016).

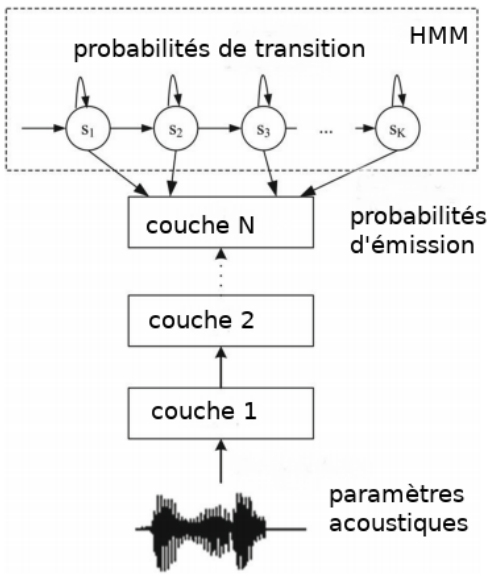


FIGURE 3.9 – Architecture DNN-HMM mise en œuvre par Kaldi

3.1.4 Une approche de bout en bout de la RAP (End to end - E2E)

Contrairement aux systèmes de RAP traditionnels dont Kaldi est une des réalisations les plus récentes (section 3.1.3), un système de Reconnaissance Automatique de la Parole de bout en bout (RAP E2E) n'a pas besoin d'un lexique ou d'un dictionnaire de phonèmes, le Modèle de Langage (ML) n'est plus indispensable mais seulement facultatif. Typiquement,

ces systèmes sont construits à partir de réseaux de neurones profonds utilisant plusieurs GPU avec un parallélisme dans la gestion des données, ce qui permet d'entraîner directement à partir de grandes masses de données, celle-ci pouvant aller jusqu'à plusieurs milliers d'heures de parole. Cependant, la RAP E2E reste encore limitée par rapport à la RAP traditionnelle. Cette dernière, en associant un MA, un dictionnaire de prononciation et un ML, conduit pour l'instant à des performances qui surpassent celles des approches de RAP E2E.

Les prédécesseurs des systèmes de RAP E2E étaient des approches hybrides HMM ANN. Renals et coll. (1994) ont proposé un système hybride HMM DNN dont l'architecture comprend une couche d'entrée, plusieurs couches cachées et une couche de sortie softmax. L'ensemble d'apprentissage était le *DARPA Resource Management speaker-independent continuous speech database* (Price et coll., 1988) qui contenait 3990 énoncés alors que l'ensemble d'évaluation contenait 600 énoncés provenant du même corpus. Cependant cette approche ne surpassait pas les performances de leur approche RAP HMM de référence.

Robinson et coll. (1996) ont introduit l'utilisation des RNN dans les modèles de RAP. Le corpus *TIMIT* speech database (Zue et coll., 1990) était utilisé pour créer les ensembles d'apprentissage et d'évaluation. Par rapport à un modèle HMM classique, le modèle RNN est capable de modéliser un contexte acoustique à long terme. Ils présentent des performances RAP équivalentes à celles d'un modèle HMM classique. L'approche de RAP E2E, intégrant des CNN (convolutional neural networks) de Sainath et coll. (2013a) surpasse les approches hybrides HMM DNN en utilisant un modèle, appris sur 400 heures de parole anglaise et un ensemble d'évaluation de 50 heures du même ensemble de données. Les mêmes ensembles d'entraînement et d'évaluation ont été utilisés pour une approche de RAP E2E RNN bidirectionnel LSTM (Graves et coll., 2013) mais cela n'a pas permis de surpasser les performances du modèle de CNN présenté par Sainath et coll. (2013a).

Dans les paragraphes suivants, nous décrivons deux systèmes de RAP E2E, Deep Speech (Hannun et coll., 2014) et ESPnet (Watanabe et coll., 2018) ; ce sont des systèmes de RAP E2E d'apprentissage multi-tâche utilisant la classification temporelle connexionniste (CTC - *Connexionist Temporal Classification*) et font correspondre les trames du signal de parole à des suites de caractères (Graves et coll., 2006; Watanabe et coll., 2017; Ueno et coll., 2018)

3.1.4.1 L'outil Deep Speech

Le noyau de Deep Speech (Hannun et coll., 2014) est un RNN composé de cinq couches de neurones cachées comme illustré parla figure 3.10. Les trois premières, ${h_t}^{(1)}$,${h_t}^{(2)}$,${h_t}^{(3)}$ ne sont pas récurrentes. La quatrième est une couche récurrente bi-directionnelle, dont ${h_t}^{(f)}$ et ${h_t}^{(b)}$ sont respectivement les couches en avant (*forward*) et en arrière (*backward*). La première couche prend en entrée les spectrogrammes des trames et prend en compte le contexte (trame précédente et trame suivante). Les deuxième et troisième couches traitent les données indépendamment pour chaque pas de temps. Les entrées de la cinquième couche non récurrente, ${h_t}^{(5)}$, sont les unités avant et arrière. La couche en sortie utilise une fonction *softmax* permettant de prédire les probabilités des caractères à chaque pas de temps.

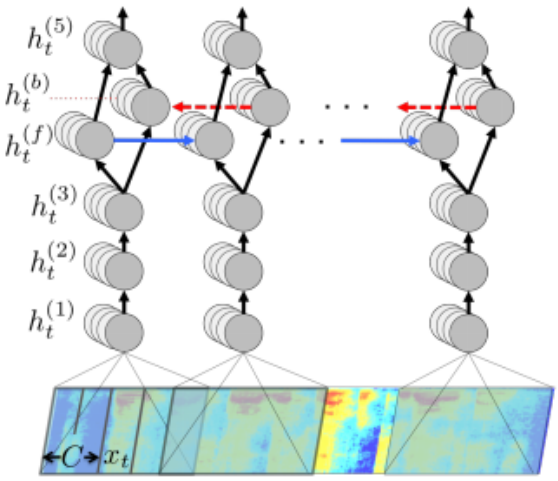


FIGURE 3.10 – Architecture de Deep Speech (Hannun et coll., 2014)

L'apprentissage d'un ML n-gramme basé sur des caractères est intégré au système mais reste facultatif. Des poids peuvent être définis pour obtenir la séquence c qui soit le meilleur compromis entre RNN, ML (α) et longueur de la phrase (β) grâce à une méthode de recherche par faisceau (*Beam Search*) :

$$Q(c) = log(P(c|x)) + \alpha\, log(P_{ml}(c)) + \beta \text{word\_count}(c) \tag{3.16}$$

où P_{ml} est la probabilité de la séquence c conformément au ML.

Afin de traiter efficacement de grandes quantités de données, la méthode utilise des lots plus grands que ceux qu'un seul GPU peut prendre en charge, ce qui permet de traiter les données en parallèle sur plusieurs GPU. Les exemples d'entraînement sont triés par longueur. Les phrases d'une taille égale apparaissent dans le même lot et sont complétées avec du silence si nécessaire afin de cibler des phrases de longueur égale par lot.

Certaines performances qui ont été obtenues avec Deep Speech sur la langue anglaise ont été publiées. Les données d'entraînement correspondantes sont constituées de 5000 heures de parole prononcées par 9600 locuteurs du *The Wall Street Journal*, de *Switchboard* et du corpus *Fisher* (Cieri et coll., 2004). L'ensemble de validation se compose de 300 heures de conversation téléphonique (Godfrey et coll., 1992). Les performances de Deep Speech sont comparées à deux systèmes de RAP hybrides DNN-HMM de référence (Hannun et coll., 2014). Les performances de ces 2 systèmes de référence, intégrant Kaldi (Povey et coll., 2011b), sont présentées dans le tableau 3.1. Les performances de Deep Speech dépassent donc celles des deux modèles de référence.

TABLE 3.1 – Comparaison des performances de Deep Speech et de Kaldi sur de la parole téléphonique en langue anglaise (Hannun et coll., 2014)

Modèle	WER (%)
Kaldi HMM-DNN(1)	19,9
Kaldi HMM-DNN(2)	18,4
Deep Speech	16

3.1.4.2 L'outil ESPnet

Au contraire de Deep Speech, ESPnet (Watanabe et coll., 2018) intègre la préparation de données et l'extraction de paramètres acoustiques de Kaldi. On peut sélectionner *Chainer* et *Pytorch* comme environnement d'ESPnet (figure 3.12). Chainer est un outil *Open Source* basé sur Python pour les modèles de réseaux de neurones profonds. Chainer mémorise le graphe de calcul implicitement avant le calcul de l'ensemble de données d'apprentissage (Tokui et coll., 2015). Pytorch s'appuie sur Chainer, et fournit un environnement de haute performance pour des modèles exécutés sur CPU et GPU (Paszke et coll., 2017). La figure 3.11 présente un aperçu de l'architecture ESPnet.

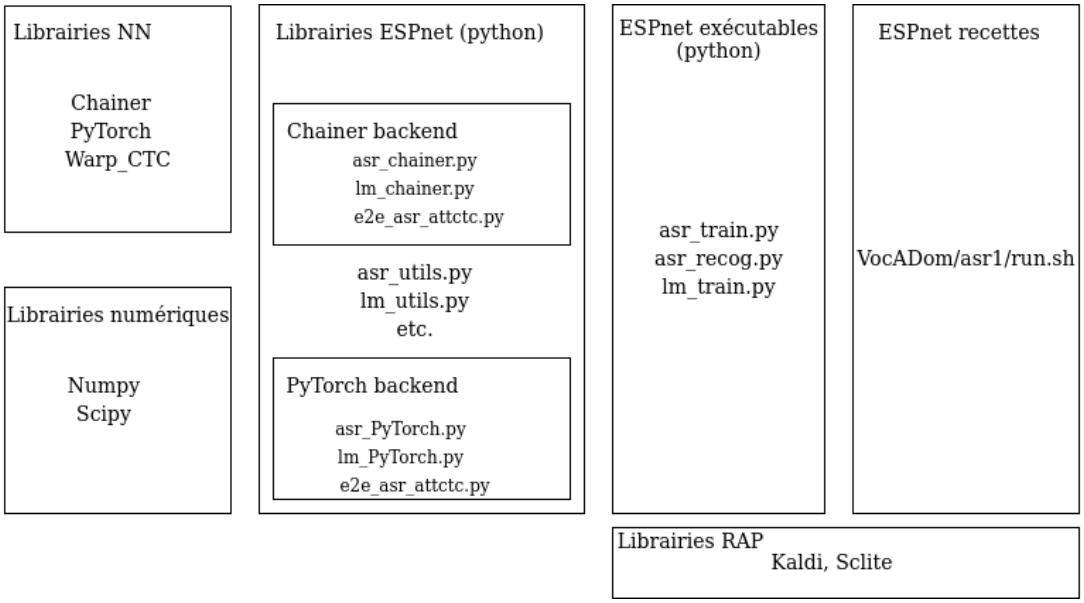


FIGURE 3.11 – Architecture logicielle de ESPnet (Watanabe et coll., 2018)

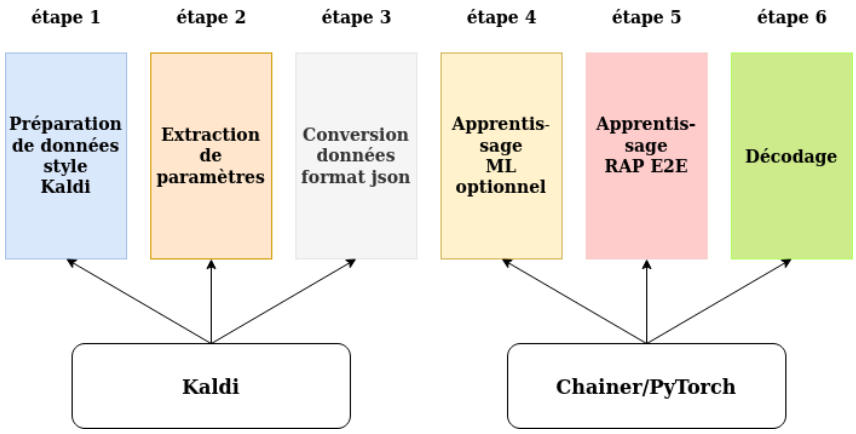


FIGURE 3.12 – Enchaînement des tâches de ESPnet (Watanabe et coll., 2018)

L'encodeur PyTorch par défaut effectue un sous-échantillonnage pyramidal bi-LSTM (BLSTM) (Chan et coll., 2016), étant donné la séquence de paramètres acoustiques de longueur T $o_{1:T}$, pour extraire la séquence de paramètres $h_{1:T'}$:

$$h_{1:T'} = \text{BLSTM}(o_{1:T}), \tag{3.17}$$

où $T' < T$ en raison du sous-échantillonnage. Le *back-end* de Chainer prend en charge les réseaux de neurones convolutifs (*Convolutional Neural Networks - CNN*). La correspondance entre paramètres acoustiques et séquences de caractères est effectuée grâce à un apprentissage *hybride* multitâche qui combine la *classification temporelle connexionniste* (Connectionst Temporal Classification - CTC) *Warp_CTC* (figure 3.11) (Amodei et coll., 2016) avec un encodeur-décodeur basé sur *l'attention* (*e2e_asr_attctc_th.py*). Les paragraphes 3.1.4.2.1 et 3.1.4.2.2 décrivent plus en détail la méthode CTC et le mécanisme d'attention.

Le mécanisme d'attention permet un alignement plus souple, qui se concentre sur les paramètres importants et les séquences de caractères tandis que l'alignement de la RAP est monotone. Il en résulte un équilibre entre l'attention et le CTC comme le montre l'expression des scores :

$$\begin{aligned} log\, p^{hyb}(y_n|y_{1:n-1}, h_{1:T'}) = \alpha\, log\, p^{ctc}(y_n|y_{1:n-1}, h_{1:T'}) \\ + (1-\alpha)\, log\, p^{att}(y_n|y_{1:n-1}, h_{1:T'}), \end{aligned} \tag{3.18}$$

où y_n désigne une hypothèse d'étiquette de sortie à la position n étant donné $y_{1:n-1}$ et la sortie de l'encodeur $h_{1:T'}$. La combinaison de scores ($log\, p^{hyb}$) pour l'architecture hybride CTC/attention, avec les probabilités d'attention p^{att} et CTC p^{ctc} log, est effectuée pendant la recherche de faisceau. Le poids α (à spécifier dans *asr_train.py*) peut être défini manuellement afin d'attribuer plus d'importance à l'attention ou au CTC.

Un ML RNN de caractères peut être fourni pour le décodage. La probabilité logarithmique p^{ml} du ML RNN peut être fusionnée avec la sortie hybride d'attention CTC par :

$$\begin{aligned} log\, p(y_n|y_{1:n-1}, h_{1:T'}) = log\, p^{hyb}(y_n|y_{1:n-1}, h_{1:T'}) \\ + \beta log\, p^{ml}(y_n|y_{1:n-1}). \end{aligned} \tag{3.19}$$

Le poids β (à spécifier dans *asr_train.py*) peut être défini manuellement afin d'attribuer plus d'importance au ML. L'interaction entre le CTC, le réseau d'attention et le ML est illustrée dans la figure 3.13, pour la séquence de paramètres acoustiques de longueur T, $X = (x_t|t = 1, \ldots, T)$.

Pour les tâches CSJ japonais (Maekawa, 2003) et HKUST en chinois mandarin (Liu et coll., 2006), deux langues où chaque caractère représente une unité sémantique, les performances d'ESPnet surpassent celles du modèle HMM DNN Kaldi, comme le montre le tableau 3.2. Dans les sections suivantes, nous illustrons les mécanismes de CTC et d'attention.

TABLE 3.2 – ESPnet - évaluation des performances WER en comparaison avec Kaldi (Watanabe et coll., 2018)

Tâche	Kaldi HMM-DNN (%)	ESPnet (%)
CSJ japonais	7.2	6.1
HKUST chinois	33.5	28.3

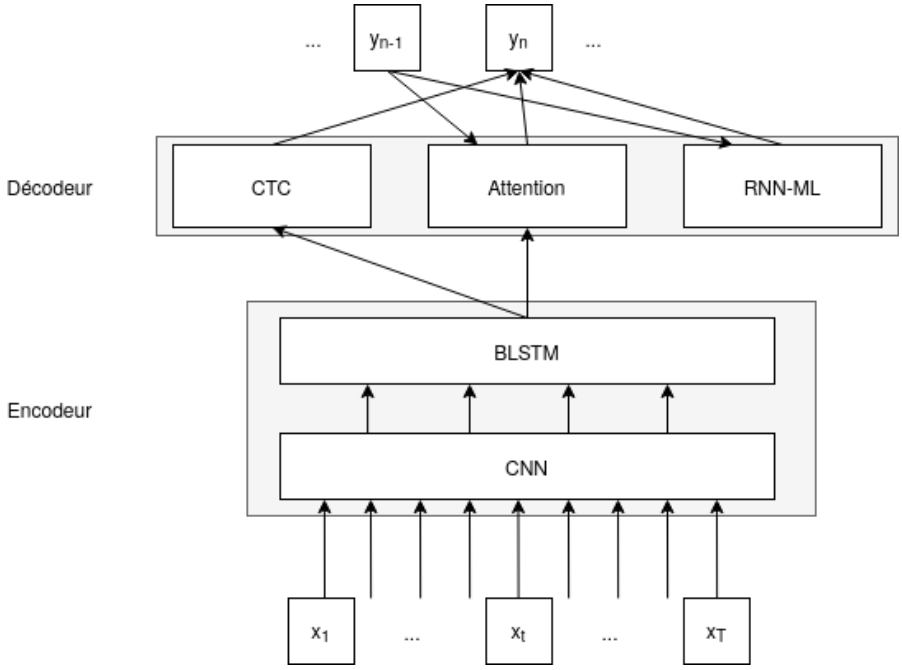


FIGURE 3.13 – ESPnet - CTC, attention et ML RNN

3.1.4.2.1 Classification Temporelle Connexionniste CTC

Dans le contexte de RAP, la Classification Temporelle Connexionniste (*Connectionist Temporal Classification* - CTC) est un mécanisme d'alignement entre paramètres acoustiques d'entrée et une chaîne de caractères de sortie. La CTC supprime le besoin de données d'entraînement pré-segmentées. Comme les séquences d'entrée et les séquences cibles de RNN n'ont généralement pas la même longueur, la CTC ajoute une étiquette spéciale, « non-caractère » (*blank*). Considérons une séquence sortie cible d'étiquettes de caractères $Y = [y_1, y_2, y_3, \cdots, y_M]$ (de longueur M) selon des pas de temps t. La séquence Y, par exemple « hello » (bonjour), peut être mise en correspondance avec une séquence d'entrée de trames acoustiques $X = [x_1, x_2, x_3, \cdots, x_N]$ de (longueur N) par insertion d'une étiquette « vide » notée ϵ. Des étiquettes identiques consécutives sont ensuite fusionnées en une seule étiquette (Graves et coll., 2006). L'alignement des trames d'entrée X sur les étiquettes de sortie Y est multiple, un ou plusieurs éléments d'entrée peuvent être alignés sur une seule sortie. Considérons les exemples de séquences suivants :

— trames d'entrée $X = [h, e, e, e, l, \epsilon, l, l, o]$,

— caractères de sortie $Y = [h, e, l, l, o]$,

$P(Y|X)$ délivre une distribution de probabilité sur la séquence de sortie « hello » pour chaque étape d'entrée. La CTC additionne la probabilité de tous les chemins d'alignement possibles A_t de la séquence d'entrée avec celle de sortie. Dans l'équation ci-dessous, la probabilité conditionnelle de la CTC somme celle de tous les alignements valides, comme le montre également la figure 3.14.

$$P(Y|X) = \sum_{A \in A_{X,Y}} \prod_{t=1}^{T} p_t(A_t|X) \tag{3.20}$$

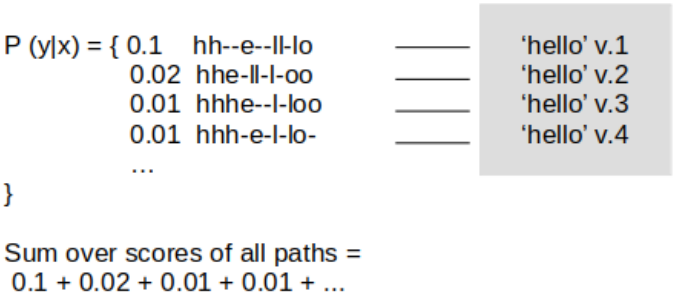


FIGURE 3.14 – Description du fonctionnement de la CTC sur un exemple

3.1.4.2.2 Mécanisme d'attention

Pour éviter le problème de la disparition du gradient (*Vanishing Gradient*) et pour modéliser de longues dépendances dans une séquence, des unités LSTM (*Long Short-Term Memory*)ou GRU ((*Gated Recurrent Unit*) - réseau récurrent à portes) sont utilisées comme unités de base de RNN. Pour permettre au décodeur de baser sa prédiction non seulement sur le mot précédent et l'état caché, mais aussi sur les états cachés de la séquence d'entrée, le mécanisme d'attention a été introduit par Bahdanau et coll. (2015, 2016).

Dans ce cas, le décodeur utilise une autre information lors du décodage à savoir le vecteur de contexte c. À chaque étape i et en fonction de la longueur de séquence d'entrée T_x :

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j. \tag{3.21}$$

Le poids α_{ij} est calculé comme suit :

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}, \tag{3.22}$$

où e_{ij} est exprimé par :

$$e_{ij} = a(s_{i-1}, h_j). \tag{3.23}$$

e_{ij} représente un modèle d'alignement ou d'attention qui indique au décodeur à l'étape i à quelle partie de l'état caché de la séquence d'entrée attribuer l'attention.

Le modèle d'alignement a peut être un RNN *feed-forward* appris conjointement avec le reste de l'architecture. La probabilité α_{ij}, reflète l'importance de h_j par rapport à l'état caché précédent $i-1$ du décodeur pour décider de l'état suivant i et générer la sortie. Par conséquent, le décodeur décide des parties de la phrase source à prendre en compte afin de leur attribuer de l'attention, ceci est particulièrement utile lorsque la sortie à générer dépend d'un mot d'entrée assez éloigné dans la séquence d'entrée.

3.1.4.2.3 Réseaux de neurones convolutifs

Des réseaux de neurones convolutifs (CNN - *Convolutional Neural Networks*) sont intégrés dans l'architecture de ESPnet. Les CNN ont permis une énorme avancée dans le domaine de la reconnaissance d'images mais ils sont maintenant de plus en plus utilisés dans le contexte de la RAP. Ils ont été initialement combinés avec des HMM/GMM, ce qui a permis

de construire des systèmes de RAP DNN hybrides (Zhang et coll., 2016; Cho et coll., 2018). Le succès de l'intégration des CNN dans les systèmes de RAP E2E peut être attribuée à l'utilisation d'un *filtrage* local et du mécanisme de *maxpool* dans l'architecture CNN. La combinaison du filtrage et de *maxpool* sur des plages de fréquences s'avère être une meilleure stratégie que la représentation entière du spectre de fréquences, comme c'est le cas de manière classique pour les modèles GMM. Cela permet une meilleure robustesse lors du traitement de la parole en présence de bruit, ainsi que de bonnes performances avec des ensembles de données d'une taille réduite (Qian et coll., 2016). Cela peut également constituer une alternative au SAT (Speaker Adaptive Training) et au MLLR (Maximum Likelihood Linear Regression) comme techniques d'adaptation au locuteur (Abdel-Hamid et coll., 2012) qui transposent les paramètres de parole dans un espace locuteur canonique.

Les paramètres MFCC ne conviennent pas car ils sont le résultat d'une DCT et ne permettent donc pas d'applique un filtrage local . Pour cette raison, ce sont les paramètres *fbank* qui sont utilisés car ils permettent un filtrage local. L'augmentation du nombre de couches CNN peut améliorer les performances de RAP (Sainath et coll., 2013b).

3.2 Compréhension du langage naturel (NLU)

Au cours des sections précédentes, nous avons donné un aperçu de l'état de l'art de RAP car le *premier* module d'une approche SLU séquentielle est un outil de RAP. Les transcriptions d'hypothèse de sortie du module de RAP sont les phrases d'entrée du deuxième module qui opère la NLU. C'est pourquoi nous décrivons également l'état de l'art concernant la compréhension du langage naturel (NLU). Les sections suivantes donnent un aperçu des modèles de NLU et comparent leurs performances pour faire ressortir les qualités et les lacunes de chacunes.

3.2.1 Approche par règles

Les premiers systèmes de NLU étaient basés sur des règles sémantiques définies manuellement (Wang et coll., 2011). Ils effectuaient un remplissage d'attributs (*Slot Filling*) ou effectuaient simplement une reconnaissance des motifs (*Pattern Matching)* pour poursuivre une conversation. Une telle approche est appliquée dans l'agent conversationnel *ELIZA*, un des premiers systèmes de dialogue homme-machine (Weizenbaum, 1966). ELIZA simule un psychothérapeute en reformulant la plupart des affirmations du patient sous forme de questions. Les phrases d'entrée sont analysées grâce à des règles de décomposition qui sont déclenchées par des mots-clés. Les phrases sont ensuite modifiées grâce à des règles de transformation. Il s'agit par exemple de règles qui changent les pronoms personnels pour qu'ils soient adaptés à la bonne réponse (Shawar et Atwell, 2002). Les défis principaux d'un tel système sont l'identification des mots-clés en tenant compte du contexte, l'application des règles de transformation, et la génération de réponses lorsqu'il y a des mots-clés qui manquent.

Un autre agent conversationnel basé sur règles par reconnaissance des motifs est *ALICE* (Wallace, 2009), qui utilise des modèles de motif (*Templates*) pour des phrases d'entrée *et* de sortie. Une technique de dialogue stimulus-réponse est appliquée. Le système répond (réponse) à la dernière question (stimulus) de l'utilisateur. Le stimulus est ensuite lié à une base de connaissance de l'agent conversationnel (Serban et coll., 2015). Chaque stimulus et chaque réponse font partie d'une catégorie. Ces catégories sont stockées dans une arborescence qui est structurée pour produire des réponses plus complexes.

Une approche à base de règles a été mise en œuvre par *TINA* qui procède en ajoutant des caractéristiques (*Features*) sémantiques aux caractéristiques syntaxiques d'une grammaire hors contexte (Seneff, 1992). Un ensemble de règles est d'abord converti en réseau probabiliste. Les probabilités sur tous les arcs du réseau sont calculées à partir d'un ensemble de phrases analysées au niveau syntaxique. Après l'analyse syntaxique, des caractéristiques sémantiques sont ajoutées. Un autre type de contrainte est l'unification. Ce mécanisme introduit des contraintes syntaxiques et sémantiques telles que l'accord entre nom, personne, sujet et verbe. La figure 3.15 montre l'exemple d'un réseau probabiliste obtenu à partir des règles suivantes pour une phrase nominale (*Noun Phrase*, NP), les parenthèses entourent la partie d'une règle qui est facultative :

[NP] ⇒ [article] ([adjective]) ([adjective]) [noun]

"the boy (le garçon)" [NP] ⇒ [article] [noun]
"a beautiful town (une belle ville)" [NP] ⇒ [article] [adjective] [noun]
"a cute little baby (un bébé mignon et petit)" [NP] ⇒ [article] [adjective] [adjective] [noun]
"the wonderful pudding (du pudding merveilleux)" [NP] ⇒ [article] [adjective] [noun]

Le nœud parent, la phrase nominale ([NP]), contient les cinq nœuds, "start", "article", "adjective", "noun", comme ses enfants avec ses probabilités.

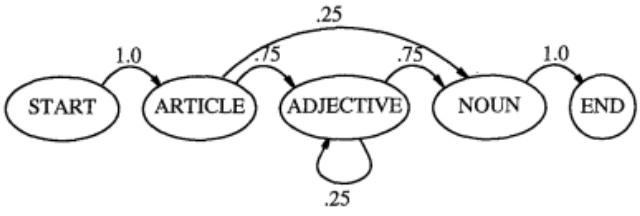


FIGURE 3.15 – TINA - réseau de probabilités (Seneff, 1992)

Le système a été évalué en utilisant la mesure de complexité :

$$-\frac{1}{N}\sum_{i=1}^{N} log_2 P(w_i|w_{i-1},...w_1) \tag{3.24}$$

Étant donné la séquence de mots w_i, le nombre total de mots N, $P(w_i|w_{i-1},...w_1)$ est la probabilité du mot i donné tous les mots précédents. Un ensemble de 791 phrases de la

TABLE 3.3 – TINA - évaluation par la complexité (%) (Seneff, 1992)

Taille de vocabulaire	Complexité sans probabilités	Complexité avec probabilités
985	368	41.5

TABLE 3.4 – GEMINI - évaluation des performances de NLU (Précision %) (Dowding et coll., 1993)

Source	Dev. (%)	Test (%)
Sans correction de disfluences	94.2	90.9
Avec correction de disfluences	96.0	93.1

tâche *RM* (*Resource Management Task*) (Pallett, 1989) ont été sélectionnées comme phrases d'apprentissage, et un ensemble de 200 phrases comme ensemble de test. Une grammaire *avec* et *sans* probabilités a été construite à partir de l'ensemble d'apprentissage. Le tableau 3.3 montre que la complexité est plus basse pour une structure de réseau probabiliste que pour une grammaire sans probabilités.

Le système *GEMINI* (Dowding et coll., 1993) utilise lui aussi une approche par règles comme TINA mais ajoute également des caractéristiques sémantiques aux caractéristiques syntaxiques. Cette approche de NLU se concentre sur la parole spontanée, en incluant des règles visant la correction des disfluences. Ses règles détectent par exemple des séquences de plusieurs mots identiques répétées et les dédouble. GEMINI a été entraîné sur un ensemble de données de 5875 énoncés du corpus ATIS (Hemphill et coll., 1990), avec 688 autres énoncés utilisés comme ensemble de test. Le tableau 3.4 compare les performances sans et avec règles de correction.

Bien que les systèmes basés sur des règles montrent de bonnes performances, leur principal désavantage est que leur construction est longue et laborieuse en termes d'efforts humains. En plus, ces règles sont très spécifiques et liées aux domaines des applications pour lesquelles elles ont été écrites et manquent de robustesse aux erreurs et irrégularités. Un autre obstacle à surmonter par les systèmes de NLU par règles est le traitement des mots hors vocabulaire. En intégrant un système de NLU dans une approche de SLU, nous sommes confrontés aux difficultés entraînées par l'oral spontané et les erreurs introduites par la RAP. C'est pourquoi l'utilisation de systèmes statistiques constitue une alternative.

3.2.2 Approche statistique

Une approche statistique cherche à trouver un sens pour une suite de mots d'entrée, ou à étiqueter les mots avec leurs concepts. Ces mots d'entrées sont d'abord convertis par une représentation qui la rende utilisable par un modèle statistique, notamment les sacs de mots (*Bag-of-Words*) ou la représentation vectorielle de mots (Word Embeddings) qui sont des vecteurs qui sont dérivés de la table de co-ocurrence des mots. Le but est de caractériser les mots par leur contexte (Mikolov et coll., 2013).

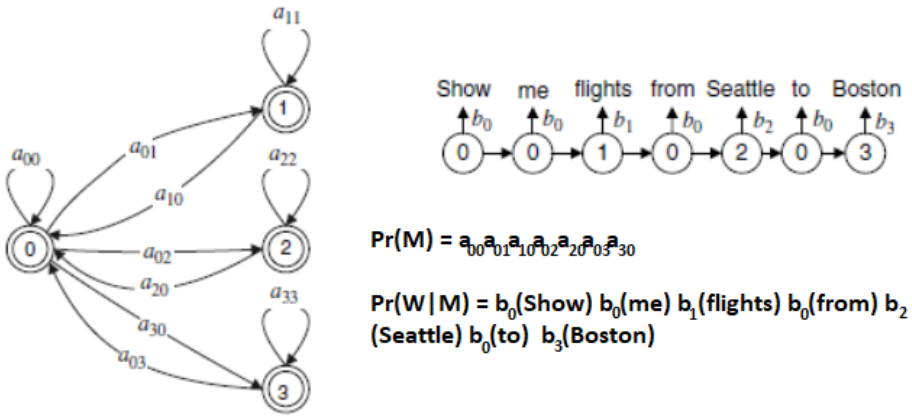


FIGURE 3.16 – Illustration du principe du fonctionnement du modèle génératif HMM

3.2.2.1 Les modèles HMM

Étant donné la suite de mots W, la représentation sémantique de la signification M est obtenue en maximisant la probabilité jointe à plusieurs variables $P(M,W)$ qui s'exprime de la manière suivante :

$$P(W,M) = P(W|M).P(M) \tag{3.25}$$

où le modèle de lexicalisation $P(W|M)$ exprime la probabilité de la la séquence des mots W étant donné la structure sémantique M, le préalable sémantique $P(M)$ attribue lui la probabilité a priori à cette structure sémantique. La figure 3.16 illustre les HMM présenté par Tur et De Mori (2011) :

— Les états 0, 1, 2 et 3 représentent les concepts auxquels un sens a été attribué.

— L'état 2 représente le concept *FromCity* (la ville de laquelle on veut partir).

— a_{00}, a_{01} ... etc. sont les probabilités de transition d'un connecteur à l'autre et font partie du modèle sémantique *a priori*.

— b_0, b_1 ... etc. sont les probabilités d'émission car il s'agit de l'émission des états observables (les mots) à partir des états cachés (qui représentent le sens). Par exemple, le mot *Seattle* est lié au concept *FromCity*.

L'étude de Levin et Pieraccini (1995) décrit *CHRONUS*, un système de NLU, comme composant d'un système de SLU qui utilise des HMM. Ce système de SLU se compose d'un module de RAP, de NLU et un module de prise de décision, qui convertit la représentation sémantique en l'action souhaitée. Le noyau du module de NLU est un modèle statistique dont les paramètres sont appris à partir du corpus *Air Travel Information Service* (ATIS). La signification d'une demande est représentée comme un modèle composé de paires mot-clé/valeur. La figure 3.17 montre un modèle pour une phrase comme "je voudrais aller de New York à San Francisco samedi matin et je préfère voler sur un Boeing 747". Cette approche statistique permet de gérer les ambiguïtés sémantiques. Dans la phrase d'exemple précédente, 747 est l'identifiant d'une avion mais peut aussi bien être un numéro de vol.

```
AIRLINE: UA
ORIGIN_CITY: NNYC
DESTINATION_CITY: SSFO
WEEKDAY: SATURDAY
ORIGIN_TIME: 0<1200
AIRCRAFT: 74M
SUBJECT: FLIGHT
```

FIGURE 3.17 – CHRONUS - modèle d'une phrase avec paire mot-clé (Levin et Pieraccini, 1995)

TABLE 3.5 – CHRONUS - évaluation de RAP, NLU et SLU (%) (Levin et Pieraccini, 1995)

Modèle	Test (taux d'erreurs %)
RAP	3.5
NLU	6
SLU	9

Pour les expérimentations, 20000 énoncés du corpus ATIS ont été utilisés pour entraîner les modèles. L'ensemble de test se compose de 1000 énoncés, sur lesquels aucune information détaillée n'est disponible. Le tableau 3.5 montre la précision, exprimée en taux d'erreur pour les performances de RAP, NLU et SLU sur l'ensemble de test.

Une autre approche statistique est décrite dans Schwartz et coll. (1996). Cette approche ne cible pas seulement la compréhension des phrases individuelles, mais tient compte du contexte des phrases précédentes. Cette approche comprend 2 étapes principales : *l'analyse sémantique* au niveau de la phrase individuelle et *la classification sémantique* en tenant compte du contexte de la phrase précédente. Ces 2 étapes sont appliquées également aux données du corpus ATIS muni d'étiquettes d'arbres d'analyse syntaxique (*Parse Trees*) et d'étiquettes sémantiques. Des probabilités ont été ajoutées aux arbres d'analyse syntaxique-sémantique lors de l'étape d'analyse sémantique. La classification sémantique, tenant compte de l'historique des phrases précédentes, est ensuite effectuée en utilisant des *arbres de décision* (la section 3.2.3 décrit l'approche d'arbres de décision). Le système complet a été entraîné sur un sous-ensemble des 4500 énoncés annotés du corpus ATIS. L'ensemble de test était un autre sous-ensemble de ce corpus. Le tableau 3.6 affiche le taux d'erreurs obtenu soit en tenant compte du contexte des phrases précédentes, soit sans tenir compte du contexte des phrases précédentes.

3.2.2.2 Combinaison d'une approche par règles avec une approche statistique

Les grammaires basées sur des règles peuvent fournir une analyse précise et détaillée lorsqu'un énoncé parlé est couvert par la grammaire. Contrairement aux approches statistiques elles ne sont pas robustes pour les phrases non couvertes. Une solution viable est une tâche NLU qui combine une approche basée sur des règles et une approche statistique. Dans l'étude de Wang1 et coll. (2002) une approche par règles est combinée avec 3 approches statistiques : un classifieur bayésien naïf, une approche n-gramme et une approche SVM.

Les Séparateurs à Vaste Marge (SVM - *Support Vector Machines*) effectuent une classifica-

TABLE 3.6 – Approche statistique tenant compte du contexte des phrases précédentes - évaluation de la NLU (%) (Schwartz et coll., 1996)

Modèle	Test (taux d'erreurs %)
Sans contexte	14.5
Avec contexte	5

TABLE 3.7 – Évaluation d'un système de NLU combinant approche par règles et approche statistique statistique dont une par SVM (Wang1 et coll., 2002)

Modèle	Taux d'erreur de classification (%)
Bigram	2.99
Réseau bayésien naïf	2.53
SVM	1.84

tion en séparant l'espace des vecteurs d'entrée par un hyperplan dans l'espace des vecteurs d'entrée. Lors de la construction du SVM, il est nécessaire de maximiser la marge, c'est à dire la distance entre l'hyperplan et les échantillons les plus proches pour que le classifieur soit optimal (Cortes et Vapnik, 1995). Afin de combiner des approches statistiques et une approche par règles, les non-terminaux d'une grammaire non-contextuelle (CFG - *Context Free Grammar*) sont inclus comme paramètres dans les classifieurs statistiques. Le classifieur SVM est d'abord appliqué pour obtenir la classe de tâches pour une commande. Ensuite la grammaire n'applique que les règles grammaticales liées à la classe identifiée.

Les données d'apprentissage se composent de 1424 phrases du corpus ATIS. Un autre sous-ensemble de 435 phrases du même corpus est l'ensemble de test. Les résultats sont présentés dans le tableau 3.7.

3.2.3 Les arbres de décision

Kuhn et De Mori (1995) décrivent un modèle de NLU utilisant des arbres de décision. Les arbres de décision, ou arbres de classification sémantique dans le contexte de la NLU, apprennent les règles sémantiques automatiquement à partir des données d'apprentissage. Pour entraîner des arbres de classification, trois éléments sont nécessaires :

1. un ensemble de questions fermées qui peuvent être appliquées aux données pour vérifier si une séquence de mots correspond à certaines expressions régulières,
2. une règle pour sélectionner la meilleure question à un certain nœud,
3. une méthode d'élagage des arbres (*Pruning*) pour éviter un sur-apprentissage.

La figure 3.18 montre un exemple d'un arbre de décision qui doit décider s'il faut montrer à l'utilisateur l'attribut de "fare" (tarif) ou non. Les phrases qui aboutissent à une feuille "YES" (OUI) auront l'attribut de "fare" dans leur liste d'attributs. Les symboles "<" et ">" correspondent au début et à la fin d'une phrase, un "+" entre deux mots ou symboles indique un écart d'au moins un mot entre eux. L'expression $M(w)$ (par exemple $M(fares)$ dans la figure) correspond à une ou plusieurs occurrences du mot w. Par exemple, l'entrée "Show me first-class fare flights to Boston" ("Montrez-moi les vols en première classe vers Boston")

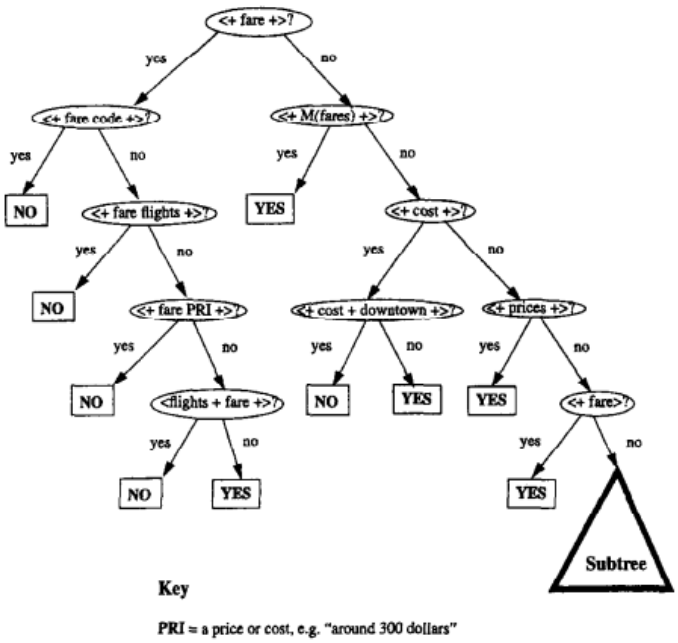


FIGURE 3.18 – Arbre de décision permettant de présenter ou non un tarif (*"fare"*) (Kuhn et De Mori, 1995)

correspond au modèle $< +fare+ >$ à la racine, ne correspond pas à $< +farecode+ >$, et correspond au modèle $< +fareflights+ >$, elle délivre donc "NO". "Show me the fare for flights to Boston" ("Montrez-moi les tarifs des vols pour Boston") correspond à l'expression racine mais à aucune autre expression qu'elle rencontre, et donne donc "YES".

Le système proposé a été appris sur 5501 phrases du corpus ATIS, et testé sur un sous-ensemble de 542 phrases du même corpus et donne une précision de 91%.

3.2.4 Les modèles conditionnels CRF

Alors que les HMM modélisent la distribution à plusieurs variables $P(y,x)$, les modèles de champs conditionnels aléatoires (CRF - Conditionnal Random Fields) visent à maximiser la probabilité conditionnelle $P(y|x)$. Le CRF modélise la dépendence entre *chaque* état et *l'entière observation*. Les modèles HMM par contre modélisent la distribution des données $P(x)$ qui pourraient imposer des caractéristiques fortement dépendantes les uns des autres. Cela rend les modèles de CRF plus adaptés à la prédiction de séquences, comme pour l'étiquetage sémantique, contenant des caractéristiques complexes, qui se chevauchent (Jeong et Lee, 2008). La figure 3.19 montre un CRF à chaîne linéaire pour l'intervalle du temps de $t-1$ à $t+1$. Chaque nœud blanc désigne une variable aléatoire, (c'est-à-dire, un état caché, ou le sens y) et le nœud gris correspondant représente sa valeur observée (le mot x). *L'étiquetage* des séquences est souvent associé à la *classification* des séquences qui est souvent une tâche conjointe. De cette façon, la prédiction peut être combinée avec une classification d'intention. Ces 2 tâches peuvent être effectuées en utilisant les Champs Aléatoires Conditionnels Triangulaires (Tri-CRF - *Triangular Conditional Random Field*).

Le modèle Tri-CRF de Jeong et Lee (2008) (figure 3.20) est une extension du CRF à chaîne

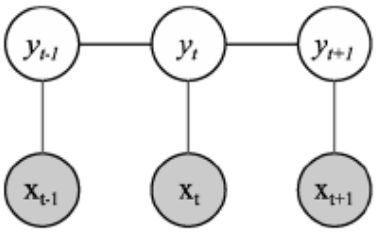


FIGURE 3.19 – Principe du fonctionnement du CRF à chaîne linéaire (Jeong et Lee, 2008)

linéaire visant la prédiction des intentions et des concepts. Ce modèle est constitué d'une séquence observée des mots x, d'une séquence cachée représentant attributs correspondants y, pour un intervalle de temps de $t-1$ à $t+1$, auxquels il convient d'ajouter l'intention associée à l'énoncé z. La séquence observée correspondant à l'ensemble de la phrase est x_0. Les liens de dépendance entre cette variable et la séquence cachée sont également ajoutés. Autrement dit, chaque mot observé x_t dans une séquence est dépendant conditionnellement de son étiquette *non observée* y_t. L'étiquette y_t est alors dépendante conditionnellement de l'étiquette précédente y_{t-1}. Chaque attribut y_t (et donc potentiellement chaque mot x_t) est dépendant de l'intention z de l'énoncé entier.

La distribution de probabilité conditionnelle du Tri-CRF est composée de 2 facteurs : le facteur observation (θ_t^{obs}) qui relie l'intention z à l'étiquette y_t et au mot x_t à une étape temporelle donnée, et le facteur de transition (θ_t^{trans}) qui relie l'intention z au changement d'étiquette entre étapes temporelles (y_t et y_{t-1}). Les poids des facteurs du modèle sont appris en maximisant la vraisemblance logarithmique conditionnelle des données d'apprentissage en utilisant une variante de la méthode quasi-Newton appellée *Limited-memory BFGS* avec régularisation L2.

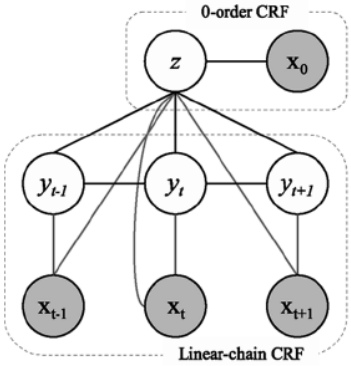


FIGURE 3.20 – Le principe du fonctionnement du tri-CRF (Jeong et Lee, 2008)

Pour évaluer les performances du modèle Tri-CRF les données du corpus ATIS sont séparées en 90% de données constituant l'ensemble d'apprentissage et les 10% de données restantes pour l'ensemble de test. Le tableau 3.8 montre les performances obtenues pour la prédiction de concepts et la classification d'intentions, pour deux modèles séparément appris (*INDÉP*). Ces performances sont dépassées par un modèle où les slots et intentions sont

TABLE 3.8 – Tri-CRF - évaluation de la prédiction de concepts et d'intentions, F-mesure (%) (Jeong et Lee, 2008)

Modèle	Attributs	Intention
INDÉP	90.67	92.09
SIMULTANÉ	94.42	93.07

appris simultanément (*SIMULTANÉ*).

3.2.5 Logiciel commercial RASA : une combinaison de CRF et SVM

Rasa NLU[1] est un outil logiciel libre de création de pipelines NLU. Selon de récentes études comparant les systèmes NLU commerciaux dominants (Braun et coll., 2017), Rasa était parmi les plus performants. Au contraire de Tri-CRF, Rasa ne prédit pas une séquence d'attributs pour chaque mot d'entrée, mais plutôt un jeu d'étiquettes d'attributs et de valeurs d'attributs associés à différents segments des entrées. Théoriquement, les prédictions peuvent se superposer et peuvent s'appliquer à des portions de mots mais c'est rarement le cas en pratique. Nous considérons la configuration '`spacy_sklearn`' de Rasa qui utilise une chaîne linéaire CRF pour classifier les étiquettes d'attributs et une table de correspondance pour déterminer les valeurs d'attributs. Le modèle utilise séparément un SVM linéaire basé sur une représentation vectorielle des mots pré-appris pour classifier les intentions.

3.2.6 RNN du type encodeur-décodeur bidirectionnel basé sur l'attention

Les modèles de CRF, ont récemment été remplacés dans leur utilisation par les *réseaux de neurones profonds* (RNN). Les RNN classiques sont des réseaux à connexions récurrentes qui prennent en compte à un pas de temps t un certain nombre d'états passés. L'étude de Liu et Lane (2016) présente un RNN du type encodeur-décodeur bidirectionnel basé sur l'attention (Att-RNN). Les tâches de détection de l'intention et de remplissage des attributs sont également effectuées simultanément. La figure 3.21 illustre ce système où la gauche de la ligne pointillée constitue la partie encodeur du système, la droite la partie décodeur, pour une phrase tirée du corpus ATIS.

Étant donné une séquence de mots $w = (w_0, w_1, w_2, \cdots, w_{T+1})$, une séquence d'attibuts $s = (s_0, s_1, s_2, \cdots, s_{T+1})$, et une séquence d'intentions $c = (c_0, c_1, c_2, \cdots, c_{T+1})$, à chaque pas de temps t, l'intention c est émise, lors de l'arrivée de la séquence de mots d'entrée w. L'intention générée à la dernière étape est utilisée comme prédiction d'intention de la phrase complète. L'intention de sortie à chaque pas de temps est renvoyée à l'état RNN. Par conséquent pour la classification d'intentions, cela conduit à l'équation suivante :

$$(C_T|w) = P(C_T|w_{\leq T}, c_{\leq T}, s_{\leq T}) \tag{3.26}$$

Pour le remplissage des attributs, à chaque étape t au cours de la séquence de mots

1. https://rasa.ai/products/rasa-nlu/

d'entrée, la sortie d'étiquette d'attribut est modélisée s_t comme une distribution conditionnelle sur les intentions précédentes $c_{<T}$, les étiquettes des attributs précédentes $s_{<T}$, et la séquence de mots d'entrée jusqu'au pas de temps t, ce qui conduit à l'équation :

$$P(s|w) = P(s_0|w_0) \prod_{t=1}^{T} P(s_t|w_{\leq t}, c_{\leq t}, s_{\leq t}) \tag{3.27}$$

Un des avantages du RNN est la capacité de trouver le lien entre une séquence de source et de cible de longueurs différentes. Dans le contexte de prédiction des attributs, le modèle doit aligner les concepts avec les segments correspondants de la phrase. On n'aligne pas les intentions car on suppose qu'une phrase n'a qu'une seule intention. Le modèle RNN de Liu et Lane (2016) applique un alignement, où chaque mot est associé à un seul attribut.

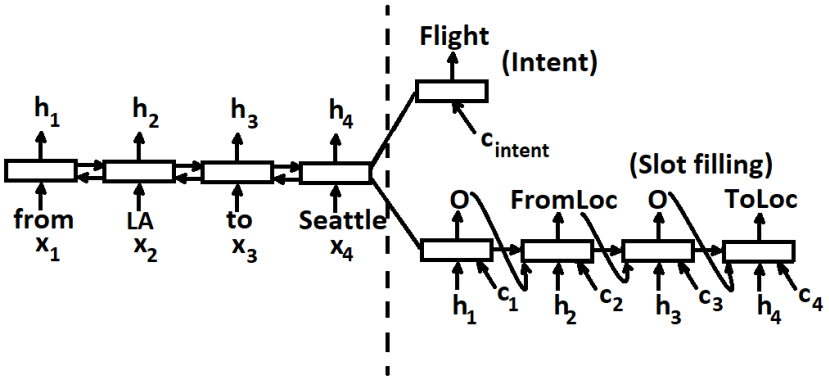


FIGURE 3.21 – Modèle RNN du type encodeur-décodeur bidirectionnel basé sur attention (Liu et Lane, 2016)

L'encodeur est un *LSTM* (*Long Short Term Memory*), un RNN composé d'unités LSTM. Le LSTM modélise mieux les dépendances longues (Long-terms Dependencies) que les RNN simples. Ce mécanisme peut contribuer à trouver des relations entre des concepts éloignés les unes des autres dans la phrase, pour mieux prédire les classes d'intentions. Le LSTM est une solution pour éviter ou réduire le problème du *Vanishing Gradient* (Chung et coll., 2014). L'unité LSTM peut mémoriser des valeurs grâce aux cellules de mémoire. La figure 3.22 montre l'unité LSTM de base : i, f et o sont respectivement la porte d'entrée, la porte d'oubli et la porte de sortie. c et $\tilde{c}$ sont respectivement la cellule de mémoire et la cellule de mémoire mise à jour. Chacune des portes agit comme un neurone dans le sens qu'elle utilise une fonction d'activation de la somme pondérée des entrées. De cette façon, elles contrôlent le flux de données qui passe par cette unité. Sur la figure, l'absence de l'attribution d'un attribut à un mot est marqué comme O.

Le fonctionnement du LSTM pour modéliser les dépendances longues peut être renforcé par le mécanisme de l'attention : au lieu d'utiliser seulement un état caché h_i à chaque pas, on utilise aussi le vecteur d'attention c_i qui peut fournir une information additionnelle sur le contexte. Les états cachés du RNN contiennent l'information de toute la séquence mais l'information peut se perdre lors de la propagation. Le vecteur d'attention c_i est un vecteur et somme de tous les états cachés de l'encodeur à l'instant i, pondérés avec des poids appris ce qui assure la prise en compte de l'attention pour les différents parties de la séquence.

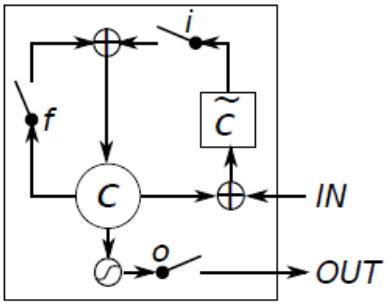


FIGURE 3.22 – Unité LSTM de base (Chung et coll., 2014)

TABLE 3.9 – RNN - évaluation de la prédiction de concepts et d'intentions (%) (Liu et Lane, 2016)

Modèle	Attributs(% F mesure)	Intention (% Taux d'erreurs)
INDÉP	94.91	2.13
SIMULTANÉ	94.64	1.79

Par exemple, dans la phrase *Ce restaurant sert de la cuisine italienne*, il faut attribuer plus d'attention au mot *italienne* pour la prédiction de l'attribut *nourriture*.

Pour ce qui concerne le modèle RNN *bidirectionnel* du type encodeur-décodeur, l'encodeur est constitué de deux parties : *forward* et *backward*. L'encodeur *forward* lit la séquence des jetons de gauche à droite et génère un état caché fh_i à chaque itération i (le temps i va de 1 à T). L'encodeur *backward* lit la même séquence de jetons de droite à gauche et génère un état caché bh_i à chaque itération i. Donc pour chaque pas du temps i, l'encodeur lit un mot x_i et émet un état caché h_i qui est une concaténation de fh_i et bh_i.

$$h_i = [fh_i, bh_i]$$

La partie décodeur est constituée de deux sous-parties : le décodage des attributs et le décodage de l'intention. Le décodage de l'intention est fait par un LSTM RNN. L'état initial du décodeur des attributs est calculé à partir du dernier état (qui contient l'information sur toute la séquence) de l'encodeur *backward*. À chaque instant i, l'état du décodeur s_i est calculé comme résultat d'une fonction qui reçoit en entrée l'état précédent du décodeur (s_{i-1}), l'état caché précédent du décodeur (h_i), la sortie précédente du décodeur (y_{i-1}, c'est-à-dire l'attribut précédent) et un vecteur *d'attention* c_i. La sortie de s_i est l'attribut prédit.

$$s_i = f(s_{i-1}, y_{i-1}, h_i, c_i)$$

Pour évaluer les performances de ce modèle RNN, 4978 phrases du corpus ATIS ont été utilisées comme ensemble d'apprentissage et 893 phrases comme ensemble de test. Le tableau 3.9 montre les performances pour la prédiction de concepts et la classification d'intentions, pour deux modèles séparément appris (Modèle INDÉP). Les performances au niveau de la classification de l'intention sont dépassées par un modèle où les attributs et les intentions sont appris simultanément (Modèle SIMULTANÉ).

3.3 Conclusion sur l'état de l'art de RAP et de NLU

Dans ce chapitre nous avons présenté l'état de l'art de la RAP et de la NLU. Concernant la RAP, nous avons décrit les composants principaux d'approches de RAP classiques statistiques HMM, GMM, hybrides HMM-DNN, et de réseaux de neurones profonds de bout en bout (E2E). Les différences architecturales entre une approche de RAP classique et une approche de RAP E2E ont un impact sur les performances de la RAP. Une approche E2E fonctionne sans lexique phonétique, et le ML est optionnel. Nous avons également comparé les paramètres acoustiques MFCC et fbank. Les paramètres MFCC font partie par défaut de l'approche de RAP classique statistique Kaldi. L'outil de RAP ESPnet par contre utilise des paramètres fbank par défaut. Deep Speech et ESPnet effectuent le mécanisme d'alignement CTC. ESPnet le combine avec le mécanisme d'attention. Cela permet un alignement plus flexible, qui se concentre en même temps sur les paramètres importants et les séquences de caractères.

Concernant la NLU, nous avons décrit des modèles de NLU par règles et des approches statistiques. Nous avons également inclus un aperçu de modèles hybrides, qui combinent une approche par règles avec une approche statistique. Finalement nous avons comparé ces derniers modèles avec des approches de réseaux de neurones profonds, notamment les modèles de RNN. La plupart de ces modèles ont été appris sur les données du corpus ATIS. Une comparaison véritable entre ces modèles reste difficile car des mesures différentes ont été utilisées. Pour certaines approches, une évaluation manquait. La mesure la plus fréquente est le taux d'erreur. D'autres performances ont été évaluées en utilisant la F-mesure ou la précision. En outre les tailles des ensembles de données diffèrent, ainsi que les proportions entre les ensembles d'apprentissage et les ensembles de test. Les approches les plus prometteuses sont les modèles multi-tâches qui prédisent conjointement les concepts et les intentions, en particulier les modèles de RNN.

CHAPITRE 4

État de l'art de SLU

Les méthodes classiques de compréhension de la parole (SLU -*Spoken language understanding*) sont des méthodes dites de SLU *séquentielles* car elles utilisent en les cascadant 2 modules de RAP et de NLU, modules dont l'état de l'art a fait l'objet du chapitre précédent. En effet, dans un système de SLU *séquentielle*, la sortie du module de RAP est injectée sur l'entrée du module de NLU. Le problème principal d'une telle approche est la dépendance de la NLU aux transcriptions issues du module de RAP, ce qui cause une propagation d'erreurs et peut affecter les performances du système dans son ensemble.

C'est pour cette raison que nous pouvons observer un intérêt croissant pour la SLU de bout en bout (E2E - *End to End*) qui vise une compréhension faite directement à partir du signal de parole par un système unique sans que les 2 tâches de reconnaissance et de compréhension ne soient séparées. Cependant, développer des systèmes de SLU E2E plus performants que les systèmes pipeline reste encore un défi à l'heure actuelle.

4.1 Compréhension séquentielle de la parole

Du fait de la dépendance des performances du système de SLU séquentielle aux erreurs de transcription du module de RAP, erreurs qui induisent à leur tour des erreurs supplémentaires du module de NLU, le thème principal de la plupart des études sur la SLU séquentielle concerne la recherche de méthodes envisageables pour pallier cette difficulté. Plusieurs des méthodes proposées s'appuient sur des mesures de confiance, des listes des N meilleures hypothèses (*n-best*), ou un vote pondéré qui améliore le lien entre ces 2 modules. Au cours du parcours de ces différentes méthodes, nous aborderons également le décalage existant entre les performances du module de NLU et le système final de SLU.

4.1.1 Systèmes basés sur règles

Le système *PHOENIX* (Ward, 1991) est un système de SLU effectuant la prédiction de concepts, ciblant le traitement de la parole spontanée. Le module de NLU est basé sur des règles. Le module de RAP est Sphinx avec un ML bigramme. Pour mieux traiter la parole spontanée, des modèles sont ajoutés qui ont été appris à base de parole bruitée. La transcription qui en résulte est ensuite passée sous forme d'une chaîne de mots au module de NLU qui fonctionne en utilisant des états finis. Chaque réseau d'états finis représente un type de

concept, et les transitions sont des mots ou des sous-réseaux. La figure 4.1 décrit les réseaux « fourchette de prix » (*PriceRange*) qui contient 4 transitions, « prix exact » (*PriceExact*), « prix approché » (*PriceApproximate*) et « limite inférieure de prix » (*PriceLowerBound*).

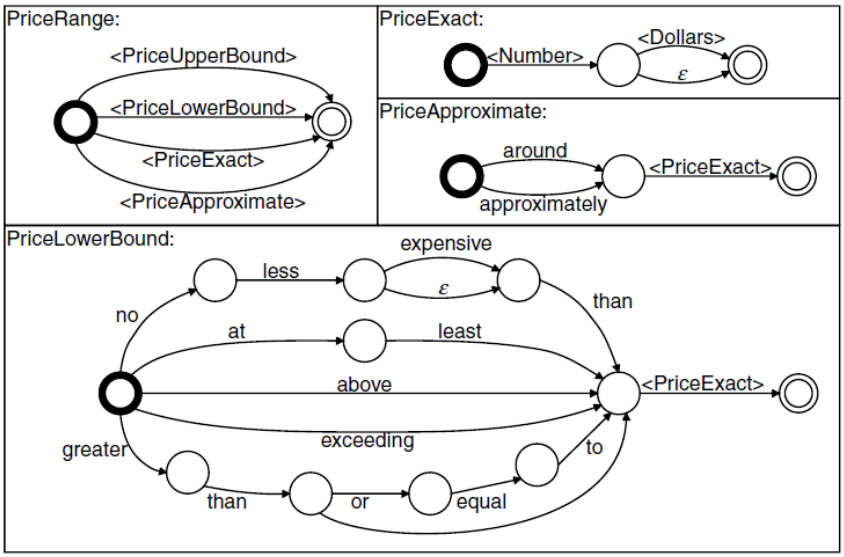


FIGURE 4.1 – PHOENIX - fragments du réseau d'états finis (Ward, 1991)

L'ensemble de données d'apprentissage se compose de 700 énoncés du corpus ATIS. L'ensemble d'évaluation se compose de 93 énoncés du même corpus. Le tableau 4.1 montre les performances de RAP, NLU et SLU pour la prédiction de concepts. Il s'avère que le décalage entre les performances de NLU et de SLU (*Diff.*) est énorme.

4.1.2 Apprentissage automatique du modèle HVS

À la différence de l'approche par règles présentée à la section précédente, le module de RAP (HMM) dans l'étude de He et Young (2003) est suivi par un module de NLU utilisant des états vectoriels cachés (HVS - *Hidden Vector State model*) pour la prédiction de concepts. Les intentions sont prédites en utilisant un classificateur utilisant un réseau bayésien naïf augmenté d'arbres.

Lors du passage du module de RAP vers le module de NLU, les taux d'erreurs augmentent. Pour cette raison, un *rescoring* est appliqué aux N meilleures hypothèses d'un treillis de mots, comme sortie du module de RAP. De cette façon, les probabilités du ML sont combinées avec les probabilités de confiance du modèle de NLU. La figure 4.2 montre le modèle de HVS correspondant. Les concepts de l'arbre d'analyse sont convertis en une

TABLE 4.1 – SLU séquentielle PHOENIX - évaluation de RAP, NLU et SLU, prédiction de concepts (Ward, 1991)

Système	Mesure	(%)
RAP	WER	22.00
NLU	Précision	80.00
SLU	Précision	53.00
NLU ↔ SLU	**Diff.**	**27**

séquence d'états d'un vecteur. Le mot 'Denver' est par exemple décrit par le vecteur sémantique [CITY, FROMLOC, SS].

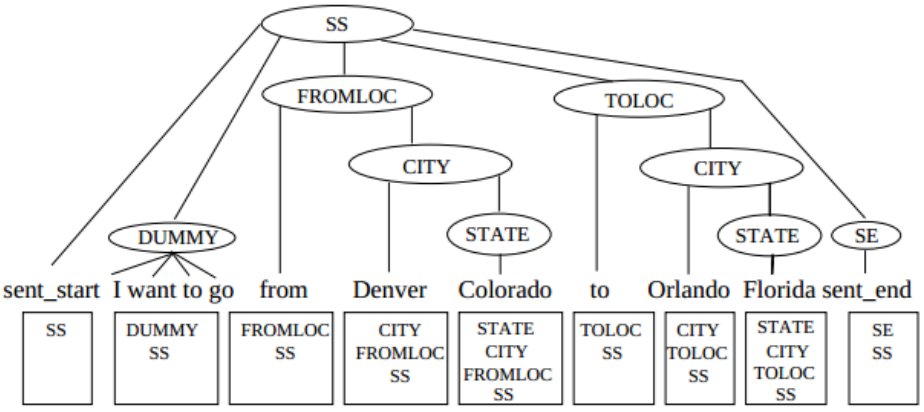


FIGURE 4.2 – Exemple d'un HVS avec arbre (He et Young, 2003)

Le classificateur d'intentions est une application d'un modèle bayésien naïf augmenté d'arbres. Pour prédire les intentions, le classificateur apprend la probabilité conditionnelle de chaque concept sémantique C_i étant donné l'intention G_u, $P(C_i|G_u)$. Finalement la classification est effectuée en sélectionnant l'intention ayant la plus haute probabilité postérieure G_u étant donné l'instance de concept, $C_1 \cdots C_n$, $P(G_u|C_1 \cdots C_n)$. La figure 4.3 montre un exemple d'un réseau bayésien naïf augmenté d'arbres.

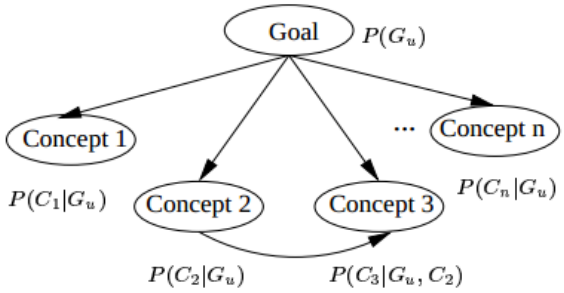


FIGURE 4.3 – Exemple d'un réseau bayésien naïf augmenté d'un arbre (He et Young, 2003)

L'apprentissage et l'évaluation du modèle ont été effectués en utilisant le corpus ATIS. Les sous-ensembles de ATIS-3 NOV93 et DEC94 ont été réservés pour être utilisés comme ensemble de test et sont exclus de l'ensemble d'apprentissage. Le tableau 4.2 montre les résultats pour le composant de RAP, les modules de NLU et de SLU, pour la prédiction de concepts et la classification d'intentions. Pour les concepts ainsi que pour les intentions, les décalages (*Diff.*) entre les performances de NLU et de SLU sont faibles.

4.1.3 Méthode de vote pondéré

Zhai et coll. (2004) combinent les mesures de confiance des transcriptions de sortie de RAP avec une liste des N meilleures hypothèses en utilisant un vote pondéré. Les transcriptions de RAP sont créées en utilisant le système de RAP *BBN Byblos* (Schwartz et coll., 1989)

TABLE 4.2 – SLU séquentielle - modèle HVS - évaluation RAP (WER), NLU, SLU, concepts et intentions (F-mesure) (He et Young, 2003)

Système	Mesure	(%)
RAP	WER	3.4
NLU (concepts)	F-mesure	91.90
SLU (concepts)	F-mesure	90.50
NLU ↔ SLU (concepts)	**Diff.**	**1.40**
NLU (Intent.)	F-mesure	91.20
SLU (Intent.)	F-mesure	90.80
NLU ↔ SLU (Intent.)	**Diff.**	**0.40**

TABLE 4.3 – SLU séquentielle - méthode de vote pondéré, prédiction de concepts, F-mesure (%) (Zhai et coll., 2004)

Modèle	Personne	Lieu	Organization
NLU	71	78	56
SLU 1$^{\text{-ière}}$ hypothèse	46	75	54
SLU Vote pondéré	48	75	55
NLU ↔ SLU Vote pondéré, **Diff.**	**23**	**3**	**1**

appliqué à 1046 phrases en chinois de *Xinhua Agence de Presse*. Les concepts sont prédits en utilisant un modèle de maximum d'entropie (ME - Maximum Entropy) à base de 20000 phrases étiquetées de concepts provenant de *People's Daily Newspaper* (janvier 1998).

Comme la n$^{\text{-ième}}$ hypothèse contient plus d'erreurs de caractère que la n-1$^{\text{-ième}}$ hypothèse, la qualité de la performance de prédiction de concepts diminue. Pour améliorer cette performance, deux mécanismes de vote ont été utilisés. En appliquant la première méthode, un concept est considéré comme correct s'il est prédit dans plus que 30% des n-meilleures hypothèses par phrase de référence. La deuxième méthode combine six mesures de confiance par hypothèse. Ces mesures de confiance concernent le MA, le ML, le nombre de mots, de phonèmes, de silences ainsi que le taux erreurs de concepts à base de maximum d'entropie. Le tableau 4.3 montre que la performance de cette méthode dépasse la F-mesure de l'approche de référence (sélection de la première hypothèse). Seuls les F-mesures pour les concepts de personne, lieu et organisation sont mentionnées. La performance globale pour toutes les étiquettes de concepts n'est pas mentionnée. Le décalage entre les performances de NLU et de SLU dépend du concept, mais est énorme pour le concept *Personne*.

4.1.4 Méthodes utilisant des mesures de confiance

Sudoh et coll. (2006) proposent une méthode qui utilise des mesures de confiance de RAP pour la prédiction de concepts. Ils ont augmenté des transcriptions japonaises issues de la RAP avec des étiquettes de concepts. Leurs données d'entraînement contenaient 10718 phrases d'articles de journaux japonais, avec des étiquettes de concept réparties en 8 catégories. Le modèle acoustique (MA) a été créé avec les mêmes énoncés, prononcés par 106 locuteurs. Leur ML de type 3-gramme basé sur des mots était entraîné sur 34 millions de mots japonais. Un modèle SVM attribue les étiquettes de concept seulement aux transcriptions

TABLE 4.4 – SLU séquentielle - performance d'une méthode utilisant des mesures de confiance (F-mesure)(Sudoh et coll., 2006)

Système	Mesure	(%)
NLU	F-mesure	84.04
SLU sans confiance	F-mesure	66.95
SLU avec confiance	F-mesure	69.02
NLU ↔ SLU (avec conf.)	**Diff.**	**15.02**

de RAP associées à une mesure de confiance dépassant un certain seuil. Comme les SVM traitent des problèmes liés à seulement deux classes, la prédiction pour des classes multiples est transformée en prédictions pour deux classes, suivant une approche *'one-against-all'*, où chaque SVM entraîné cible à tour de rôle une seule classe par rapport aux autres classes. Les données sur lesquelles les classificateurs sont entraînées contiennent des mesures de confiance basées sur des mots, des étiquettes morpho-syntactiques et des probabilités postérieures de RAP. Comme le montre le tableau 4.4, bien que les performances de l'approche cible dépassent celles de l'approche *sans* mesures de confiance (*SLU sans conf.*), il reste un grand décalage entre les performances de NLU et les approches de SLU.

Simonnet et coll. (2017) utilisent également des mesures de confiance issues d'un système de RAP pour améliorer la prédiction de concepts. Cette approche utilise la représentation vectorielle *acoustique* en apprenant des réseaux de neurones convolutifs (CNN) comme décrits dans l'étude de Ghannay et coll. (2016). Les auteurs combinent cette approche avec des réseaux de confusion de mot et des probabilités postérieures comme mesures de confiance pour la prédiction de concepts. Les transcriptions de RAP sont générées en utilisant l'outil de RAP du LIUM (Rousseau et coll., 2014). Le MA est entraîné sur les corpus *ESTER1* (Galliano et coll., 2005), *ESTER2* (Galliano et coll., 2009), *ETAPE* (Gravier et coll., 2012) et *REPERE* (Giraudel et coll., 2012), qui contiennent un total de 565 heures d'enregistrements de parole. Un ML générique est entraîné sur 77 millions de mots provenant de journaux français et interpolé avec un ML spécifique au domaine entraîné sur le corpus *Media* (Bonneau-Maynard et coll., 2005).

Les auteurs comparent les performances d'une approche de SLU utilisant un modèle de CRF avec les performances des RNN bidirectionnels. Ces deux modèles sont appris sur le corpus Media, annoté manuellement de concepts. Les paramètres (*features*) utilisés sont la représentation vectorielle de mots, les paramètres morpho-syntactiques, les mesures de confiances des probabilités postérieures de la RAP. Le tableau 4.5 montre les performances pour les modèles de SLU CRF et RNN. Les performances du modèle de CRF dépassent celles du modèle de RNN. L'intégration de mesures de confiance réduit le *taux d'erreur de concept* (CER - *Concept Error Rate*) et le *taux d'erreur de valeurs de concept* (CVER - *Concept Value Error Rate*) (*SLU CRF avec conf.*) par rapport au modèle de CRF sans mesures de confiance (*SLU CRF sans conf.*). Cette étude ne mentionne pas les performances de NLU.

TABLE 4.5 – SLU séquentielle - comparaison de méthodes avec ou sans mesure de confiance pour la prédiction de concepts (Simonnet et coll., 2017)

Modèle	WER (%)	CER (%)	CVER (%)
RAP	23.6	-	-
SLU RNN sans confiance	-	22.3	28.8
SLU CRF sans confiance	-	20.9	26
SLU CRF avec confiance	-	19.9	25.1

TABLE 4.6 – SLU séquentielle - comparaison de méthodes avec meilleure hypothèse, treillis d'hypothèses et réseaux de confusion (Hakkani-Tür et coll., 2006)

Modèle	Meilleure hypothèse F-mesure (%)	Treillis d'hypothèses F-mesure (%)	Réseaux de confusion F-mesure (%)
téléphone 1	86.00	**93.70**	93.60
téléphone 2	79.80	87.30	**87.60**
date	63.10	71.00	**73.60**

4.1.5 Méthodes utilisant treillis d'hypothèses et réseaux de confusion

Au lieu d'utiliser la meilleure hypothèse de RAP, Hakkani-Tür et coll. (2006) proposent d'améliorer la transition entre le module de RAP et celui de NLU en utilisant des réseaux de confusion de mots (Mangu et coll., 2000) qui sont obtenus à partir du treillis d'hypothèses de la RAP. Les réseaux de confusion de mots fournissent une représentation compacte de plusieurs hypothèses de RAP alignées avec les scores de confiance de mots. Leurs transitions sont pondérées par les probabilités du modèle acoustique et du modèle de langage. Une approche de SLU utilisant des treillis incluant plusieurs hypothèses peut surpasser les performances de la RAP utilisant seulement la meilleure hypothèse.

À la différence des treillis d'hypothèses, les réseaux de confusion de mots appliquent un alignement des mots se produisant au cours de la même intervalle de temps dans le réseau. Les transcriptions de RAP sont générées en utilisant le module de RAP du système de dialogue AT&T (Gupta et coll., 2005). Un outil de NLU statistique est ensuite entraîné sur le corpus 'AT&T Spoken Dialog' pour lequel la taille des données d'apprentissage n'est pas mentionnée. Le modèle de SLU qui en résulte a été testé sur 2 formats de numéros de téléphone *téléphone 1* et *téléphone 2* sur respectivement 617 et 26k énoncés. Un troisième ensemble de test était composé de dates (*date*). Les résultats présentés dans le tableau 4.6 montrent que les approches des treillis d'hypothèses et des réseaux de confusion de mots surpassent l'approche de la meilleure hypothèse.

Dans leur étude, Liu et coll. (2020) présentent et évaluent des modèles de SLU séquentiels qui intègrent des réseaux de confusion de mots, la meilleure hypothèse et les listes des N-meilleures hypothèses (N=10) pour la prédiction de concepts et valeurs de concept. Les probabilités postérieures de RAP sont intégrées dans un modèle de SLU à base d'une représentation de BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et coll., 2019). Le réseau de confusion de mots est introduit dans l'encodeur de BERT et ensuite en

TABLE 4.7 – SLU séquentielle - comparaison de méthodes avec meilleure hypothèse, N-meilleures hypothèses, réseaux de confusion et BERT (Liu et coll., 2020)

Modèle	F-mesure (%)
Meilleure hypothèse	84.06
N-meilleures hypothèses	85.05
Réseaux de confusion	85.01
Réseaux de confusion + BERT	**87.91**

représentation vectorielle. La couche de sortie est un classificateur de concepts et de ses valeurs. Cette approche est évaluée à l'aide du jeu de données DSTC2 (*Second Dialog State Tracking Challenge*) (Henderson et coll., 2014). Les résultats qui sont affichés au tableau 4.7 montrent que le modèle qui combine BERT et les réseaux de confusion surpassent les autres modèles étudiés.

4.2 Compréhension de la parole de bout en bout (SLU E2E)

Bien que les systèmes séquentiels de SLU soient les plus largement utilisés, de nouvelles approches pour la compréhension de la parole E2E ont récemment été explorées. Ce dernier type d'approche tente de combiner la RAP et la NLU en une tâche unique, en essayant d'éviter la cascade d'erreurs cumulées par une approche SLU séquentielle. L'un des avantages de la compréhension de la parole de bout en bout (E2E - *End to End*) provient du fait que la reconnaissance de *chaque mot* dans la phrase n'est *pas toujours nécessaire* pour extraire les concepts et les intentions. En outre, un tel système a accès aussi aux niveaux acoustique et prosodique qui peuvent avoir un impact positif sur les performances de SLU.

4.2.1 Classification d'intentions à base des paramètres acoustiques MFCC

Serdyuk et coll. (2018) infèrent les intentions directement des paramètres acoustiques MFCC, guidés par l'intuition que nous ne comprenons pas nécessairement la parole en reconnaissant et comprenant chacun des mots d'un énoncé. Par contre, la parole est directement comprise lorsque une attention est accordée aux concepts directement liés à la tâche. Une telle approche permet aux composantes prosodiques du signal de parole d'être exploitées par le modèle E2E pour la classification d'intention. Les différences prosodiques entre une question et par exemple la voix impérative peuvent contribuer à aider la classification des intentions.

Les auteurs ont proposé d'utiliser un modèle SLU E2E entraîné à l'aide d'un GRU RNN bidirectionnel à 4 couches d'encodeur-décodeur. La longueur de la séquence est réduite avec un LSTM pyramidal bidirectionnel (Chan et coll., 2016), pour extraire les représentations au niveau de la syllabe selon la représentation de la figure 4.4. Une couche *softmax* est utilisée pour calculer les probabilités d'intention postérieures. Les données d'apprentissage sont composées de 320 heures de données du corpus ATIS étiquetées avec 35 types d'intentions, dix autres heures du même corpus sont les données de validation. Le tableau 4.8 montre que

TABLE 4.8 – SLU E2E - évaluation de la classification d'intentions à partir des paramètres acoustiques MFCC (Serdyuk et coll., 2018)

Modèle	Précision (%)
Référence	80.00
E2E	74.10
E2E bruité	72.00

les performances du modèle E2E ne dépassent pas le modèle de référence séquentiel. Les performances d'un deuxième modèle E2E qui est appris sur la parole bruitée, baissent encore plus. Par contre, cette approche montre la faisabilité d'effectuer de la SLU à partir du signal acoustique.

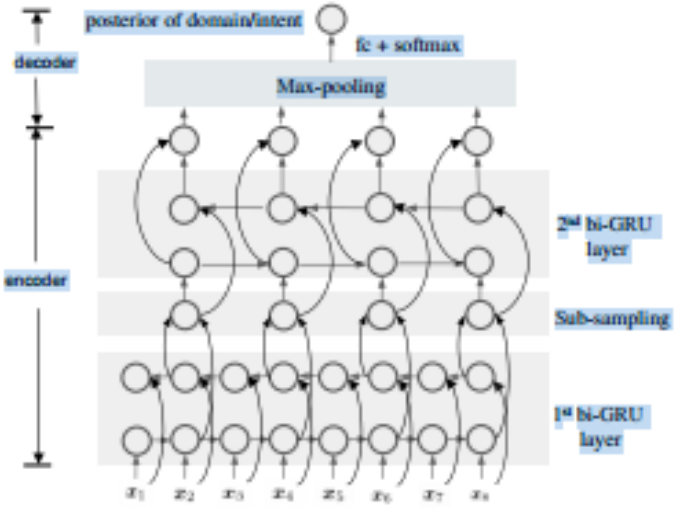


FIGURE 4.4 – SLU E2E - GRU RNN bidirectionnel à 4 couches (Serdyuk et coll., 2018)

4.2.2 Apprentissage multitâche à base de transcriptions augmentées de concepts symboliques

Ghannay et coll. (2018) enrichissent leurs données de parole d'apprentissage avec des étiquettes de concepts. Ces étiquettes de concepts sont injectées dans les transcriptions de RAP sous forme d'étiquettes symboliques. Ces données sont transmises au système de RAP par réseaux de neurones de *Baidu Deep Speech* (chapitre 3, section 3.1.4.1), qui se compose de deux couches CNN et de six couches récurrentes bidirectionnelles, utilisant la fonction de coût CTC (*Connectionist Temporal Classification*) (Amodei et coll., 2016). Pour le décodage, le CTC est lié à un ML basé sur caractères, puis une recherche de faisceau (*Beam Search*) est appliquée. Au lieu d'appliquer le schéma BIO d'étiquetage de concepts (*Begin-Inside-Outside*) aux données d'apprentissage, des balises qui représentent huit concepts sont injectées dans les transcriptions de la RAP.

Huit balises différentes marquent le début de chacun des huit concepts, tandis qu'une balise identique marque la fin du concept. Dans l'exemple,

TABLE 4.9 – SLU E2E multitâche - transcriptions augmentées de concepts symboliques - prédiction de concepts (Ghannay et coll., 2018)

Modèle SLU	F-mesure (%)
Séquentiel	64.00
E2E	67.10
E2E*	69.00

```
"le sculpteur [antoine] est mort # hier]",
```

le concept personne est précédée de la balise `[`, la balise représentant l'entité 'temps' est précédée de `#`. Les deux entités sont suivies d'une seule et même balise `]` qui marque la fin de l'inclusion.

Afin de réduire l'importance que la fonction de coût CTC accorde à chaque caractère et d'attirer davantage l'attention sur les concepts, toutes les suites de caractères qui ne contiennent pas de concepts sont remplacées par le symbole $\star$. L'exemple ci-dessus est transformé en,

```
"* [antoine] * # hier]"
```

après avoir remplacé les suites de caractères sans concepts par le symbole $\star$. Les données d'entraînement sont composées de parties des corpus ESTER1, ETAPE et *QUAERO* (Névéol et coll., 2014), pendant que les données de test de développement et de validation sont composées respectivement de parties des corpus ESTER1 et ESTER2, et d'ETAPE. Les données contiennent 107 heures de données d'entraînement, 24 heures de données de test et 30 heures de données de développement.

Une approche d'apprentissage multitâche est appliquée. Dans un premier temps, le réseau est entraîné uniquement pour la RAP sans émettre de caractères représentant des concepts. Dans une seconde étape, la couche *softmax* est réinitialisée pour prendre en compte les marqueurs de concepts et un deuxième apprentissage est fait. Les performances de cette approche sont comparées à une approche SLU séquentielle de référence. Le tableau 4.9 montre que les performances des deux modèles E2E, *E2E sans*, et *E2E* avec* symboles d'astérisque ajoutés au données d'apprentissage, dépassent le modèle de référence séquentiel CRF.

Hatmi et coll. (2013) décrivent une méthode de RAP qui prédit également des concepts, pour améliorer la qualité des transcriptions de sortie de RAP. Les transcriptions du lexique et du ML sont augmentés d'étiquettes de concepts. Puis le système de RAP génère des transcriptions étiquetées de concepts au niveau du mot. Les transcriptions de RAP sont générées en utilisant le système de transcription de la parole LIUM, basé sur le système de RAP Sphinx (Deléglise et coll., 2005, 2009). Le MA est entraîné sur 240 heures de données des corpus ESTER1 et ESTER2. 16 émissions du corpus ESTER2 ont été utilisées comme ensemble de test. Les modèles de langage de quadri-grammes, tri-grammes bi-grammes interpolés ont

TABLE 4.10 – SLU E2E - transcriptions augmentées de concepts, évaluation de RAP et de SLU (%) (Hatmi et coll., 2013)

Modèle	WER	F-mesure (%)
RAP de référence	20.23	-
SLU E2E (RAP+concept)	21.17	63
SLU séquentielle	-	58

été créés à base des corpus AFP, APW, Le Monde, Afrik, L'humanité, et de ESTER1.

Les corpus utilisés pour créer les ML et le lexique ont été annotés manuellement de concepts. L'exemple suivant montre le schéma de balises de concepts *IOB*, intérieur, extérieur, début, (*Inside-Outside-Begin*), utilisé pour l'une des 7 catégories de concepts :

Il est vingt **-time-B** heures-**time-I** (4.1)

Ils comparent 2 systèmes de RAP *sans* et *avec* reconnaissance de concepts dont le composant d'étiquetage de concept est *LIANE* (Béchet et Charton, 2010). Cette approche de RAP qui utilise également l'information de concepts, et que l'on peut considérer comme approche SLU E2E, est comparée avec une approche SLU séquentielle.

Au niveau de RAP, le tableau 4.10 montre que le système de RAP qui intègre la reconnaissance de concepts (RAP+concept), ne surpasse pas le système de RAP de référence. Par contre, au niveau de SLU, le système de RAP *avec* reconnaissance de concept surpasse l'approche séquentielle.

4.2.3 Apprentissage de curriculum par transfert

L'intuition de l'apprentissage de curriculum par transfert est basée sur une analogie faite avec le comportement humain, en effet notre apprentissage est plus efficace lorsqu'une tâche complète est décomposée en sous-tâches, lorsque les concepts et les exemples à apprendre sont présentés progressivement, du plus simple au plus complexe. La motivation de ce type d'apprentissage est que l'ordre dans lequel les données d'apprentissage sont présentées, des exemples les plus faciles aux plus difficiles, aide les algorithmes d'apprentissage, en accélérant la convergence (Krueger et Dayan, 2009). Plus récemment, Caubrière et coll. (2019) décrivent une approche SLU E2E où ils ont mis en œuvre ce type d'apprentissage, en utilisant l'outil de RAP *Baidu Deep Speech* (décrit au chapitre 3 en section 3.1.4). Les poids appris à l'étape t sont réinjectés comme poids préinitialisés à l'étape $t+1$. Il consiste en une phase d'apprentissage de RAP, et trois phases d'apprentissage de concepts. Le modèle de RAP est appris sur les corpus *EPAC* (Esteve et coll., 2010), ESTER2, ETAPE, QUAERO et REPERE. Un modèle de SLU est entraîné sur les mêmes données. Pour que la fonction de coût CTC se concentre davantage sur les concepts que sur les mots de contexte, les mots à l'extérieur des étiquettes de concept ont été également remplacés par le symbole d'astérisque $\star$ comme dans l'étude de Ghannay et coll. (2018). Le tableau 4.11 montre une comparaison de 3 performances d'apprentissage par transfert :

TABLE 4.11 – SLU E2E - apprentissage par transfert - prédiction de concepts (CER %) (Caubrière et coll., 2019)

Modèle	CER (%)
Transfert1	20.10
Transfert2	19.00
Transfert3	18.10
Transfert2*	17.00
Transfert3*	16.40

TABLE 4.12 – SLU E2E - apprentissage par transfert, classification d'intentions (Lugosch et coll., 2019)

Modèle	Précision (%)
Sans apprentissage par transfert	96.6
Avec apprentissage par transfert	97.2

— Un *premier* apprentissage par transfert (*Transfert1*) sur les étiquettes de concept du corpus Media

— Un *deuxième* apprentissage par transfert (*Transfert2*) sur les étiquettes de concept des corpus Media et Port-Media.

— Une *troisième* phase d'apprentissage par transfert (*Transfert3*) sur toutes les données du modèle de RAP, étiquetées de concepts.

*Transfert2** et *Transfert3** montrent les performances sur les modèles d'apprentissage par transfert équivalents, mais avec des symboles $\star$ insérés. Les données de test sont 3500 énoncés, enlevés des 17700 énoncés de Port-Media.

Lugosch et coll. (2019) proposent une méthode d'apprentissage par transfert de SLU E2E, pour lequel un modèle de RAP est d'abord également appris, suivi par un apprentissage par transfert au niveau des intentions. À cette fin, le corpus *Fluent Speech Commands* est utilisé (corpus décrit au chapitre 2 en section 2.6.5). Les intentions sont composées de 31 combinaisons différentes possibles de valeurs de concept. Pour chaque intention, il existe plusieurs combinaisons possibles. Par exemple, l'intention {action : "activer", objet : "lumières", emplacement : "aucun"} peut être exprimée comme "allumer les lumières", "lumières allumées", etc. Les données sont divisées en un ensemble de développement, de test et d'apprentissage, sans mentionner la taille exacte de chaque ensemble. Le tableau 4.12 montre les résultats de performance pour les prédictions d'intentions *sans* et *avec* apprentissage par transfert.

4.2.4 Conclusion

Le problème principal des systèmes séquentiels de compréhension de la parole est la dépendance des transcriptions sorties du module de RAP. Bien que cette dépendance réduise les performances du module de NLU, leurs performances surpassent en général celles des approches de bout en bout. Néanmoins le décalage entre les performances de leurs mo-

dules de NLU et du système de SLU séquentielle considéré dans son ensemble restent souvent élevés, malgré les stratégies mises en œuvre pour réduire ces décalages. Cependant, la technique d'apprentissage par transfert, appliquée dans l'approche SLU E2E de l'étude Lugosch et coll. (2019) montre des performances prometteuses qui surpassent la plupart des performances de SLU séquentielle. Les études de Ghannay et coll. (2018) et de Caubrière et coll. (2019) ont montré que la reconnaissance de chaque mot par phrase n'est pas nécessaire pour déduire les concepts ou les intentions. En outre il est avantageux qu'un modèle de SLU E2E puisse exploiter le niveau acoustique, comme l'étude de Serdyuk et coll. (2018) l'a considéré.

CHAPITRE 5

Méthode

Cette section introduit les questions de recherche et détaille les différentes étapes de la méthode suivie pour y répondre. Comme vu dans les chapitres de l'état de l'art, la compréhension automatique de la parole (SLU) a souvent été abordée en considérant deux sous-tâches séparées, celle de la reconnaissance automatique de la parole (RAP) et celle de la compréhension automatique du langage (NLU), le problème principal étant résolu en cascadant ces tâches dans un pipeline ce qui conduit à une approche séquentielle. Dans cette thèse nous cherchons à **comprendre quels avantages une approche SLU de bout-en-bout (E2E) peut offrir par rapport à une approche en pipeline classique**.

L'état de l'art des approches SLU du chapitre 4 montre que l'un des inconvénients principaux des approches de SLU séquentielle est l'effet cascade d'erreurs en raison des imperfections du composant RAP qui impacte la NLU. De fait, l'objectif principal de ces approches est de réduire l'influence des erreurs de RAP sur la NLU. En utilisant un modèle d'inférence qui extrait les intentions et concepts directement du signal – approche de bout-en-bout (*End-to-End, E2E*) –, **peut-on éviter la cascade d'erreurs de l'approche SLU séquentielle ?** L'approche de Ghannay et coll. (2018) semble confirmer cette hypothèse et montre également qu'une RAP parfaite n'est pas nécessaire pour obtenir de bonnes performances NLU.

Par ailleurs, étant donné que l'approche E2E infère des concepts et des intentions transportés par un énoncé directement à partir du signal acoustique, on peut se poser la question de savoir **si le modèle E2E exploite l'information prosodique et permet ainsi d'améliorer les performances de prédiction d'intentions et de concepts**. Par exemple, l'étude de Serdyuk et coll. (2018) qui utilise une approche SLU E2E par un modèle neuronal a montré qu'il est possible de prédire certaines intentions en utilisant uniquement des informations *acoustiques*.

Un autre avantage potentiel d'une approche E2E est le pouvoir d'abstraction. En effet, dans une approche séquentielle la transcription en mot du signal d'entrée est une étape fondamentale. Or les mots hors vocabulaires, les syntaxes inusuelles ne sont pas bien modélisées par les systèmes de RAP courants. **Est-ce qu'un modèle SLU E2E serait capable d'être plus robuste aux variations de vocabulaire et de syntaxe ?**

Sur un autre plan, nous avons vu que les modèles SLU à base de réseaux de neurones profonds nécessitent des corpus d'apprentissage de grande taille et adaptés aux besoins de l'application visée. Cependant, comme indiqué au chapitre 2, il y a un manque de corpus dans le domaine domotique en français. On peut donc se demander **comment nous pou-**

vons apprendre des modèles profonds avec une faible quantité de données initiales.

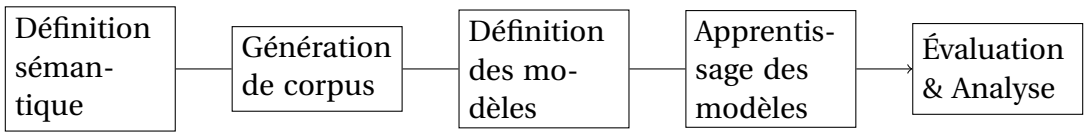


FIGURE 5.1 – Représentation schématique de la démarche proposée

Pour répondre à toutes ces questions, la démarche, que nous avons adoptée et que nous décrivons sur le diagramme de la figure 5.1, consiste à poser le problème de SLU comme un problème de *slot-filling* appliqué au domaine de la commande vocale dans l'habitat. Pour cela, nous nous appuierons sur l'expérience de l'équipe de recherche dans le domaine et sur les environnements intelligents du laboratoire. La première tâche de la méthode est donc de définir l'espace sémantique des commandes vocales. Cette tâche est introduite en section 5.1.

Une fois la sémantique et les commandes possibles clarifiées, il est possible de recueillir un premier corpus de *test* afin de pouvoir évaluer les approches SLU sur des données réalistes. Concernant le corpus d'*apprentissage*, nous avons choisi une approche de génération de corpus artificiel à partir d'expertises. Afin de reproduire une situation réaliste, aucune donnée du corpus de test n'est utilisée pour l'apprentissage des modèles. Cette étape est introduite en section 5.3 et fera l'objet du chapitre 6.

Les modèles SLU de bout-en-bout et séquentiel sont introduits en section 5.4. Il s'agit principalement de concevoir une approche pipeline de référence se basant sur un système de RAP à l'état de l'art couplé à un modèle NLU neuronal. Le modèle E2E est composé d'un réseau pyramidal de neurones récurrents. Ce réseau est appris par une méthode hybride multi-tâche qui combine une fonction de coût CTC et un encodeur-décodeur basé sur l'attention. Théoriquement ce type de modèle est capable de gérer les mots hors vocabulaire. Nous comptons donc étudier si l'interaction entre l'attention et le CTC peut renforcer la robustesse d'un tel modèle sur des données de test contenant une *variabilité linguistique*.

En section 5.5, nous décrivons succinctement les métriques utilisées pour mesurer les performances des modèles SLU. En ce qui concerne les caractéristiques contextuelles, prosodiques et phonétiques, celles-ci ont été mesurées avec des méthodes de type test d'hypothèse et sont introduites en section 5.6. Enfin, afin d'évaluer la robustesse des modèles aux variations lexicales, la section 5.7 décrit la méthode à base de génération employée. Ces analyses font l'objet du chapitre 9.

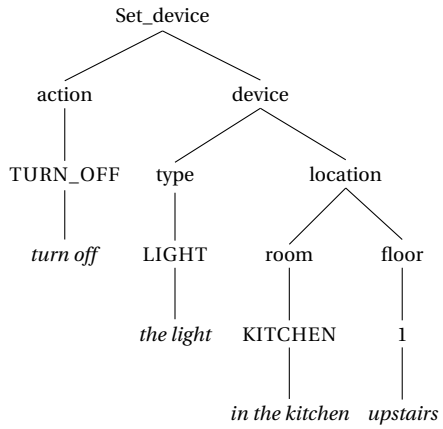
5.1 Représentation des informations extraites des commandes vocales

Pour extraire des intentions et des concepts des commandes vocales, il est nécessaire de bien les définir. L'intention peut être définie comme l'objectif du locuteur lorsqu'il ou elle énonce une commande vocale. La notion d'intention est proche de celle de *l'acte de langage*

(*Speech Act*). Dans ce cadre, il s'agit de l'activité communicative définie par le fait qu'un locuteur énonce une parole dans le but de produire un effet sur son interlocuteur (Crystal, 2011). Dans cette étude, l'intention est le type d'acte recherché par le locuteur lorsqu'il s'adresse au système domotique. Par exemple, l'énoncé "allume la lumière" transporte l'intention de faire modifier l'état d'un objet. Une fois identifiée, cette intention est analysée par un module de décision qui agit en conséquence (p.ex., allumer la lumière ou demander de quelle lumière il s'agit).

Pour caractériser la commande vocale, il faut également identifier les informations pertinentes dans les énoncés, ce que l'on appelle *concepts* ou *slots* en anglais. On appelle le processus d'identification des concepts, la reconnaissance d'attributs ou manière plus usuelle le *slot-filling* en anglais (Tur et De Mori, 2011). Une bonne identification des concepts peut contribuer à une meilleure classification des intentions, et vice versa, surtout par les modèles de réseaux de neurones profonds multi-tâches.

Les concepts composant l'énoncé peuvent être représentés de façon hiérarchique ou sous leur forme plane (Tur et De Mori, 2011). La figure 5.2a montre une représentation *hiérarchique* des slots pour la commande "turn off the light in the kitchen upstairs" (éteins la lumière dans la cuisine à l'étage) d'une part, La figure 5.2b montre une représentation *plane* des concepts pour la même commande.



(a) Représentation hiérarchique

set_device(action = TURN_OFF = turn off,
device = LIGHT = *the light*,
location-room = KITCHEN = "in the kitchen",
location-floor = 1 = "upstairs")

(b) Représentation linéaire

FIGURE 5.2 – Représentations des concepts d'une commande vocale

Le désavantage principal de la représentation plane comme 5.2b, est que les concepts ne sont pas liés entre eux. Par exemple *location* n'est pas rattaché à *device*. Dans le cas où il y a des objets identiques dans des pièces différentes, la représentation hiérarchique peut contribuer à la résolution des ambiguïtés et spécifier la localisation correcte de l'objet. Dans le cas où le système doit par exemple choisir entre la lampe de la cuisine, ou la lampe du salon, la structure imbriquée de la figure 5.2b montre qu'il s'agit de la lampe de la cuisine. Par contre, il est plus difficile d'établir un modèle qui pourrait extraire des intentions/slots et à la fois les structurer dans une hiérarchie. En outre, dans cette étude, la cible est tout d'abord de créer un modèle qui est capable de tenir compte de la variabilité linguistique, syntaxique des utilisateurs. Enfin, extraire la structure hiérarchique complète est plus difficile pour un sys-

tème automatique. Par conséquent, pour cette étude, nous avons choisi une représentation plane.

L'état est une structure de données qui contient les connaissances acquises jusqu'à présent (Williams et coll., 2014). Étant donné que notre application concerne uniquement des commandes vocales isolées, nous n'avons pas appliqué de compréhension des énoncés au fil de l'eau (DST - *Dialogue State Tracking*) dans ces travaux.

5.2 Définition de l'espace sémantique des commandes vocales

Dans le chapitre 2, nous avions décrit les habitats intelligents Amiqual4Home (section 2.2.8) et DOMUS (section 2.2.7). En tenant compte du contexte de l'habitat intelligent Amiqual4Home (chapitre 2, section 2.2.8), nous avons défini les catégories générales des intentions et concepts de notre espace sémantique pour les commandes vocales.

5.2.1 Intentions

Nous répartissons les intentions en quatre catégories :

— l'intention `contact` qui permet à un utilisateur de passer un appel ;

— l'intention `set` qui permet de modifier l'état des objets dans l'habitat intelligent ;

— l'intention `get` qui permet à l'utilisateur d'interroger l'état des objets ainsi que les caractéristiques du monde dans son ensemble ;

— et enfin l'intention `check` qui permet à l'utilisateur de vérifier si un objet se trouve dans un certain état.

Par exemple, la commande "l'aspirateur est-il allumé" est une intention `check`, qui est une question fermée, tandis que "où est l'aspirateur", est une question ouverte et que `get` ("Comment est la porte") est une intention qui attend un état d'appareil comme réponse.

5.2.2 Concepts

Les concepts sont divisés en huit catégories de base :

— l'action (`action`) à exécuter,

— l'appareil (`device`) à actionner,

— l'emplacement (`location`) de l'appareil ou de l'action,

— la personne (`person`) ou l'organisme (`organization`) à contacter,

— un composant (`component`) de l'appareil,

— un attribut (`setting`) d'appareil,

— une propriété (`property`) d'un appareil ou d'ambiance.

Sur la base de ces catégories et de leur combinaison, nous avons pu définir 17 étiquettes de concept. À partir des 4 catégories générales d'intention, nous définissons 8 intentions, que nous détaillerons dans le chapitre 6 consacré à la création du corpus de test VocADom@A4H et la génération du corpus artificiel VocADom@ARTIF, qui sera lui automatiquement annoté de concepts et d'intentions. Nous avons créé ces 2 ensembles de données pour surmonter le défi d'un manque de données d'apprentissage et d'évaluation spécifique au domaine domotique en langue française.

5.3 Acquisition de corpus

Le corpus spécifique au domaine réaliste VocADom@A4H (chapitre 6, section 6.2) a été acquis dans un habitat intelligent en utilisant une technique de magicien d'Oz pour simuler un appartement réagissant aux commandes vocales de participants naïfs. Ce corpus a été enregistré non seulement pour développer et tester les méthodes de SLU, mais aussi pour servir à des recherches sur la RAP en conditions distantes. VocADom@A4H constituera donc le corpus de test par défaut pour toutes nos expérimentations décrites au cours des chapitres à venir.

Bien que VocADom@A4H présente l'avantage d'être un corpus enregistré dans des conditions réalistes, il s'est avéré bien trop restreint pour l'apprentissage supervisé de réseaux de neurones. Pour augmenter la taille des données d'apprentissage, nous avons donc généré artificiellement un corpus d'énoncés annotés d'étiquettes sémantiques qui sera présenté au chapitre 6 en section 6.3.

En générant ce corpus artificiel VocaADom@Artif nous avons ciblé l'extraction d'intentions de commandes vocales *sans* contexte linguistique et contenant une seule intention par énoncé. Nous nous sommes concentrés sur la question de la syntaxe et la variabilité linguistique comme c'était le cas pour le corpus de test VocADom@A4H. La partie acoustique du corpus artificiel a été générée par synthèse vocale.

Nous décrivons ces deux corpus dans le chapitre 6. Ils serviront comme données d'apprentissage et d'évaluation pour notre approche de référence de SLU séquentielle et notre approche cible de SLU E2E. Il convient de souligner que les données *réelles* du corpus de test ne font *pas* partie des données *artificielles* de l'ensemble d'apprentissage. Ceci nous place dans le cas difficile mais réaliste de données d'apprentissage ne provenant pas de la même distribution que les données d'évaluation.

5.4 Architectures de SLU retenues

Pour vérifier quels avantages l'approche SLU E2E peut offrir par rapport à une approche séquentielle, nous construisons une approche SLU séquentielle que nous comparerons à une approche SLU E2E. La figure 5.3 comprend un aperçu schématique des deux approches.

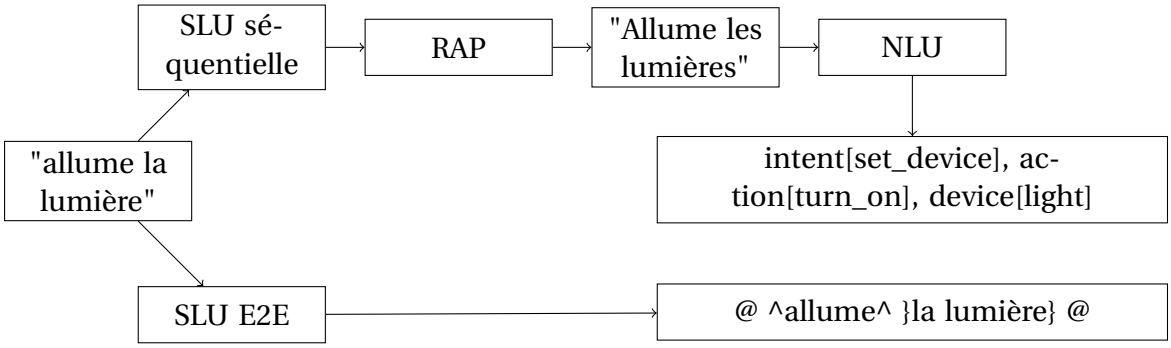


FIGURE 5.3 – Comparaison entre les architectures SLU séquentielle et SLU de bout en bout

5.4.1 SLU séquentielle

Notre approche de référence de SLU séquentielle se compose d'un module de RAP et d'un module de NLU

L'approche de RAP que nous avons employée, repose sur un modèle HMM-DNN qui reste à l'état de l'art en ce qui concerne le français. Il s'agit de l'outil Kaldi (Povey et coll., 2011a, 2015) (chapitre 3, section 3.1.3). Nous avons suivi la recette standard Kaldi pour la modélisation acoustique basée sur DNN nnet2, que nous avons également appliquée dans le chapitre 7. Le module de RAP a été appris sur un grand ensemble de données de parole du français extérieur à la tâche. Les hypothèses de transcriptions de sortie sont les entrées du module de NLU.

Concernant la NLU, les modèles de l'état de l'art, qu'ils soient CRF (Jeong et Lee, 2008) ou neuronaux (Mesnil et coll., 2015; Bapna et coll., 2017; Liu et Lane, 2016; Huang et coll., 2017), abordent le problème de NLU comme une *tâche d'étiquetage de séquence*. Cela signifie que les données d'apprentissage doivent être *alignées* pour associer chaque mot à une étiquette de concept. Malgré l'efficacité des modèles NLU aligné, ce type d'alignement ne peut pas être appliqué dans le cas d'une approche de bout-en-bout lorsque les données d'entrée consistent en une parole spontanée avec des disfluences qui provoquent souvent des erreurs de suppression et d'insertion de RAP. Par ailleurs, les modèles alignés sont forcés de travailler au niveau des *tokens* ce qui limite grandement leur capacité d'abstraction voire même leur capacité de traiter des concepts dont les tokens s'entrecroisent.

Notre approche a donc consisté à définir le problème de NLU comme un problème de traduction en utilisant un modèle de séquence à séquence avec attention pour rendre l'apprentissage indépendant des données alignées (Desot et coll., 2019b,a). L'utilisation de données non alignées offre la flexibilité nécessaire pour extraire des étiquettes de concept à partir de transcriptions imparfaites. C'est notre stratégie pour réduire les erreurs de transcription provenant de la RAP. Nous décrivons le modèle plus en détail dans le chapitre7.

5.4.2 SLU de bout en bout (E2E)

Notre approche E2E repose sur un modèle neuronal encodeur-décodeur où l'encodeur est composé de couches de CNN suivies de couches de bi-LSTM avec sous-échantillonnage.

Le décodage est réalisé à l'aide d'un mécanisme hybride d'attention et de CTC (*Connectionist Temporal Classification*). Le principe de notre approche consiste, à partir des signaux de parole, à produire une transcription "enrichie" par des symboles indiquant quelles intentions et quels concepts sont exprimés et délimitant les *token* concernés.

Puisqu'il s'agit de transcriptions enrichies, nous avons utilisé l'outil de RAP ESPnet (Watanabe et coll., 2018) (chapitre 3, section 3.1.4.2). Le modèle E2E est appris à base de transcriptions *enrichies* de symboles représentant les intentions et les *slots*. Notre approche pour prédire des concepts à partir du signal d'entrée en utilisant des transcriptions enrichies a été inspirée par Ghannay et coll. (2018) (chapitre 4, section 4.2.2). Nous montrerons que, contrairement à l'approche séquentielle, les intentions et les concepts peuvent être déduits directement de l'entrée vocale brute, inspirée de l'approche de (Serdyuk et coll., 2018) (chapitre 4, section 4.2.1), comme nous l'avons précisé dans l'introduction de ce chapitre.

L'ensemble de données de parole utilisé pour le module de RAP de l'approche de SLU séquentielle est également utilisé comme données d'apprentissage pour l'approche SLU E2E. Néanmoins, ces données sont complétées par des données de synthèse vocale. À cette fin, de la parole *synthétique* a été générée sur le corpus artificiel complet, qui a été ajoutée à l'ensemble de données d'apprentissage acoustique *réel*, dans l'esprit de l'étude de Li et coll. (2018) (chapitre 6, section 6.3.4).

5.5 Méthode d'évaluation

Pour mesurer les différences entre les approches de SLU E2E et séquentielle, nous utilisons les indicateurs usuels de performance NLU. Par ailleurs, pour prendre en compte l'effet des données artificielles, nous proposons des indicateurs au niveau acoustique et prosodique. Dans la suite, nous expliquons la méthode d'évaluation et les mesures utilisées.

5.5.1 Évaluation de la RAP

Les deux outils de RAP que nous avons utilisés (Kaldi et ESPnet en mode RAP) dans cette étude sont évalués à l'aide du taux d'erreurs WER. Cette mesure est définie comme la somme des nombres de mots *insérés I* (Insertions), de mots *substitués S* (Substitutions) et de mots *supprimés D* (Deletions) rapportée au nombre de mots N d'une transcription de référence qui a été produite manuellement par un expert. L'alignement entre la référence et l'hypothèse de transcription est obtenue par programmation dynamique avec comme objectif de trouver l'alignement amenant au WER minimal. Le WER se calcule comme suit :

$$\text{WER} = \frac{I+S+D}{N} \times 100 \tag{5.1}$$

5.5.2 Évaluation de la compréhension du langage

L'évaluation de la reconnaissance d'intention peut s'effectuer en considérant cette tâche comme une tâche de *classification*. Il s'agit d'une comparaison entre les classes d'intentions

de référence et d'hypothèse. Comme mesure sur les intentions nous avons donc utilisé les mesures classiques en classification, c'est à dire le rappel = $\frac{TP}{TP+FN}$ et la précision = $\frac{TP}{TP+FP}$, où TP sont les prédictions vrais positifs (*True Positive*), FP les faux positifs (*False Positive*), et FN les faux négatifs (*False Negative*). La *précision* exprime la proportion des intentions correctement prédites parmi l'ensemble des *prédictions* tandis que le *rappel* exprime le taux de prédictions correctes parmi l'ensemble des instances à prédire. Pour faire une synthèse de ces deux mesures, il est usuel de calculer la *F-mesure* ou *mesure F1* (*F1 score*) comme suit,

$$F1 = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} \tag{5.2}$$

En ce qui concerne les concepts, nous avons utilisé deux stratégies. L'une consiste à voir la reconnaissance de concept comme un problème de classification. C'est à dire que chaque *token* doit recevoir une étiquette de classe. On peut ainsi utiliser les mesures de précision, rappel et F-mesure. Cette stratégie est particulièrement adaptée aux approches alignées pour lesquelles la tâche est vue comme une tâche d'étiquetage de séquence. Cependant, cette stratégie ne peut s'appliquer à l'approche de NLU *non-alignée*. En effet, dans ce cas, la tâche de NLU est vue comme une tâche de génération et non comme un étiquetage. Dans ce cas, à l'instar de Simonnet et coll. (2017), nous utilisons le *Concept Error Rate* (CER) [1]. Dans Hahn et coll. (2008), le CER est défini comme le rapport de la somme des *concepts* supprimés, insérés et substitués par rapport à un alignement de Levenshtein pour une chaîne de concepts de référence donnée. Nous avons calculé le CER d'une manière similaire au calcul de WER, mais nous n'avons pas pris en compte l'ordre des séquences d'étiquettes de concept, car une séquence de référence comme par exemple, `action, device`, fournit les mêmes informations qu'une hypothèse « inverse »`device, action`.

L'exemple dans la figure 5.4 explique comment nous avons calculé le CER pour une transcription de sortie de RAP *'minouche allume la lumière dans la cuisine au premier étage'*, avec la séquence d'hypothèse d'étiquettes de concept (erronée), `action, device-setting, device`, par rapport à la séquence de référence d'étiquettes de concept `action, device, location-room, location-floor`.

Le CER est calculé pour les approches séquentielle et de bout en bout. Dans le cas de bout en bout, vu qu'il s'agit de transcriptions enrichies de symboles, représentant les concepts et les intentions, nous ne prenons en compte que les séquences de symboles de concepts de référence et d'hypothèse. La figure 5.5 montre ainsi comment la référence et l'hypothèse sont « vidées » des *tokens* de la transcription pour ne garder que les symboles de concept sur lesquels le CER est calculé. Dans cet exemple, l'hypothèse a supprimé le symbole `}` et obtient donc un score WER de 50%.

1. À ne pas confondre avec le *Character Error Rate*.

– Transcription de sortie du module de RAP :

'minouche allume la lumière dans la cuisine au premier étage'

– Référence :
`action, device, location-room, location-floor`

'allume' : `action`
'la lumière' : `device`
'dans la cuisine' : `location-room`
'au premier étage' : `location-floor`

– Prédiction (erronée) :
`action, device-setting, device`

Phase 1 : triage des étiquettes des séquences en ordre alphabétique,

– Référence :
`action, device, location-floor, location-room`
– Prédiction :
`action, device, device-setting`

Phase 2 : distance de Levenshtein,

– Référence :
`action, device, location-floor, location-room`
– Prédiction :
`action, device, device-setting, _____________`
`C C S D`

Phase 3 : calcul du CER,

```
((S + D) / N) * 100
((1 + 1) / 4 ) * 100 = 50%
```

FIGURE 5.4 – Exemple de calcul du *Concept Error Rate*.

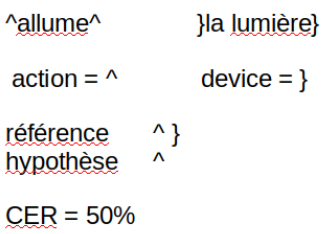


FIGURE 5.5 – SLU E2E - exemple de calcul du CER lorsque le concept est représenté par un symbole

5.5.3 Évaluation de la pertinence des données acoustiques générées

Étant donné que nous avons utilisé un synthétiseur de parole pour obtenir le signal acoustique du corpus artificiel, nous avons cherché à évaluer la qualité de ces signaux en comparant la parole artificielle aux données de parole réelle de l'ensemble de test VocADom@A4H. Nous avons procédé par mesure de la distance acoustique entre ces deux ensembles de données. À cette fin nous avons appliqué la technique de déformation temporelle dynamique (DTW - *Dynamic Time Warping*). Le calcul de la fonction DTW est une technique qui a été introduite en RAP il y a quelques décennies par Sakoe *et al.* (Sakoe et Chiba, 1978) et qui est toujours utilisée (Dhingra et coll., 2013; Su et coll., 2019). Étant donné que l'alignement temporel de différents énoncés est un problème central pour la mesure de distance des séquences vocales, la DTW mesure la similitude entre deux séries temporelles qui peuvent varier ou être *déformées* au cours du temps. L'alignement optimal est cherché pour une série temporelle qui est déformée de manière non linéaire en l'étirant ou en la rétrécissant le long de son axe temporel. Cette similitude est mesurée en trouvant la distance d'édition minimale. Par conséquent, pour deux séquences identiques le résultat du calcul de DTW sera zéro (Muda et coll., 2010; Sakoe et Chiba, 1978).

En utilisant la fonction de DTW, les séquences temporelles Q et C de longueur respectivement n et m, $Q = q_1, q_2, ..., q_i, q_n$ et $C = c_1, c_2, ..., c_j, c_m$, sont alignées dans une matrice de (n, m). n est le nombre de trames du premier signal et m le nombre de trames du deuxième signal. L'élément (i, j) de la matrice contient la distance $d(q_i, c_j)$ entre les deux points q_i et c_j . Chaque élément (i, j) de la matrice correspond à l'alignement entre les points q_i et c_j . La distance cumulée est mesurée par :

$$D(i,j) = min[D(i-1,j-1), D(i-1,j), D(i,j-1)] + d(i,j) \tag{5.3}$$

La distance absolue entre les valeurs des deux séquences est calculée en utilisant la distance euclidienne :

$$d(q_i, c_j) = (q_i - c_j)^2 \tag{5.4}$$

5.6 Analyse fine des propriétés para-linguistiques et acoustiques utilisées par le modèle

Un autre objectif est de déterminer quelles sont les propriétés du signal acoustique qui contribuent à la prédiction des concepts et des intentions lorsqu'on utilise une approche de SLU de bout en bout, c'est ce que nous présenterons au (chapitre 9). Pour cette tâche, nous mesurons la corrélation existante entre les performances de SLU séquentielle et E2E et les valeurs de *F0* et *d'énergie* du signal de nos données de test. Nous avons calculé ces valeurs en utilisant l'outil Praat [2] en considérant également l'impact du bruit de fond et les différences au niveau *inter-locuteur*.

Nous avons utilisé les *coefficients de corrélation de Pearson* et de corrélation des *rangs de Spearman* entre les valeurs de *F0* et *d'énergie* d'une part, et le WER et le CER (*Concept Error Rate*), d'autre part. Le coefficient de corrélation de Pearson r mesure la force de l'association entre deux variables x et y en supposant une distribution normale des valeurs. Plus les points de coordonnées x, y sont disposés à proximité immédiate d'une ligne droite, plus la force d'association entre les variables est élevée. Une corrélation est positive lorsque les valeurs des variables évoluent dans la même direction et négative lorsque elles évoluent dans une direction opposée. La corrélation de Pearson est calculée comme suit où n dénote le nombre d'éléments :

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{n}\sqrt{\left(x_i - \overline{x}\right)^2}\sum_{i=1}^{n}\sqrt{\left(y_i - \overline{y}\right)^2}} \tag{5.5}$$

r varie entre -1 et 1. Il n'y a pas de corrélation lorsque r est égal à 0, 1 indique une forte corrélation et -1 une forte corrélation négative.

Le *coefficient de corrélation des rangs de Spearman*, r_s , mesure la corrélation entre les *valeurs de rang* de deux variables. Tandis que la corrélation de Pearson évalue les relations linéaires, la corrélation de Spearman évalue les relations monotones. Dans une relation monotone, les variables ont tendance à se déplacer dans la même direction relative, mais pas nécessairement à une vitesse constante, ce qui est bien le cas dans une relation linéaire.

La corrélation de Spearman est calculée comme suit, n dénote le nombre d'éléments :

$$r_s = 1 - \frac{6\sum_{i=1}^{n}\left(x_i - y_i\right)^2}{n(n^2-1)} \tag{5.6}$$

Pour déterminer si le coefficient de corrélation est significatif, la valeur de p est souvent utilisée. Elle se calcule dans les tests d'hypothèse pour déterminer s'il faut rejeter ou non une hypothèse nulle. La valeur de p pour le coefficient de corrélation de Pearson ou de Spearman ($coef$) utilise la loi de distribution t,

$$t = r\sqrt{\frac{n-2}{1-coef}} \tag{5.7}$$

2. https://www.fon.hum.uva.nl/praat

Nous rejetons l'hypothèse nulle, $H_0 : coef = 0$, si la valeur de p est inférieure à 0.05 ($p < 0.05$). La valeur de p où l'hypothèse $H_1 : coef \neq 0$ (test bilatéral), se calcule comme suit :

$$p = 2 * P(T > |t|), \tag{5.8}$$

où P désigne la probabilité et T suit une distribution t avec $n-2$ degrés de liberté. Au cours du chapitre 9, nous distinguons entre des corrélations pour lesquelles $p < 0.05$ que nous marquons avec *, et des corrélations pour lesquelles $p < 0.01$ qui sont dénotées avec **.

5.7 Analyse de la robustesse aux variations lexicales et grammaticales

Pour mesurer la robustesse des modèles aux variations lexicales et grammaticales, nous avons défini deux tests.

Le premier test s'intéresse à l'impact des *mots hors vocabulaire* (OOV - *Out Of Vocabulary*) de l'ensemble de test VocADom@A4H, sur les performances de SLU séquentielle et de SLU E2E (Chapitre 9). Pour mesurer l'impact d'un taux de mots hors vocabulaire augmenté, nous avons progressivement, en 4 étapes, remplacé le vocabulaire des énoncés par des types de mots qui n'apparaissent pas dans l'ensemble de données d'apprentissage. Par étape, nous avons ciblé un type de concept avec une haute fréquence dans le corpus de test, et ensuite remplacé ses mots de valeur par des mots hors vocabulaire, mais appartenant au même concept. À chaque nouvelle étape, nous avons accumulé les mots hors vocabulaire de l'étape précédente.

Le deuxième test s'intéresse à la variabilité syntaxique de la parole de nos utilisateurs cibles. Nous mesurerons la robustesse des modèles de SLU séquentielle et de SLU E2E, en prédisant les concepts et les intentions sur des énoncés des données d'évaluation dont les syntagmes sont placés d'une manière différente de celle des énoncés du corpus d'apprentissage. Là encore, la stratégie est d'augmenter le nombre de perturbations de manière graduelle, en 2 étapes, pour mettre en évidence une tendance. Dans une première étape, des verbes fréquents ont été remplacés par des structures syntaxiques plus complexes et moins fréquentes. Dans l'étape suivante, des disfluences telles que des répétitions, des corrections etc. ont été ajoutées, et cumulées avec les modifications introduites lors de la première étape.

L'hypothèse de test est que l'apprentissage multi-tâche du modèle de SLU E2E, qui combine CTC et attention, peut renforcer la robustesse sur des données d'évaluation présentant une *variabilité linguistique* augmentée.

5.8 Conclusion

L'objectif de cette thèse est de comprendre quels avantages une approche SLU de bout-en-bout(E2E) peut offrir par rapport à une approche en pipeline classique. Dans ce chapitre,

nous avons posé les questions de recherche et donné un aperçu des étapes de la démarche que nous avons définie pour y répondre. Nous avons également présenté les outils nécessaires pour construire notre approche de SLU E2E cible et l'approche de référence séquentielle.

Le point essentiel de cette thèse est que la démarche d'évaluation ne se limite pas seulement à concevoir un modèle SLU performant mais aussi à analyser quelles propriétés du signal acoustique entrent en jeu pour obtenir cette performance.

Par ailleurs, la démarche se place dans un cadre résolument difficile en considérant, au contraire des démarches d'apprentissage classique, le cas réaliste où les données d'apprentissage et de test sont issues de sources différentes.

CHAPITRE 6

Collecte et génération de corpus oral pour la commande vocale

Au cours des chapitres précédents, nous avons montré que les méthodes de SLU, notamment les modèles de réseaux de neurones profonds nécessitent des corpus de taille suffisante et adaptés aux besoins de notre étude. Dans le chapitre 2, nous avons pu constater que, pour une tâche de NLU, la taille du vocabulaire et le nombre d'intentions des corpus disponibles sont plutôt réduits. Pour les commandes vocales, il s'agit très souvent de phrases très courtes qui doivent suivre un motif syntaxique strict pour faciliter l'interprétation par un système de NLU ou de SLU.

Cependant, nos utilisateurs cibles sont des personnes âgées et des études ont montré que ce type de population a tendance à s'écarter d'une grammaire trop rigide (Möller et coll., 2008; Takahashi et coll., 2003; Vacher et coll., 2015). Cela signifie que nos modèles doivent être entraînés sur des corpus de commandes vocales avec des énoncés assez variés tant du point de vue syntaxique que lexical. Dans ce chapitre, nous apporterons une réponse à la question de recherche énoncée au chapitre 5 : **comment apprendre des modèles profonds à partir d'une faible quantité de données initiales ?**

Nous commencerons par définir quelles sont les caractéristiques que devra satisfaire le corpus que nous utiliserons en s'appuyant sur notre définition de l'espace sémantique présentée au chapitre 5 en section 5.2). Ensuite, nous présenterons le processus que nous avons suivi pour enregistrer le corpus réaliste VocADom@A4H. Celui-ci implique plusieurs locuteurs interagissant par commande vocale avec un habitat intelligent grâce à un magicien d'Oz.

Le corpus obtenu, VocADom@A4H, étant de taille limitée, nous avons été amenés à créer un second corpus en mettant à profit l'expérience acquise lors de ce premier enregistrement notamment en ce qui concerne le contenu sémantique. Ce second corpus, VocADom@ARTIF, se compose d'un ensemble de phrases automatiquement étiquetées de concepts et d'intentions appartenant au domaine de la domotique. Dans ce chapitre, nous décrirons la méthode utilisée pour créer ce corpus d'apprentissage artificiel afin de l'utiliser dans un contexte de SLU séquentielle. Pour l'utiliser dans un contexte de SLU de bout en bout, nous expliquerons comment nous avons enrichi les transcriptions d'étiquettes symboliques de concepts et d'intentions. Nous estimerons également la distance entre ce corpus et l'ensemble réel VocADom@A4H. Enfin, nous décrirons le processus par lequel nous

avons sélectionné des énoncés sans intention afin de constituer un ensemble d'exemples d'apprentissage négatifs pour rendre le système capable d'identifier des énoncés hors du domaine de la domotique et pour étendre la quantité d'énoncés de parole réelle.

6.1 Caractéristiques attendues

Dans le cadre du projet VocADom, il a été décidé qu'une commande vocale se compose d'un mot-clé suivi d'une intention. La présence du mot-clé (ou *Wake-up Word*) est nécessaire pour que le système comprenne que l'on s'adresse à lui et que la suite de l'énoncé est une commande vocale. Considérons la commande vocale suivante : « Ichefix, est-ce que la porte est ouverte ? ». Le mot-clé « *Ichefix* », suivi d'une commande, activera le système domotique et est utilisé comme identifiant de l'habitat intelligent. Son utilisation permet à la maison intelligente de savoir si la commande vocale s'adresse bien à elle et non à un autre habitant de la maison. Mise à part le mot-clé, la commande vocale doit transporter une intention claire. Dans l'exemple « Ichefix, est-ce que la porte est ouverte ? », l'intention doit être classifiée comme `check_device`, une requête demandant de vérifier si un appareil/objet (*la porte*) est dans un état précis (*fermée*). Enfin, la commande vocale doit faire usage de termes liés à des concepts tels que `device` (la porte) ou `device_setting` (ouverte).

Par la suite, pour faciliter le travail de l'étape de décision, nous partons du principe que :

— le mot-clé est toujours le premier de la phrase (nous verrons que cette contrainte est difficile à respecter à l'usage) ;

— la commande vocale ne concerne qu'un seul `device` (la porte) ou un groupe de `device` (les portes) de même type. Par exemple la commande » Ichefix, est-ce que la porte et la télé sont ouvertes ? » n'est pas une commande acceptable ;

— chaque énoncé contient une seule commande. Par exemple, la commande « Ichefix, éteint la lumière et ferme les volets » ne sera pas reconnue. Il s'agit de deux commandes qui doivent faire l'objet de deux énoncés séparés.

6.1.1 Mots-clés

Le choix des mots-clés possibles doit obéir à des critères bien précis. Les mots-clés devront comporter de 3 à 4 syllabes ce qui permet une durée suffisante pour assurer une RAP correcte. De plus, comme les utilisateurs cibles sont des personnes âgées, les mots-clés doivent être particulièrement faciles à prononcer et à reconnaître. L'étude Aman (2014) a montré que, dans l'alphabet phonétique international (IPA), les consonnes s, ʃ, m, ʁ et *l* et les voyelles i, y, u, ɛ et *e* étaient les mieux reconnues par les systèmes de RAP quand il s'agissait de voix âgées. Nous avons donc privilégié ces dix phonèmes dans le choix des mots-clés.

Les mots respectant ces contraintes ont été extraits automatiquement d'un dictionnaire, ensuite filtrés et finalement discutés entre les chercheurs du projet VocaDom pour obtenir la liste suivante de mots-clés : *téraphim, ulysse, ichefix, chanticou, vocadom, écirrus, hé cirrus, allo cirrus, allo messire, dis vesta, dis hestia, dis bérénio, dis téraphim, dis vocadom, mi-*

nouche. Ces mots clés conservés ont soit un rapport avec le projet (*vocadom*) soit un rapport avec la maison (Vesta – déesse romaine du foyer –, Théraphim – dieu du foyer sémitique –, Minouche – nom populaire donné à un chat).

6.1.2 Intentions

Dans le chapitre 5, section 5.2.1 nous avons décrit et défini les composants sémantiques de base de notre espace sémantique. Au niveau des intentions nous avons défini 8 classes d'intentions à base des 4 catégories générales définies par l'espace sémantique. Le tableau 6.1 présente un aperçu des intentions en donnant pour chaque intention un exemple ainsi que sa fréquence dans le corpus artificiel VocADom@ARTIF et l'ensemble de test réaliste VocADom@A4H.

TABLE 6.1 – Intentions dans les corpus VocADom artificiel et réaliste - exemples et fréquences

Intention	**Énoncé**	**Fréquence**	
		Artif.	**Réel.**
`Check_device`	*minouche est-ce que la fenêtre est ouverte?*	2754	284
`Contact`	*vocadom appelle un médecin*	567	114
`Get_room_property`	*bérénio quelle est la température?*	9	3
`Get_world_property`	*ulysse quelle heure est-il?*	9	3
`None`	*la fenêtre est ouverte*	-	4135
`Set_device`	*hestia baisse les stores*	63,288	2178
`Set_device_property`	*ichefix diminue le volume de la télé*	7290	9
`Set_room_property`	*chanticou diminue la température*	3564	21

6.1.3 Concepts

Nous avons défini 17 catégories à base des 8 catégories générales définies par l'espace sémantique au chapitre 5 en section 5.2.2). Ces concepts peuvent être représentés de manière hiérarchique comme le montre la figure 6.1. Le tableau 6.2 présente un aperçu des concepts avec pour chaque intention un exemple et sa fréquence dans le corpus artificiel VocADom@ARTIF et dans l'ensemble de test réaliste VocADom@A4H. Beaucoup de concepts sont liés à la localisation alors que d'autres apparaissent en fait assez peu dans le corpus réel. Le corpus artificiel pourrait donc permettre de compenser ce déséquilibre. Bien entendu, toutes les commandes sont conçues pour être compatibles avec une approche de "*slot-filling*".

TABLE 6.2 – Concepts dans les corpus VocADom artificiel et réaliste - exemples et fréquences

Concept	**Énoncé**	**Fréquence**	
		Artif.	**Réel.**
`action`	ichéfix *diminuer* chauffage	70332	2211
`device`	téraphim stoppe *la télévision*	68769	2473
`device-component`	hestia change *de chaîne*	7290	5
`device-setting`	hé cirrus est-ce que la bouilloire est *éteinte*	2579	284
`location-floor`	dis vocadom ferme les stores de la chambre *du haut*	47643	805
`location-house`	euh minouche éteins toutes les lumières *de l'appartement*	47643	9
`location-inroom`	vocadom éteins la lumière *du plafond*	522	34
`location-room`	éteindre la lumière *de la salle de bain*	71475	1055
`organization`	minouche appelle *le supermarché*	273	4
`person-name`	vocadom appelle *Marie*	269	0
`person-occupation`	minouche appelle *un médecin*	60	2
`person-relation`	téraphim s'il te plaît appelle *ma fille*	29	0
`room-property`	dis bérénio diminuez *la température*	3573	24
`value-artist`	vocadom joue *David Bowie* dans la radio	1215	0
`value-numeric`	hestia mets *la deux*	1944	1
`value-qualitative`	ulysse baisse *un peu* la lumière	3840	2
`world-property`	ichéfix quelle est *la température*	21	3

6.2 Enregistrement du corpus réaliste VocADom@A4H

6.2.1 Procédure et déroulement des enregistrements

L'objectif de la collecte de données du corpus multimodal VocADom@A4H était d'être utile au développement et à l'évaluation de nouveaux systèmes domotiques d'interaction vocale en contexte. Le défi a consisté à enregistrer des données collectées nécessaires pour les différentes tâches de traitement tout en étant suffisamment représentatives d'une variété de situations domestiques réalistes. Le protocole a donc été conçu pour collecter des données pour plusieurs tâches :

— localisation humaine,

— reconnaissance de l'activité humaine,

— rehaussement de la parole,

— détection d'activité vocale,

— reconnaissance automatique de la parole

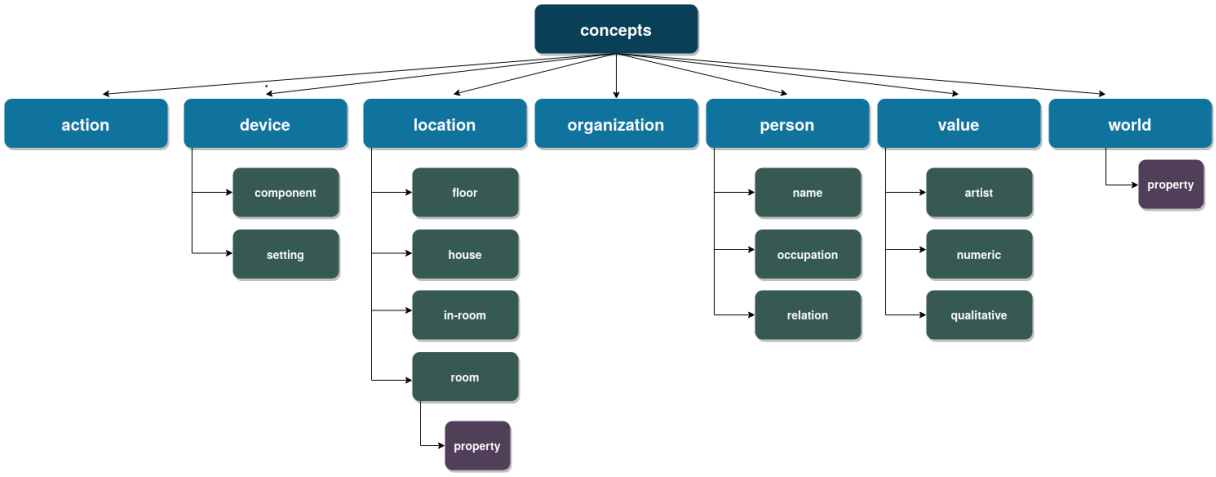


FIGURE 6.1 – Représentation hiérarchique des concepts et leurs attributs

— compréhension du langage naturel et finalement,

— prise de décision automatique.

Chaque tâche peut s'appuyer sur différentes sources de données multimodales, telles que les commandes vocales (extraites de la parole), les bruits (sons émis par les appareils domestiques) et les données fournies par les capteurs domotiques (Portet et coll., 2019).

Les enregistrements en conditions réalistes ont été effectués dans l'habitat intelligent Amiqual4Home. Cet habitat intelligent est décrit en détail dans le chapitre 2, section 2.2.8.

Pour collecter des données « réalistes », le défi consistait à provoquer chez le locuteur une émission de parole spontanée présentant une variation linguistique lexicale et syntaxique, tout en gardant un contrôle sur les énoncés. Pour répondre à ce défi, trois phases d'enregistrement ont été définies dans lesquelles les participants –jouant le rôle d'un habitant ou d'un visiteur– ont utilisé des commandes vocales pour communiquer avec la maison en effectuant des activités de la vie quotidienne. Les trois phases ont introduit différents degrés de spontanéité dans le comportement des participants et ont incorporé plusieurs conditions de perturbation sonore (conversations environnantes, aspirateur, télévision, ventilateur et douche).

— Phase 0 : Phase d'accueil du participant et choix du mot-clé. Dans cette phase le participant visitait l'appartement et prenait connaissance des possibilités d'interaction. C'est durant cette phase que la personne choisissait un mot-clé après les avoir tous lus à voix haute.

— Phase 1 : Instruction graphique donnée au participant. Le participant disposait seulement d'une instruction sous forme graphique afin qu'il puisse éliciter les commandes vocales spontanément sans qu'on lui fournisse des biais lexicaux. Pour cela, des images illustrant l'activité à effectuer avaient été fournies au participant qui a dû utiliser son propre vocabulaire pour construire des commandes vocales permettant d'accéder à une action disponible dans l'habitat. La figure 6.2 montre un participant en situation d'élicitation. La figure 6.3 montre un exemple d'image utilisée pour éliciter des ordres suivants :

— « Allume la lumière dans la cuisine »

— « Allume la bouilloire »

— « Allume le ventilateur dans le salon »

— « Allume la lumière à l'étage »

— « Est-ce que le ventilateur est allumé ? »

— « Éteins la lumière dans la chambre à coucher »

— « Éteins le ventilateur »

— « Éteins la lumière au rez-de-chaussée »

FIGURE 6.2 – Phase 1 de VocADom@4H - participant élicitant une commande vocale de la domotique

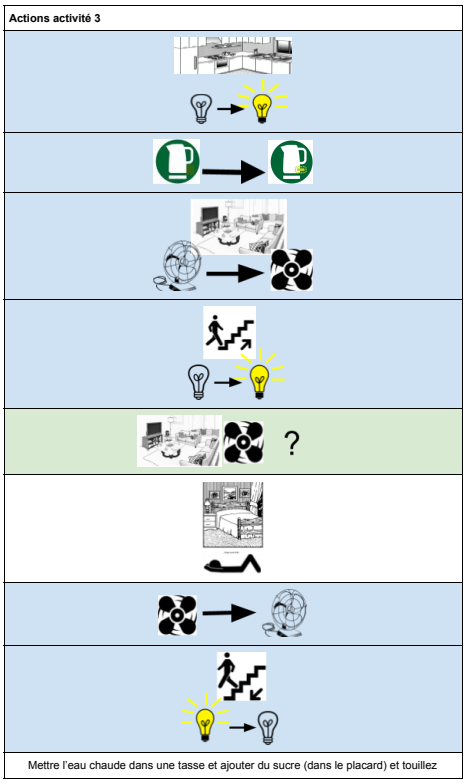


FIGURE 6.3 – Exemple d'image utilisée pour éliciter des ordres dans le cadre du recueil du corpus VocADom@A4H

— Phase 2 : Scénario à deux habitants : l'habitant de la maison intelligente reçoit un visiteur. Pour provoquer des dialogues et des situations à plusieurs habitants, une simulation de visite par un ou deux expérimentateurs était effectuée. Pour guider cette phase, un scénario était fourni aux participants sur les commandes à effectuer mais de manière descriptive sans qu'aucune grammaire stricte n'ait été imposée. Les deux locuteurs émettaient donc des commandes vocales sans restrictions de grammaire tout en interagissant avec la maison intelligente. Les activités à réaliser, écrites en forme de scénario, étaient par exemple « Allez dans la chambre et écoutez la radio », « Assurez-vous que les stores sont fermés avant de partir ».

FIGURE 6.4 – Phase 2 de VocADom@4H - l'habitant et son invité interagissant avec l'habitat

— Phase 3 : Commandes vocales en environnement sonore perturbé. Pour recueillir un corpus important en situation bruitée, les participants devaient lire des commandes vocales en présence d'un bruit de fond (aspirateur, radio, télévision, etc.) (figure 6.5). Chaque participant devait lire une liste de phrases contenant des commandes vocales et des phrases anodines comme par exemple :
  — « je ferme le verrou »
  — « non merci bien »
  — « vocadom baisser lumière »
  — « diminuer le chauffage »
  — « téraphim arrête la radio de la salle de bain »

FIGURE 6.5 – Phase 3 de VocADom@4H - commandes vocales avec un aspirateur en fonctionnement

6.2.2 Annotation du corpus

L'enregistrement d'un corpus dans un habitat intelligent est en soi un défi, mais les données obtenues ne seront utilisables qu'après une seconde tâche d'annotation. Les données vocales enregistrées ont été transcrites manuellement en utilisant l'outil Transcriber. Le format d'annotation est illustré ci-dessous (les balises indiquent les repères temporaires) :

<Sync time="2780.720000"/>
c'est ici
<Sync time="2781.572000"/><Sync time="2783.472000"/>
vocadom baisse le store du bureau
<Sync time="2785.348000"/><Sync time="2786.817"/>
dis hestia stoppe les stores du rez-de-chaussée
<Sync time="2789.316000"/>

Les transcriptions ont été annotées de classes d'intention et d'étiquettes de concept d'une manière semi-automatique. La grammaire de génération du corpus artificiel et l'annotation automatique de ses classes d'intentions et de concepts, nous ont servi pour créer l'outil d'annotation semi-automatique basé sur des expressions régulières.

Ensuite, ces pré-annotations sémantiques ont été corrigées à la main en utilisant une interface qui facilite les annotations des concepts et d'intentions (cf. figure 6.6).

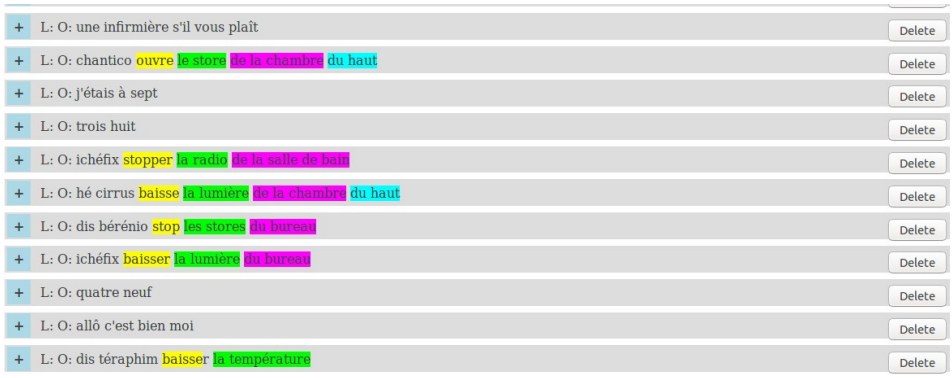


FIGURE 6.6 – Interface permettant d'annoter le corpus VocADom@4H

Le résultat de l'annotation était au format `Json`. L'exemple ci-dessous montre l'annotation de la phrase « *diminue la lumière* » :

{"text" : "diminue la lumière",
"sync_trs" : "1232.753",
"entities" : [
{
"start" : 0,

"end" : 7,
"entity" : "action",
"value" : "TURN_DOWN",
"text" : "diminue",
"color" : "#FFFF00"
},
{
"start" : 8,
"end" : 18,
"entity" : "device",
"value" : "light",
"text" : "la lumière",
"color" : "#00FF00"
}
],
"intent" : "set_device"
}

Cette phrase est annotée avec l'intention et les concepts (sous la liste *entities*). Chaque concept est représenté par une étiquette *slot-label* (le champ *slot*), sa valeur (*slot value*), son texte correspondant (*text*), l'index de début (*start*) et de fin (*end*) de ce concept dans la chaîne du texte. On indique aussi la couleur *color* qui est utilisée dans l'interface servant à l'annotation. Le champ *sync_trs* indique l'instant d'enregistrement que l'on peut utiliser pour des besoins de synchronisation avec le signal original. Le corpus a été normalisé, c'est-à-dire, traité automatiquement pour remplacer tous les mots-clés énumérés dans la section 6.1.1 et utilisés pour s'adresser au système par *KEYWORD* pour les expérimentations NLU.

6.2.3 Contenu du corpus VocADom@A4H

Dans son ensemble, le corpus VocADom@A4H comprend environ douze heures de signaux audio de 11 locuteurs (7 hommes et 4 femmes), et de traces provenant du système domotique, ses caractéristiques principales sont résumées dans le tableau 6.3.

En ce qui concerne les énoncés, leur répartition est affichée sur le tableau 6.4. Le corpus contient au total 6747 énoncés, dont 2612 commandes vocales et 4135 autres énoncés anodins. Les commandes vocales ont été étiquetées avec les classes d'intention et de concepts définis en section 6.1.2. Les tableaux 6.1 et 6.2 affichent respectivement un aperçu des fréquences des intentions et des concepts présents dans le corpus de test réaliste VocADom@A4H (Réal.)

Ce corpus *réaliste* VocADom@A4H servira de *corpus de test* pour évaluer l'ensemble de nos expérimentations de SLU comme indiqué dans les chapitres 7 et 8.

TABLE 6.3 – VocADom@A4H - caractéristiques des enregistrements par les 11 participants

Participant	Âge/Genre	Durée	Mot-clé choisi
S00	20-23 ans, M	01 :03 :54.29	vocadom
S01	20-23 ans, M	00 :48 :53.14	vocadom
S02	20-23 ans, M	01 :12 :26.27	hé cirrus
S03	20-23 ans, M	01 :11 :51.63	ulysse
S04	23-25 ans, F	01 :04 :45.98	téraphim
S05	<20 ans, F	01 :22 :59.19	allo cirrus
S06	23-25 ans, M	00 :55 :54.19	ulysse
S07	25-28 ans, M	01 :03 :53.78	ichefix
S08	23-25 ans, M	01 :13 :01.11	ulysse
S09	23-25 ans, F	01 :20 :06.13	minouche
S10	23-25 ans, F	01 :11 :03.03	hestia
Tous : 11	Moyenne : 23-25 ans Ensemble : 4 F - 7 M	Total :12h 28mn 45s	8/14

TABLE 6.4 – VocADom@A4H - répartition entre énoncés avec ou sans intention

VocADom@A4H	**Énoncés**	**Mots**	**Intentions**	**Concepts**	**Valeurs de concept**
Avec intention	2612	430	7	14	60
Sans intention	4135	1326	1	-	-
Ensemble des énoncés	6747	1462	8	14	60

6.2.4 Limites du corpus VocADom@A4H

L'espace sémantique de ce corpus réaliste est lié à l'équipement de l'habitat intelligent Amiqual4Home et aux actions rendues possibles par cet équipement. Les intentions étiquetées représentent également le type de commandes vocales que les utilisateurs prononceraient pour interagir avec cette maison intelligente.

Cependant, la taille totale de cet ensemble de données (environ 6k énoncés) est très limitée malgré le temps et les moyens très importants qui ont été nécessaires pour l'acquérir et l'annoter. Elle est insuffisante pour apprendre des modèles profonds de bout-en-bout effectuant un étiquetage automatique des concepts et une classification d'intentions à partir d'un signal acoustique. Par ailleurs, bien que l'accent ait été mis sur la spontanéité, les situations enregistrées ne couvrent pas l'ensemble des possibilités lexicales et syntaxique auxquelles on peut s'attendre dans un habitat intelligent avec des locuteurs de différentes générations.

Pour cette raison, nous avons mis en place une technique de génération automatique de textes pouvant créer un corpus artificiel à large couverture. Les énoncés générés ont été automatiquement étiquetés avec des concepts et des classes d'intention en faisant varier le vocabulaire et le style syntaxique. Le processus de développement de ce corpus est décrit dans les sections suivantes.

6.3 Génération du corpus artificiel VocADom@ARTIF

6.3.1 Génération automatique de texte

Pour générer un vaste ensemble de commandes vocales, nous avons utilisé une approche à base d'une grammaire hors-contexte (*Context free grammar* – CFG). Cette grammaire permet la génération de phrases ayant des caractéristiques (*features*) sémantiques qui sont attachées aux phrases de sortie finales.

La grammaire est regroupée par classe d'intentions dont chacune définit les énoncés comme une composition de leurs constituants possibles, avec des contraintes de génération. Par exemple, la *règle de grammaire générative* dans le tableau 6.5, définit les concepts de l'intention `set_device` et peut générer la commande *"ouvrir la fenêtre dans la cuisine"*. Le non-terminal `Slot_action` contient l'attribut `ACTION=?s` tandis que le non-terminal `Slot_device` contient l'attribut `ALLOWABLE_ACTION=?s`. Ces deux attributs sont contraints par la variable nommée `?s` d'avoir la même valeur, ce qui garantit que nous ne générons pas de phrases avec des appareils qui ne peuvent être actionnés par une action particulière. Par exemple " allumé les stores" ne peut pas être généré car l'action "allumé" ne fait pas partie des actions permises.

Les règles de plus bas niveau, contiennent d'autres attributs linguistiques tels que des contraintes d'accord en genre et en nombre.

TABLE 6.5 – Corpus artificiel VocADom@Artif- variation syntaxique, grammaire générative et annotation présentées sur un exemple

Énoncé *Ouvre la fenêtre dans la cuisine*
Variation syntaxique *Est-ce que tu peux ouvrir la fenêtre dans la cuisine?*

Annotation
`SET_DEVICE(ACTION=open="open",DEVICE=window="window",`
`LOCATION=room="kitchen")`

Règle de grammaire générative
`Intent_set_device[ACTION=?s,Location=?l, Device=?d]` →
`Slot_action[ACTION=?s, :ACTION_TYPES={}, AGR=?a]`
`Slot_device[ALLOWABLE_ACTION=?s,`
`Location=?l, Device=?d, ARTTYPE=def]`

De plus, des contraintes ont été utilisées sur l'emplacement des objets dans l'habitat pour éviter la production d'énoncés absurdes tels que *"allumer le lave-vaisselle dans la chambre à coucher"*. Pour permettre à ces contraintes de fonctionner, *l'unification* d'attributs a été appliquée. C'est le processus par lequel différents symboles dans les règles sont liés en fonction de leurs attributs. Si la règle définissant par exemple l'appareil *"lave-vaisselle"* contient la caractéristique "location = [room = "cuisine"]", la grammaire doit unifier cette caractéristique avec la même caractéristique attribuée à une pièce, afin de générer une phrase comme par exemple *"allume le lave-vaisselle dans la cuisine"*.

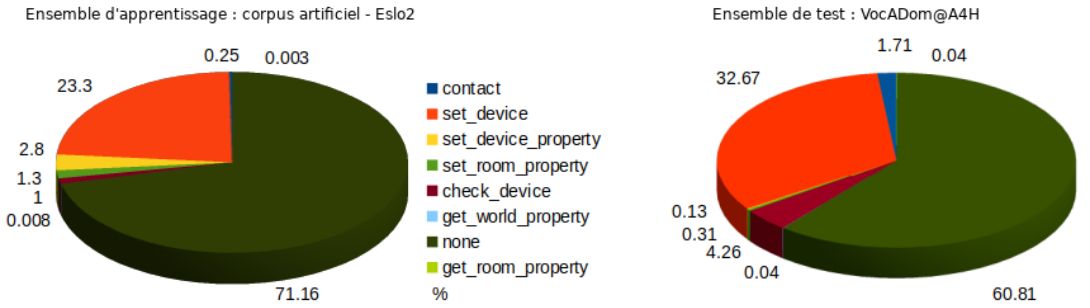


FIGURE 6.7 – Répartition des intentions pour le corpus d'apprentissage réunissant VocADom@Artif et ESLO2, et pour celui de test VocADom@4H

TABLE 6.6 – Caractéristiques des ensembles d'apprentissage considérés (OOV = mots d'ensemble de test hors d'ensemble d'apprentissage, intent. = intention, val. = valeur, perplex. = perplexité), les ensembles de test sont extraits du corpus VocADom@A4H

Ensembles d'apprentissage	**énoncés**	**mots**	**intent.**	**concept**	**val.**	**perplex.**	**OOV**	**Ensembles de test**
VocADom@Artif	77,481	187	7	17	69	124.41	307	Avec intent.
SWEET-HOME	1412	480	6	7	28	49.33	343	Avec intent.
ESLO2	161,699	29,149	1	-	-	151.90	211	Avec intent.*none*
Tous	240,592	30,821	8	17	69	372.06	235	Corpus complet

La grammaire du générateur du corpus artificiel a également été conçue pour générer une variation syntaxique, telles que les constructions interrogatives françaises introduites par l'expression (*"est-ce que"*) ou par l'inversion entre sujet et verbe (*"sont-elles…"*, *"sont-ils …"*). Cela comprend également des verbes modaux comme dans (*"peux-tu…"*), des marqueurs de politesse tels que (*"s'il vous plaît"*) et d'autres (le tableau 6.5 en présente un exemple). Tout comme pour l'ensemble de test (section6.2), chaque commande vocale comprend un mot-clé pour activer le système de commande vocale. En maximisant toutes les combinaisons d'étiquettes sémantiques qui produisent des énoncés, la grammaire génère environ 77000 phrases uniques, dont chacune est annotée avec une intention et ses concepts (Artif. dans le tableau 6.6). Un aperçu des intentions est présenté dans le tableau 6.1 et la figure 6.7.

Afin d'introduire des énoncés sans intention de commande vocale de la domotique (intentions *none*), nous avons ajouté le corpus ESLO2 au corpus VocADom@Artif dans notre ensemble d'apprentissage. ESLO2 est un corpus de français parlé acquis par l'université d'Orléans qui sera décrit en détail dans les sections suivantes.

6.3.2 Format aligné et non aligné

Pour l'évaluation du corpus artificiel VocADom@Artif, des modèles NLU CRF de l'état de l'art (Jeong et Lee, 2008) ainsi que des modèles DNN (Mesnil et coll., 2015; Bapna et coll., 2017; Liu et Lane, 2016) ont été utilisés. Ces modèles abordent la tâche de NLU comme un *étiquetage de séquence*. Cela signifie que les données artificielles d'apprentissage doivent

être *alignées* pour associer chaque terme à une étiquette de concept comme dans le schéma d'étiquetage *IOB* : intérieur (*Inside*), extérieur (*Outside*), début (*Begin*). Le préfixe B devant une étiquette indique que l'étiquette est le début d'un concept et un préfixe I est utilisé pour une étiquette à l'intérieur d'un concept. Une balise O représente un terme en dehors d'un concept. Cette approche d'étiquetage est différente d'une *tâche de génération de séquence*. En utilisant une tâche de génération de séquence avec des données *non alignées*, le modèle devrait apprendre à associer plusieurs mots à une étiquette de concept sans données alignées. Par exemple, pour la phrase source « Allumer la lumière » un modèle séquence-à-séquence générerait la séquence cible `intention [set_device], action [TURN_ON], device [light]`, sans spécifier le concept associé à l'article défini. Pour la génération du corpus artificiel *aligné* et *non aligné*, nous avons adapté le générateur artificiel pour générer des phrases en deux formats : *alignés* pour l'apprentissage des modèles de NLU, Rasa-NLU, Tri-CRF et Att-RNN en utilisant le schéma d'étiquetage de concept *IOB* (Desot et coll., 2018) et deux formats *non alignés* ciblant la SLU séquentielle et E2E (Desot et coll., 2019a,b). L'exemple suivant est en format *json* (Rasa-NLU) pour l'énoncé « pouvez-vous fermer le store » :

```
   "vocadom tu peux fermer le store"
"intent": "set_device"
  {
  "start": 16,
  "end": 22,
  "entity": "action",
  "value": "close",
  "text": "fermer",
  },
  {
  "start": 23,
  "end": 31,
  "entity": "device",
  "value": "blind",
  "text": "le store",
  }
  ]
```

Le tableau 6.7 permet de faire, sur un exemple, une comparaison entre ces différentes approches. Sous l'onglet (Tri-CRF), ce tableau montre une phrase de la version du corpus utilisant le schéma d'étiquetage *IOB* de concept avec une fenêtre de deux mots précédant et suivant le mot cible associé à une étiquette de concept pour l'apprentissage du modèle Tri-CRF de (Jeong et Lee, 2008, 2009). Sous l'onglet (Att-RNN), le tableau 6.7 montre la version du corpus aligné avec le schéma d'étiquetage *IOB* de concept pour l'apprentissage du modèle Att-RNN (Liu et Lane, 2016). Il y a une association entre les mots de la transcription source

et les étiquettes cibles : l'article défini français 'le' est lié à l'étiquette du concept cible 'B-person'.

Les approches alignées sont moins pertinentes pour un modèle de SLU classique étant donné que les données d'entrée consistent en parole spontanée avec des disfluences. Celles-ci provoquent fréquemment des erreurs au niveau de la RAP, notamment des erreurs de *suppression* et *d'insertion*. Par conséquent, pour une approche de *génération de séquence*, une version du corpus artificiel a été générée sans alignement entre les phrases source d'une part et les étiquettes d'autre part. En utilisant une approche non alignée, les étiquettes de concept peuvent être déduites de transcriptions de RAP imparfaites car le modèle Seq2seq de NLU génère l'ensemble des étiquettes à partir d'un état abstrait de l'énoncé. Il n'est donc pas dépendant du contexte immédiat du terme d'entrée.

C'est pourquoi, le tableau 6.7 inclut l'exemple d'une phrase en format non aligné dont nous nous servirons pour entraîner un modèle de Seq2seq dans le chapitre suivant. Dans ce format, l'intention est incluse (entre crochets) dans la séquence de concepts comme premier élément. Nous avons supposé que l'intention en position initiale améliorera la prédiction des concepts suivants, car ceux-ci ont tendance à dépendre de l'intention. Ces concepts sont séparés des valeurs, qui sont elles entre crochets, afin que le modèle puisse les apprendre séparément (Mishakova et coll., 2019).

6.3.3 Transcriptions enrichies de symboles représentant les étiquettes de concepts et les classes d'intention

En prévision de l'apprentissage d'un modèle de bout-en-bout, nous avons également généré une version du corpus artificiel qui enrichit les transcriptions des énoncés avec des symboles de classes d'intentions et d'étiquettes de concept, automatiquement injectés dans les phrases du corpus (Desot et coll., 2019a,b). Une méthode similaire a été appliquée dans Ghannay et coll. (2018) (chapitre 4, section 4.2.2). Des transcriptions symboliquement enrichies d'étiquettes de concept ont été utilisées pour entraîner un modèle en utilisant le système de RAP *Baidu Deep Speech* (Hannun et coll., 2014). Notre approche est également inspirée de Serdyuk et coll. (2018) où les intentions ont été directement déduites des paramètres MFCC, en entraînant un modèle seq2seq sur des données vocales sans bruit de fond d'une part et bruitées d'autre part. Contrairement à ces deux approches, nos transcriptions sont enrichies de symboles d'étiquettes d'intention *ainsi que* de concept. Le tableau 6.8 comprend un aperçu d'étiquettes symboliques par classe d'intention et le tableau 6.9 comprend un aperçu d'étiquettes symboliques pour les concepts.

Un exemple d'une transcription enrichie est inclus dans le tableau 6.7 et figure 6.8.

6.3.4 Synthèse vocale

La méthode de génération du corpus artificiel présentée plus haut ne génère que des énoncés textuels. Or, pour effectuer la tâche de SLU il est nécessaire d'obtenir des énoncés acoustiques. La deuxième phase nécessaire de création de corpus est alors la génération

TABLE 6.7 – Corpus artificiel - présentation des formats aligné, non aligné, et des transcriptions enrichies de symboles

Aligné

Tri-CRF (“vocadom appelle médecin")

```
(Source) vocadom appelle médecin
(Cible)
O vocadom /                    -1=<s>     +1=appelle +2=médecin
action-B appelle / -2=<s>      -1=vocadom +1=médecin +2=</s>
person-B médecin / -2=vocadom -1=appelle +1=</s>
CONTACT
```

Att-RNN (“vocadom appelle le médecin")

```
(Source) vocadom appelle  le       médecin
(Cible)  O       B-action B-person I-person
CONTACT
```

Non aligné

Seq2seq (“vocadom ferme la porte")

```
(Source) vocadom ferme la porte
(Cible)  intent[set_device], action[close], device[door]
```

Transcription enrichie de symboles

E2E (ESPnet) (“vocadom allume la lumière")

```
(transcriptions + étiquettes cibles insérées)

@ VocADom  ^allume^ }la lumière} @

SET_DEVICE = @
Action = ^
Device = }
```

TABLE 6.8 – Étiquettes symboliques associées à chaque classe d’intention

Intention	Symbole
check_device	#
contact	[
get_room_property	{
get_world_property	]
set_device	@
set_device_property	_
set_room_property	&

d’une base d’énoncés de parole artificielle. À cette fin, nous avons utilisé une technique de synthèse vocale.

6.3.4.1 Intérêt de la synthèse vocale pour la compréhension et la reconnaissance automatique de la parole

La technique d’augmentation de données par synthèse vocale est très proche de Lugosch et coll. (2020) qui ont utilisé de la synthèse vocale pour générer des données de parole artificielle en anglais pour l’apprentissage d’un modèle de SLU E2E. Ce modèle a été généré à l’aide de *VoiceLoop* (Taigman et coll., 2018) de Facebook qui contient 22 voix synthétiques.

TABLE 6.9 – Étiquettes symboliques associées à chaque concept

Concept	Symbole
action	^
device	}
device-component	*
device-setting	,
location-floor	;
location-house	!
location-inroom	?
location-room	>
organization	§
person-name	+
person-occupation	=
person-relation	-
room-property	/
value-artist	.
value-numeric	%
value-qualitative	|
world-property	∘

TABLE 6.10 – Comparaison de performances SLU sur des données réelles (réel) et sur un mélange de données réelles et artificielles (réel + artif.) (Lugosch et coll., 2020)

Modèle	Précision (%)
réel	65.5
réel + artif.	71.4

Ils ont combiné les données synthétiques qui en résultent avec des données d'entraînement de parole réelle, en utilisant l'ensemble de données English Fluent Speech Commands (Lugosch et coll., 2019). Les performances de leur approche SLU ont été comparées pour un modèle combinant des données artificielles et réelles d'une part, et un modèle avec uniquement des données réalistes d'autre part. Les résultats du tableau 6.10 montrent les meilleures performances de leur modèle avec les données augmentées.

(Li et coll., 2018) ont présenté un modèle de RAP de bout en bout qui est entraîné avec des données de synthèse. Les auteurs rapportent que leur modèle obtient des performances de RAP optimales en utilisant 50 % de données de synthèse et 50 % de parole réelle dans le corpus d'apprentissage.

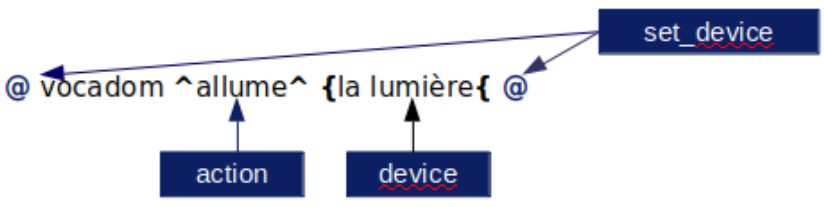


FIGURE 6.8 – Transcription enrichie de symboles de concept et d'intention

Ce court état de l'art montre que la synthèse vocale est une approche intéressante pour combler le manque de données réelles. Nous avons donc utilisé la voix de synthèse vocale Open Source Ubuntu *SVOX*[1] française féminine[2] pour générer de la parole artificielle.

Ce moteur de synthèse prend un énoncé textuel en entrée et génère en sortie un fichier au format MS-WAV échantillonné à 16 kHz. Son style conversationnel rend cette voix intéressante car le style de la parole des locuteurs de notre ensemble de test est également *conversationnel*. Cependant, sa qualité prosodique est médiocre et des erreurs de prononciation fréquentes se produisent entre autres pour les noms propres des mots-clés. Nous avons comparé cette voix à la voix de synthèse vocale Open Source Google (*gTTS*) pour le français[3]. Pour chaque phrase d'entrée, un fichiers au format MP3 à 24 kHz est produit, fichier que nous avons ensuite converti au format MS-WAV à 16 kHz. Bien que sa prosodie et sa qualité phonétique soient meilleures par rapport à la voix française de synthèse vocale SVOX, son style de parole n'est pas conversationnel et il s'agit plutôt d'un style de *lecture*. En comparant les deux voix en tant que données d'apprentissage, la voix SVOX conversationnelle s'est avérée plus adaptée à notre approche SLU. Ces expérimentations et évaluations sont décrites en détail dans le chapitre 8.

6.3.4.2 Évaluation de la qualité de la synthèse vocale

Pour mesurer la pertinence de l'approche par synthèse vocale, nous avons comparé la parole artificielle avec les données de parole réelle VocADom@A4H, en calculant les distances acoustiques entre les deux ensembles de données. Nous avons ainsi généré de la parole artificielle à partir des 6747 énoncés de l'ensemble de test VocaDom@A4H et calculé la distance acoustique entre les énoncés de la parole *réelle* et les énoncés de la parole *artificielle* en appliquant la technique de *déformation temporelle dynamique* (DTW) que nous avons décrit dans le chapitre 5, section 5.5.3.

La figure 6.9 illustre cette distance d'édition minimale pour un échantillon de la parole réelle (VocADom@A4H) avec 20 MFCC sur 168 trames, et un échantillon de synthèse vocale avec 20 MFCC sur 179 trames pour la commande « chanticou arrêtez les stores de la salle de bains ». Les figures 6.10 et 6.11 montrent respectivement les signaux et spectrogrammes des deux échantillons de parole réelle et de synthèse vocale. La ligne bleue de la figure 6.9 montre le chemin de déformation qui minimise la distance entre le signal vocal artificiel et le signal vocal réel. Les régions plus foncées symbolisent un coût et une distance plus élevées.

Le tableau 6.11 montre les résultats du calcul de DTW entre les échantillons de synthèse vocale (féminine) et réels (*Synthèse vocale ↔ réel*) d'une part, et les résultats au niveau inter-locuteur pour les échantillons de parole réelle pour des phrases identiques entre les locuteurs d'autre part (*Inter-locuteur-réel*). Les distances ont été normalisées de trois manières différentes. Nous avons divisé la distance totale par la longueur de la séquence la plus longue (*Long norm.*), par la longueur la plus courte (*Court norm.*) et par la longueur du chemin de

1. https://launchpad.net/ubuntu/+source/svox
2. https://doc.ubuntu-fr.org/svoxpico
3. https://pypi.org/project/gTTS

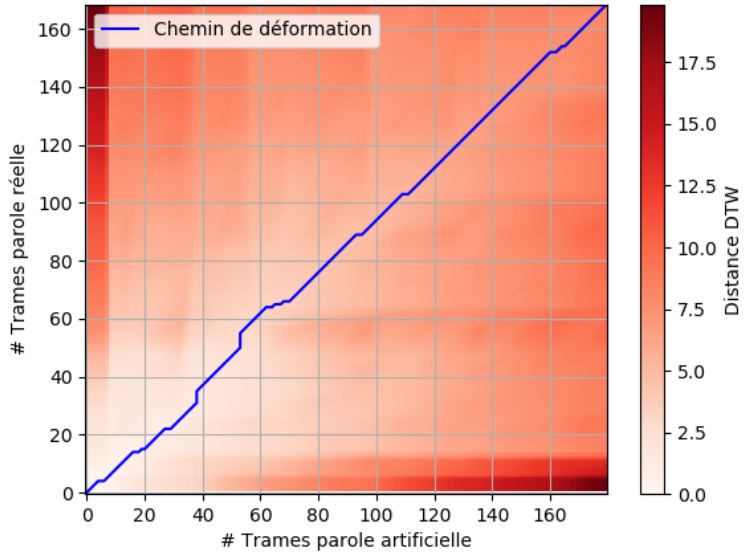


FIGURE 6.9 – Distance DTW entre VocADomA4H et VocADomArtif pour la phrase « chanticou arrêtez les stores de la salle de bains »

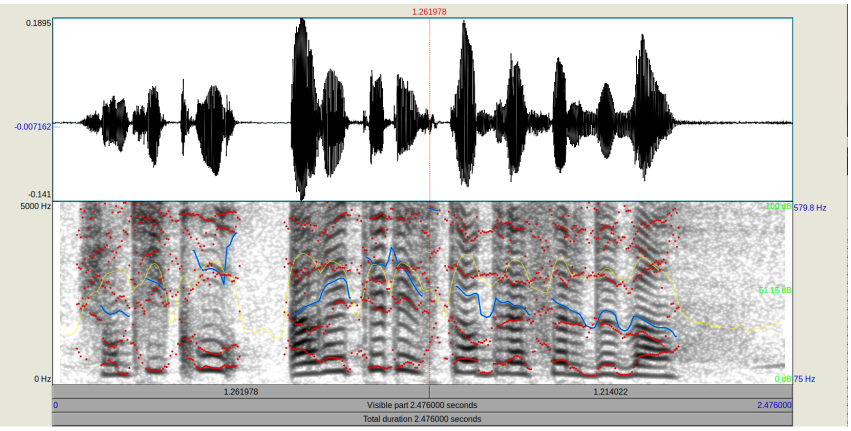


FIGURE 6.10 – Échantillon de parole naturelle « chanticou arrêtez les stores de la salle de bains »

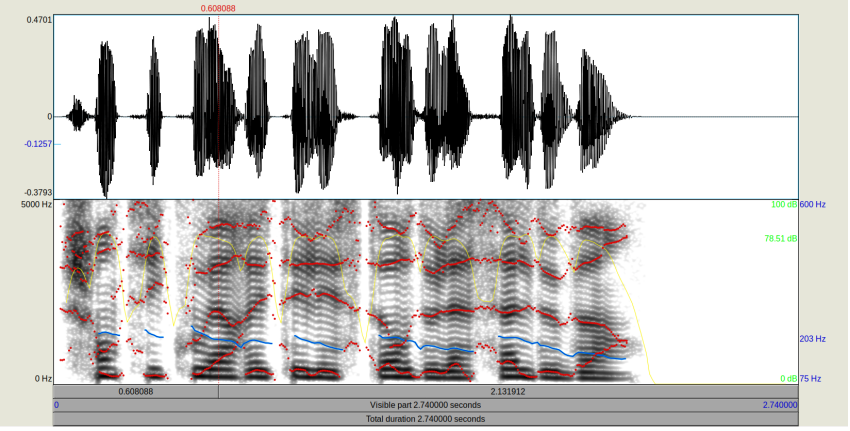


FIGURE 6.11 – Échantillon de parole synthétique « chanticou arrêtez les stores de la salle de bains »

déformation optimal (*Opt. norm.*) (tableau 6.11) (Ratanamahatana et Keogh, 2004). Le tableau 6.11 présente la distance moyenne et l'écart type pour tous les échantillons comparés.

TABLE 6.11 – DTW entre données de parole réelle VocADom@4H et de synthèse VocADom@Artif

DTW	**Long norm.**	**Court norm.**	**Opt norm.**
Synthèse vocale ↔ réel :			
tous (6747)	$5.58^{\pm 4.42}$	$7.20^{\pm 8.32}$	$4.71^{\pm 3.61}$
masculin (4372)	$5.72^{\pm 4.02}$	$7.31^{\pm 7.86}$	$4.83^{\pm 3.28}$
féminin (2375)	$\mathbf{5.32^{\pm 5.04}}$	$\mathbf{7.01^{\pm 9.07}}$	$\mathbf{4.51^{\pm 4.12}}$
Inter-locuteur-réel :			
en commun (2806)	$1.85^{\pm 2.22}$	$4.16^{\pm 9.05}$	$1.67^{\pm 2.07}$

La distance entre les 6747 échantillons de synthèse vocale et leurs échantillons de parole réelle équivalents (*tous*) est présentée sur le tableau 6.11. Le tableau comprend également une comparaison entre tous les échantillons masculins de parole réelle et le nombre équivalent (mis entre parenthèses) d'échantillons de synthèse vocale (*masculin*). Finalement, la troisième ligne du tableau comprend une comparaison entre tous les échantillons féminins de parole réelle et les échantillons de synthèse vocale équivalents (*féminin*). Les distances sont calculées uniquement sur des énoncés identiques.

Le tableau 6.11 montre qu'en général les distances inter-locuteurs de parole réelle sont significativement plus faibles que dans le cas où la comparaison est faite entre de la synthèse vocale et de la parole réelle. Comme la parole artificielle est générée pour une voix française *féminine*, les distances entre les échantillons de la synthèse vocale féminine et les échantillons réels féminins sont plus faibles que les distances entre les échantillons de la parole réelle masculine et la synthèse vocale (féminine).

Ces mesures montrent que, bien que la voix issue d'une synthèse vocale soit bien plus éloignée des participants que lorsqu'ils sont comparés entre eux, cette distance reste faible. Par ailleurs, une partie de la mesure acoustique est influencée par les conditions d'enregistrement qui sont les mêmes entre les participants alors que la synthèse vocale ne contient aucun bruit de fond. Enfin, nous verrons plus loin que la synthèse vocale a bien joué son rôle de données d'apprentissage complémentaires.

6.4 Sélection des données d'apprentissage sans intention

Le corpus artificiel ne contient que des commandes vocales et par conséquent seulement des intentions. Étant donné que 75% des données de test VocADom@A4H ne sont en fait pas des commandes vocales, il devient necessaire d'ajouter des énoncés *sans* intention (intention `none`) dans le corpus d'apprentissage. Nous avons utilisé pour cela des énoncés du corpus ESLO2 de discours conversationnel en français (126h) (Serpollet et coll., 2007) pour modéliser l'intention `none`. De la même facon que pour les corpus VocADom@A4H et SWEET-HOME, il contient des disfluences fréquentes. Étant donné que les données ESLO2

contiennent un discours conversationnel quotidien, certains de ces énoncés contenaient en fait une intention (par exemple, "vous devriez ouvrir la porte"). Par conséquent, ces énoncés ont été filtrés et mis de côté. En utilisant un modèle n-gramme appris sur le corpus artificiel, des phrases avec un vocabulaire spécifique au domaine ont été sélectionnées. Ces énoncés ont ensuite été filtrés manuellement et seulement les énoncés hors domaine ont été gardés afin de collecter des énoncés sans commande vocale, ou avec classe d'intention `none` (figure 6.7 et tableau 6.6).

6.5 Conclusion

Dans ce chapitre nous avons décrit l'enregistrement du corpus réaliste VocADom@A4H que nous avons utilisé comme ensemble de test, et la génération d'un corpus artificiel VocADom@ARTIF dont les phrases sont automatiquement générées et étiquetées de concepts et d'intentions. Sur la base de ce corpus artificiel, nous avons utilisé une technique d'augmentation de données par synthèse vocale en utilisant une voix de synthèse vocale française féminine *Open Source*. Nous avons pu mesurer qu'il existe une distance entre la parole réelle et artificielle. Il est cependant difficile de dire quel impact aura cette distance lorsqu'un modèle de SLU aura été appris sur ces données artificielles. Les expérimentations que nous présenterons au cours des chapitres 7 et 8 montreront l'influence de cette distance au niveaux *acoustique* et *syntaxique*. Nous verrons qu'il est possible de compenser ce décalage grâce à l'utilisation de données réelles complémentaires que nous avons utilisé pour enrichir les données d'entraînement avec des énoncés *sans* intention. **La création et génération de ces deux corpus nous permettront d'entraîner et d'évaluer des modèles profonds pour effectuer de la tâche de SLU**.

CHAPITRE 7

Approche séquentielle de la compréhension de la parole

La question de recherche principale de cette thèse est de vérifier **quels avantages une approche SLU de bout-en-bout (E2E) peut offrir par rapport à une approche en pipeline classique**.

Une approche pipeline classique est constituée d'un pipeline avec un module de RAP dont les hypothèses de transcription de sorties alimentent un module de NLU qui extrait les concepts des transcriptions. C'est ce que nous appelons par la suite l'*approche séquentielle*. Cette approche reste aujourd'hui standard dans les systèmes industriels performants. Cette séparation en module permet une optimisation indépendante des modules sur des données qui n'ont pas besoin d'être synchronisées entre les deux tâches. Par ailleurs, une telle architecture est plus facile à maintenir d'un point de vue industriel. Cette architecture reste cependant très sensible aux cascades d'erreurs. Dans ce chapitre, nous décrivons l'approche séquentielle de SLU que nous avons conçue et qui constituera notre système de référence. Les corpus VocADom@ARTIF et VocADom@A4H décrits au cours du chapitre 6 serviront respectivement comme corpus d'apprentissage et d'évaluation. Le chapitre 3 sur l'état de l'art de la SLU séquentielle présente les techniques utilisées pour réduire les décalages entre les modules de RAP et de NLU. En utilisant des données d'entraînement *non-alignées* nous essaierons de diminuer l'effet cascade d'erreurs causé par les imperfections du composant RAP qui impactent la NLU. Nous positionnerons cette approche par rapport à des approches de NLU *alignées*.

Dans les sections suivantes, nous décrivons le module de RAP que nous avons construit en utilisant la boite à outils Kaldi, et comment nous avons appris les modèles acoustiques, le lexique et le modèle de langage de ce moteur de RAP HMM-DNN classique. Nous obtenons des performances de RAP à l'état de l'art sur les données considérées dans cette thèse. Puis nous décrivons les modules de NLU en mode *aligné* et *non aligné* et comment nous les avons optimisés. Au cours de ce chapitre, mais aussi dans les chapitres suivants, nous nous référerons régulièrement à figure A.1 de l'annexe A qui schématise l'utilisation de nos ensembles d'apprentissages utilisés.

7.1 Système de RAP KALDI

Le module de RAP de l'approche SLU séquentielle de référence repose sur le système hybride HMM-DNN Kaldi que nous avons décrit dans le chapitre 3 section 3.1.3. Nous avons

TABLE 7.1 – Corpus utilisés pour l'apprentissage et la validation de notre système de RAP basé sur Kaldi

Corpus	**# heures**
ESTER1	100
ESTER2	100
REPERE	60
ETAPE	30
BREF120	51.50
AD	0.5
SWEET-HOME	2.5
CIRDO	2
ESLO2	126
Total	472.65

utilisé la version Kaldi 5.5 [1]. Nous nous sommes inspirés de l'approche des études de Elloumi et coll. (2018) et de Elloumi (2019). Pour entraîner nos modèles acoustiques, nous avons utilisé 472.65 heures de données de parole qui se composent des corpus français ESTER1 (Galliano et coll., 2005) et 2 (Galliano et coll., 2009), REPERE (Giraudel et coll., 2012), ETAPE (Gravier et coll., 2012), BREF120 (Tan et Besacier, 2006), AD (Vacher et coll., 2008), SWEET-HOME (Vacher et coll., 2014), CIRDO (Vacher et coll., 2016) (chapitre 2) et le corpus de parole spontanée ESLO2 (Serpollet et coll., 2007) (tableau 7.1). Par la suite nous comparerons les performances *avec* et *sans* le corpus ESLO2 dans l'ensemble d'apprentissage. Tout comme VocADom@A4H, ESLO2 est un corpus de parole spontanée qui contient beaucoup de disfluences. L'ensemble de ces données à été séparé en 90% de données d'entraînement et 10% de données de développement. Les tests ont été effectués sur VocADom@A4H.

7.1.1 Méthode utilisée

7.1.1.1 Processus de pré-traitement

Pour améliorer la tâche d'apprentissage, les données textuelles ont été normalisées pour unifier les formes et réduire le vocabulaire. Ces données ont également été utilisées pour créer le Modèle de Langage (ML) et le Modèle Acoustique (MA). Ce pré-traitement consiste en :

1. convertir tous les caractères en minuscule,
2. normaliser et convertir les chiffres en lettres,
3. convertir les unités de mesures,
4. les symboles en lettres,
5. transformer les abréviations en mots,
6. segmenter en unités de référence (*tokenisation*) et,
7. supprimer les ponctuations.

1. https://github.com/kaldi-asr/kaldi

Certaines séquences de mots se prononçant comme une expression figée, nous les avons considérées comme une entité unique en réunissant chacun des termes par l'insertion d'un tiret bas. Par exemple, `"parce que"` est converti en `"parce_que"`.

7.1.1.2 Modèles acoustiques

Les MA ont été entraînés à l'aide des scripts faisant partie de l'outil KALDI (Povey et coll., 2011b), en suivant la configuration standard. Nous avons utilisé des paramètres MFCC à 13 dimensions, leurs dérivées premières Δ, leurs dérivées secondes $\Delta\Delta$ et l'énergie, avec une analyse discriminante linéaire (*Linear Discriminant Analysis* – LDA) et une transformation linéaire à maximum de vraissemblance (*maximum likelihood linear transformation* – MLLT) appliquée sur une fenêtre de trames de largeur 7 (3 contextes gauches et 3 droits) et projetées dans un espace de 40 dimensions.

Pour l'apprentissage d'un MA HMM-DNN, il faut d'abord créer un modèle HMM-GMM pour générer les alignements au niveau des données acoustiques. Les états des phonèmes sont divisés en sous-groupes et chacun de ces sous-groupes représente une partie de la probabilité totale de l'état du phonème modélisé. D'une telle façon, un triphone prend en compte un contexte de trois phones, un précédent, un courant et un suivant.

Dans une première étape nous avons effectué un entraînement d'un MA mono-phone (tableau 7.4, *Mono-phone*), suivi par des modèles tri-phone (*Tri-phone*, *Tri-phone* delta delta). Ensuite une analyse discriminante linéaire (LDA) et une transformation linéaire à vraisemblance maximale (MLLT) (*Tri-phone LDA + MLLT*) ont été appliquées comme méthodes de transformations *indépendantes* du locuteur (*speaker independent transformation*). Finalement un modèle *dépendant* du locuteur a été appris en appliquant une régression linéaire à maximum de vraisemblance fMLLR (*Feature space Maximum Likelihood Linear Regression*) aux paramètres acoustiques (*GMM LDA + MLLT + fMLLR + SAT*). Les paramètres du modèle cible HMM-DNN (*DNN*) sont donnés dans le tableau 7.2

TABLE 7.2 – Paramétrage de notre système basé sur Kaldi

Paramètre	**Description**
Architecture	4 couches cachées de taille 1024 couche de sortie de 4748 unités Softmax
Optimiseur	SGD taux d'apprentissage qui varie entre 0.01 et 0.001
Taille de lot	128
Paramètres acoustiques	MFCC à 13 dimensions
ML	basé sur mots interpolation ML spécifique au domaine (poids = 0,6) avec ML générique (poids = 0,4)
Apprentissage	15 époques durée totale 253 heures (10.5 jours) 1 GPU, GPUGeForce GTX TITAN Black

Comme le tableau 7.2 le montre, la couche *softmax* produit un plus grand nombre de probabilités postérieures que nécessaire, comme ce nombre devrait dépasser (2 fois) le nombre

TABLE 7.3 – Description des données monolingues utilisés pour construire les modèles de langage

Corpus	Lignes	# mots	Mots unique
EUbookshop	18M	432M	1,71M
TED2013+wit3	0,16M	2M	0,06M
GlobalVoices	0,37M	7M	0,18M
Giga	18M	57M	1,23M
Europarl-v7	2,24M	60M	0,13M
MultiUN	13M	404M	0,41M
OpenSubtitles2016	90M	534M	0,87M
DGT	3.1M	61,7M	0,28M
News-Commentary11+News-WMT	30M	661M	1,43M
Lemonde	13M	368M	1,12M
Trames	0,21M	0,79M	0,03M
Wikipedia	20M	502M	2M
Total	208.08M	3089.49M	5.14M

de feuilles de l'arbre de décision.

7.1.1.3 Modèle de langage

Nous avons construit deux modèles de langage 3-grammes en appliquant un lissage Kneser-Ney en utilisant l'outil *SRILM*[2] (Stolcke, 2002). Ces deux modèles, le modèle générique et le modèle adapté à notre application, ont ensuite été utilisés pour obtenir le modèle de langage de notre système.

Le premier modèle (ML générique) a été appris sur une grande quantité de données monolingues de plusieurs corpus français, EUbookshop (Lison et Tiedemann, 2016), TED2013 [3], wit3 (Cettolo et coll., 2012), GlobalVoices [4], Gigaword [5], Europarl-v7 (Koehn, 2005), MultiUN (Ziemski et coll., 2016), OpenSubtitles2016 [6], DGT [7], News-Commentary [8], News-WMT [9], Lemonde [10], Wikipedia [11]. Le tableau 7.3 présente les principales caractéristiques des corpus utilisés pour l'apprentissage de ce modèle.

Le second modèle (ML spécifique) a été appris avec les phrases de VocADom@ARTIF, spécifique au domaine et celles du corpus SWEET-HOME.

Ces deux modèles sont ensuite interpolés avec un poids plus fort donné aux données spécifiques au domaine. Le modèle final résulte d'une interpolation entre le ML spécifique au domaine (poids = 0,6) et le le ML générique (poids = 0,4). Le modèle obtenu étant volumineux, nous avons effectué une opération de filtrage sur le modèle interpolé en ne gardant

2. http://www.speech.sri.com/projects/srilm/download.html
3. http://opus.nlpl.eu/TED2013.php
4. http://casmacat.eu/corpus/global-voices.html
5. https://catalog.ldc.upenn.edu/LDC2011T10
6. http://opus.nlpl.eu/OpenSubtitles-v2018.php
7. http://opus.nlpl.eu/DGT.php
8. http://opus.nlpl.eu/News-Commentary.php
9. http://opus.nlpl.eu/WMT-News.php
10. http://catalog.elra.info/en-us/repository/browse/ELRA-W0015/
11. https://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/

que les n-grammes ayant une probabilité supérieure à 10^{-9}.

7.1.1.4 Dictionnaire de phonétisation

Notre dictionnaire de prononciation a été construit à l'aide de la ressource lexicale *BDLEX* (Perennou, 1986) et de l'outil de conversion automatique de graphèmes-à-phonèmes *LIA_Phon*[12] afin de trouver les variantes de prononciation des mots. Il a été produit en 3 étapes :

1. **Sélection du vocabulaire** : Nous avons sélectionné le vocabulaire en extrayant les 80k mots les plus probables du ML, et nous avons ajouté la liste des mots composés proposée dans la ressource *BDLEX* (de-la, afin-de,etc.).
2. **Phonétisation BDLEX** : Attribuer à chaque mot de notre vocabulaire la liste des phonétisations possibles proposée par *BDLEX*. Étant donné que notre corpus est hétérogène, nous avons ajouté des variantes de prononciation.
3. **Phonétisation automatique** : Si un mot de notre corpus n'existe pas dans *BDLEX*, nous utilisons l'outil *LIA_PHON* pour obtenir automatiquement les prononciations correspondantes du mot. Étant donné que *LIA_PHON* est un outil automatique pouvant produire des erreurs, nous avons vérifié manuellement les prononciations des 1000 mots les plus fréquents.

7.1.2 Évaluation et résultats

Le système a été évalué sur les 6747 énoncés du corpus VocADom@A4H. Les performances des modèles acoustiques ont été évaluées à l'aide du taux d'erreurs de mot (WER) (chapitre 5, section 5.5.1). Le tableau 7.4 présente les résultats de RAP dans différentes conditions de modélisation acoustique. On peut voir que les performances du modèle DNN (tableau 7.4, *DNN*) surpassent les performances de tous les autres modèles. Par ailleurs, l'effet du corpus ESLO2 est très variable en fonction des modèles. On peut cependant constater que le modèle DNN peut en tirer parti pour améliorer son score global. Les transcriptions des hypothèses de sortie de KALDI, seront les entrées de test pour le module de NLU. Nous décrivons son développement dans la section suivante.

7.2 Système de compréhension de langage naturel (NLU)

Nous avions terminé le chapitre 5 en concluant que le problème principal des systèmes de SLU séquentielle de l'état de l'art est leur dépendance des transcriptions sorties du module de RAP. Dans cette section nous proposons une manière de rendre la NLU moins sensible au taux d'erreurs d'un module de RAP en considérant un module de NLU appris sur des transcriptions *non-alignées*. Un modèle appris sur des transcriptions non-alignées devient

12. http://lia.univ-avignon.fr/chercheurs/bechet/download/lia_phon.v1,2.jul06.tar.gz

TABLE 7.4 – Performances de notre système de RAP basé sur Kaldi. Tests sur le corpus VocADom@A4H

Modèle Acoustique	WER(%)	
	sans ESLO2	avec ESLO2
Mono-phone	47.27	49.22
Tri-phone	35.56	36.54
Tri-phone + Δ + ΔΔ	36.03	36.70
Tri-phone LDA + MLLT	35.44	36.33
GMM LDA + MLLT + fMLLR + SAT	28.08	27.99
DNN	23.25	**22.92**

plus flexible et peut mieux traiter des transcriptions d'hypothèse de RAP contenant des mots supprimés, substitués ou ajoutés. Pour *valider* cette approche, nous avons d'abord évalué la qualité de l'ensemble d'apprentissagge VocADom@ARTIF en effectuant plusieurs approches de NLU *alignées*.

Comme nous l'avons expliqué au chapitre 6, section 6.3.2, les modèles alignés abordent la tâche de NLU comme un *étiquetage de séquence* où chaque terme des données d'apprentissage est associé à une étiquette de concept. Comme étape *préparatoire*, ciblant la validation de notre approche de NLU, nous avons entraîné d'abord trois modèles alignés sur le corpus Port-Media (Lefèvre et coll., 2012) qui est annoté au niveau sémantique. La comparaison entre les performances des modèles et entre les ensembles de données doit nous permettre de déterminer si les limites de performance sont dues aux limitations des modèles sur cette tâche ou aux limitations du jeu de corpus artificiel.

7.2.1 Approches alignées

Nous examinons les performances des modèles NLU à l'état de l'art, entraînés sur le corpus artificiel et testés sur le corpus réaliste VocADom@A4H. Nous avons ainsi construit deux modèles NLU correspondant aux méthodes de Tri-CRF (Jeong et Lee, 2008), et d'Att-RNN (Liu et Lane, 2016) ; notre approche de référence est obtenue à partir de l'outil commercial *Open Source* Rasa (Braun et coll., 2017) que nous avons présenté au chapitre 3, respectivement en sections 3.2.4, 3.2.6 et 3.2.5. Le modèle RASA contient des modèles *séparés* pour prédire l'intention et les concepts. Les concepts sont associés à des morceaux de texte, en utilisant une approche de CRF, alors que l'intention est associée à l'ensemble de la phrase, en utilisant une approche de classification par SVM. Par contre les modèles Tri-CRF et Att-RNN appliquent un apprentissage multi-tâche. Ils peuvent prédire une intention associée à l'ensemble de la séquence d'entrée et une séquence de concepts *simultanément*.

7.2.1.1 Paramétrisation des outils de NLU

Nous avons utilisé ces trois outils de NLU de l'état de l'art en utilisant les paramètres suivants :

— **Tri-CRF** : Afin de réduire le temps d'apprentissage, nous avons élagué les intentions à

TABLE 7.5 – Less corpus VocADom@ARTIF, VocADom@A4H et Port-Media utilisé pour l'évaluation du système de NLU

Ensemble de données	Énoncés	Mots	Intent.	Concept	Concept valeurs
VocADom @Artif 28k	28000	156	7	16	60
VocADom @Artif 42k	42195	157	7	16	60
VocADom @Artif (complet)	77481	187	7	17	69
VocADom@A4H.	6747	1462	8	14	60
Port-Media	18026	3062	4	32	378

faible probabilité (< 0.1%) et initialisés les poids en utilisant la pseudo-vraisemblance (pseudo *likelihood*) pour 30 itérations d'apprentissage. L'apprentissage comportait 200 itérations.

— **Att-RNN** : Dans notre implémentation de l'outil Att-RNN, l'encodeur et le décodeur d'un BLSTM sont des couches de 128 unités. L'optimisation est faite par descente de gradient stochastique (SGD – *Stochastic Gradient Descent*) avec une taille de lot (*batch size*) de 16, un écrêtage de gradient (*gradient clipping*) avec une norme de 0,5, une régularisation par abandon (*dropout*) avec une probabilité de maintien de 0,5. L'apprentissage pouvait se poursuivre jusqu'à 10000 étapes (*steps*).

— **RASA** : L'outil utilise un modèle CRF pour prédire les étiquettes de concepts et une table de correspondance pour déterminer leurs valeurs. Le modèle utilise séparément un SVM basé sur une représentation vectorielle des mots (*Word Embeddings*) préappris en appliquant word2vec sur des données Wikipedia, OpenSubtitles et Wikinews. Le vocabulaire final contient 1 184 651 mots et les vecteurs ont une longueur de 300 unités.

7.2.1.2 Étape préparatoire de validation sur le corpus Port-Media

Pour valider l'implantation des modèles NLU *alignées*, nous utilisons l'ensemble de données Port-Media (Lefèvre et coll., 2012) contenant des informations touristiques et de réservations de billets en français pour le festival de musique d'Avignon de 2010. L'ensemble de données contient des énoncés naturels de 140 locuteurs dans une tâche de réservation téléphonique simulée. Il contient des annotations de concept et d'étiquettes de valeur. Dans ce corpus, les intentions sont également des concepts, associés à des sous-ensembles d'énoncés - nous extrapolons simplement ces attributs comme des intentions au niveau de l'énoncé. Une comparaison entre le corpus Port-Media, les différentes versions du corpus VocADom@ARTIF et VocADom@A4H est fournie dans le tableau 7.5. Comme indiqué, Port-Media est riche en termes d'étiquettes et de valeurs de concept et d'une taille comparable à notre première version de l'ensemble de données artificielles (*Artif. 28k*[13]). Il est donc adéquat de comparer les performances du modèle de NLU avec les performances des ensembles de données obtenues en habitat intelligent.

13. Première version du corpus artificiel, VocADom@ARTIF, de 28K phrases

TABLE 7.6 – Performances de NLU aligné (%) sur le corpus Port-Media

Modèle NLU	**Intention**			**Concept**		
	Précision	**Rappel**	**F-Mesure**	**Précision**	**Rappel**	**F-Mesure**
Rasa-NLU	92.20	92.52	92.26	95.17	94.22	94.16
Tri-CRF	96.42	96.43	96.36	95.31	95.74	95.39
Att-RNN	**97.56**	**97.56**	**97.56**	**95.96**	**96.36**	**96.11**

Le tableau 7.6 montre les résultats pour Port-Media, ce jeu étant partitionné entre un ensemble d'apprentissage (90%) et de développement (10%). Les performances pour la prédiction des concepts et les intentions, affichées concernent seulement le jeu de développement. Att-RNN atteint sans surprise les meilleures performances pour les 2 tâches. Rasa est moins performant que les 2 autres méthodes. Ces résultats sur Port-Media montrent le niveau de performance qui peut être atteint avec les 3 modèles NLU à l'état de l'art sur des tâches de complexité similaire à celles rencontrées en habitat intelligent [14].

En ce qui concerne le domaine de l'habitat intelligent, le jeu de données d'apprentissage artificielles (tableau 7.5, *Artif. 28k* et annexe A, figure A.1) a été réparti aléatoirement entre apprentissage (90%) et développement (10%). Nous avons normalisé tous les mots-clés (vocadom, minouche, bérénio etc.) des ensembles d'apprentissage artificiel et de test VocADom@A4H, en les remplaçant par "KEYWORD". Enfin, le corpus de test était constitué de 2612 énoncés de VocADom@A4H. Ce corpus contenant des énoncés sans intentions domotique (p.ex., "d'accord"), ceux-ci ont été exclus ce qui explique que seuls 2612 énoncés ont été retenus.

Les résultats obtenus sur les 2612 phrases *avec* intention (tableau 6.4), du corpus réaliste VocADom@A4H sont affichés sur le tableau 7.7 et la figure A.1 (*RASA-NLU(1), Tri-CRF(2), Att-RNN1(3)*). Les performances des modèles *Att-RNN (3)* et *Att-RNN (4)* dans le tableau 7.7, surpassent celles de *RASA-NLU (1)* et de *Tri-CRF (2)*. Les meilleures performances de *Att-RNN2 (4)* par rapport à Att-RNN1 (3) montrent l'impact d'un ensemble d'entraînement d'une plus grande taille. Ces résultats sont aussi une première indication qu'on peut utiliser des données d'apprentissage NLU artificielles et de données de test réalistes, malgré la distance linguistique entre ces deux ensembles de données.

TABLE 7.7 – Performances des systèmes RASA, Tri-CRF, Att-RNN NLU aligné et Seq2seq NLU non-aligné (%) sur les données VocADom@A4H

Modèles NLU	**Intention**			**Concept**		
	Précision	**Rappel**	**F-Mesure**	**Précision**	**Rappel**	**F-Mesure**
Artif. 28k :						
RASA-NLU(1)	90.48	71.39	76.57	**85.72**	**73.54**	**79.03**
Tri-CRF(2)	84.11	79.47	76.36	77.28	52.65	60.64
Att-RNN1(3)	**93.77**	**90.28**	**91.30**	69.19	66.24	66.09
Artif. 42k :						
Att-RNN2(4)	96.81	96.63	96.70	77.32	73.67	74.27
Seq2seq1(5)	95.37	94.59	94.74	48.95	55.27	51.06

14. Il convient de noter que ces experiences ont été initiées en 2018 et que le modèle Rasa a évolué depuis.

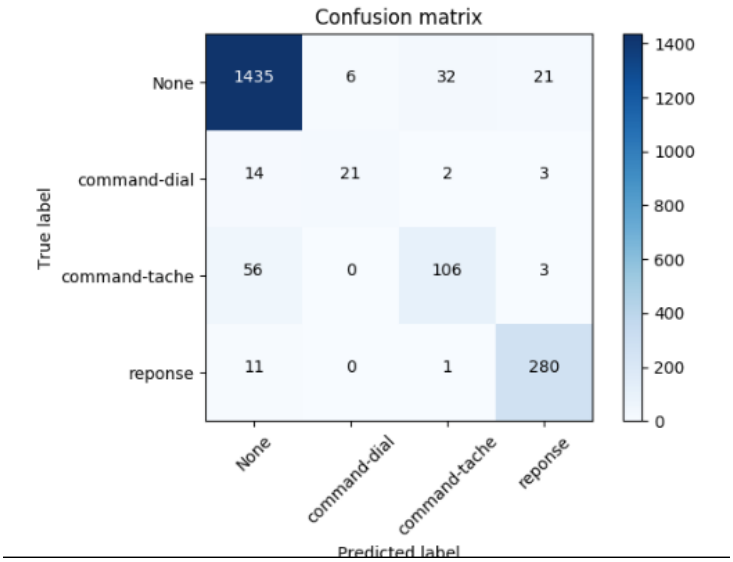


FIGURE 7.1 – Matrice de confusion entre l'intention *none* et les autres intentions dans le corpus Port-Media

7.2.1.3 Bilan de l'approche de NLU alignée

Pour évaluer les prédictions des concepts et des intentions, nous avons utilisé les mesures standards de précision, rappel et F-mesure présentées au chapitre 5 en section 5.5.2. Les performances sur le corpus réaliste VocADom@A4H sont moins bonnes que sur Port-Media et particulièrement pour la prédiction des étiquettes de concept. Cependant, nous devons considérer la forte précision de prédiction des intentions sur Port-Media comme biaisée par la présence d'une forte proportion d'intentions `none` comme le montre la matrice de confusion de la figure 7.1. Le corpus artificiel VocADom@ARTIF ne contient pas de phrases sans intention. Remarquons aussi que Port-Media contient seulement 4 classes d'intention alors que le corpus VocADom@A4H dans sa totalité en contient 7, la classe `none` intention n'étant pas prise en compte, ce qui rend la prédiction plus difficile.

Les résultats de reconnaissance de concepts sur VocADom@A4H sont particulièrement insatisfaisants. La raison la plus probable est qu'il contient des variations significatives de vocabulaire et de syntaxe par rapport au corpus artificiel. Les répétitions, les disfluences et les interjections (ex. "euh") conduisent à des énoncés syntaxiquement différents de ceux du corpus artificiel. C'est le modèle Att-RNN qui a montré la meilleure performance sur les intentions, mais sur les concepts, c'est RASA qui a été le plus performant. Cela peut être dû au fait que contrairement à Tri-CRF et à Att-RNN, RASA utilise une représentation vectorielle des mots (*Word Embeddings*) pré-entraînés sur des données externes ce qui lui permet de prendre en compte les mots hors vocabulaire. Par ailleurs, la taille de vocabulaire des intentions du corpus réaliste VocADom@A4H complet (430 mots) est plus du double de celui du corpus artificiel complet (187 mots). La perplexité 3-gram du ML artificiel calculée sur le corpus VocADom@A4H est 124, ce qui est assez élevé pour un vocabulaire aussi restreint. Le nombre de mots OOV est important, avec 307 mots absents du corpus artificiel (tableau 6.6).

7.2.2 Approche non-alignée

Bien qu'il s'avère que le modèle de NLU Att-RNN soit le plus efficace dans un contexte de NLU isolé, ce type d'alignement ne peut pas être pris en compte pour une approche de SLU séquentielle. Dans un contexte de SLU, les données d'entrée consistent en une parole spontanée avec des disfluences qui provoquent souvent des erreurs de suppression et d'insertion par la RAP. Déjà au niveau de la NLU, il y a un écart considérable entre les performances sur l'ensemble de test VocADom@A4H, et sur l'ensemble de développement qui est beaucoup plus homogène au niveaux lexical et syntaxique, par rapport à l'ensemble d'apprentissage. Les énoncés du corpus VocADom@A4H contiennent des disfluences fréquentes. Par conséquent, il faut un module de NLU plus flexible et robuste pour traiter des énoncés de test contenant des disfluences et une syntaxe divergente. L'utilisation de données non alignées contraint l'apprentissage à induire un modèle suffisamment flexible pour déduire des concepts à partir de transcriptions imparfaites.

7.2.2.1 L'outil de NLU seq2seq

Le module de NLU que nous avons utilisé pour l'approche de NLU non alignée est un modèle de séquence à séquence (seq2seq). Le problème de compréhension est vu comme un problème de traduction où l'entrée doit être abstraite pour guider une génération complète de la sortie. Nous avons utilisé un RNN du type encodeur-décodeur (Britz et coll., 2017) avec attention. Nous sommes parti des travaux que nous avions entrepris (Mishakova et coll., 2019) dans lesquels nous avions appris un modèle seq2seq avec attention sur les annotations non-alignées de VocADom@ARTIF (tableau 7.7, *Seq2seq1(5)*). Pour réduire la distance syntaxique et lexicale entre les données d'apprentissage artificielles et l'ensemble de test réaliste, nous avons ajouté aux données d'apprentissage, 727 énoncés du corpus réaliste, spécifique au domaine domotique SWEET-HOME (chapitre 2, section 2.7.5). Des performances significatives ont été obtenues avec le modèle Att-RNN (tableau 7.7, *Artif. 42k*[15], *Att-RNN2(4)*).

7.2.2.2 Expérimentation de NLU sur le corpus VocADom@A4H

Au cours de nos travaux de thèse, nous avons ré-encodé l'approche seq2seq en utilisant une bibliothèque que nous avons développée au laboratoire [16] et qui est basée sur la bibliothèque PyTorch. Notre modèle seq2seq est composé d'un codeur et décodeur LSTM bidirectionnel et utilise l'attention de Luong. Nous énumerons ci-dessous les paramètres du modèle seq2seq et de l'apprentissage :

— Les mots d'entrée ont d'abord été transmis à une couche de 300 unités. L'encodeur et le décodeur étaient chacun une seule couche de 500 unités.

— L'algorithme d'optimisation Adam a été utilisé avec une taille de lot de 10, en utilisant un écrêtage de gradient à une norme de 2.0. une régularisation par abandon (*dropout*) a été fixé à 0.2.

15. Deuxième version améliorée du corpus artificiel, VocADom@ARTIF, de 42K phrases
16. https://gricad-gitlab.univ-grenoble-alpes.fr/getalp/seq2seqpytorch

— L'apprentissage s'est poursuivie sur 10000 étapes avec un taux d'apprentissage de 0.0001.

— La longueur de séquence d'entrée a été réglée à 50 mots et la longueur de séquence de sortie à 20 étiquettes.

— Pour le décodage, une recherche de faisceau de taille 4 a été effectuée.

L'étude de Mishakova et coll. (2019) a démontré qu'une approche NLU non alignée peut être compétitive avec une approche alignée, en ajoutant un minimum de données spécifiques au domaine à l'ensemble d'apprentissage. Cependant injecter des données de test dans le corpus d'apprentissage est une hypothèse non réaliste dans notre cadre. C'est pourquoi nous avons ajouté le corpus SWEET-HOME complet aux données d'apprentissage artificielles. Comme le montre le tableau 6.4, 61.28% du corpus VocADom@A4H consiste en intentions `none`. Les sections précédentes sur l'approche NLU alignée, ont rapporté des expérimentations sur l'ensemble de test VocADom@A4H, sur les phrases *avec* intention (38.71%) uniquement. Afin d'évaluer le corpus VocADom@A4H entièrement, nous avons ajouté les données ESLO2 aux 77K phrases du corpus artificiel complet comme indiqué dans le chapitre 6, section 6.4. Comme les transcriptions d'hypothèse du module de RAP Kaldi, seront les phrases d'entrée du modèle seq2seq, nous avons de nouveau utilisé les noms propres des mots-clés d'origine, au lieu de la version normalisée par 'KEYWORD'.

Le tableau 7.8 montre les mesures de l'évaluation des prédictions d'intention et des concepts sur VocADom@A4H. L'analyse des résultats montre une forte tendance vers les prédictions d'intention `none`, étant donné qu'il s'agit de la classe d'intention majoritaire. Une modification du poids dans la fonction de coût d'entropie croisée du modèle NLU a partiellement résolu le problème de données déséquilibrées. Cela a été calculé sur les données d'apprentissage complètes. Les poids résultants des classes ont été multipliés avec la perte d'entropie croisée, calculée par lot, comme le montre l'équation (7.1).

$$\text{poids\_classe\_i} = \frac{\text{instances\_total}}{\text{instances\_classe\_i}} \tag{7.1}$$

Par conséquent, la pénalité pour les classes d'intention *majoritaires* était moins forte que pour les classes d'intention moins représentées dans les données d'apprentissage. Par conséquent, l'apprentissage a augmenté pour les classes d'intention *minoritaire*. Cette méthode a amélioré les performances.

Les performances de NLU pour les prédictions d'intention sur les transcriptions d'hypothèse de sortie du module de RAP Kaldi, (*Pondéré-hyp. (Kaldi-Seq2seq-complet(7))*) sont légèrement moins bonnes que les prédictions de transcription de référence (*Pondéré-réf. (Seq2seq2(6))*), et montre 3% d'erreurs en plus pour les prédictions de concepts.

7.2.2.3 Bilan de l'approche NLU non-alignée

La non-normalisation des mots-clés dans les données d'apprentissage et de test, et l'ajout des phrases sans intention, ou avec la classe d'intention `none` aux données d'apprentissage et de test a diminué considérablement les performances par rapport à nos ap-

proches alignées, et l'approche non alignée sans classe d'intention `none`. Les performances chutent en particulier pour les classes d'intention. Nous avons essayé de gérer la situation des données déséquilibrées pour les classes d'intention `none` surreprésentées, en modifiant la fonction de coût d'entropie croisée, en attribuant plus de poids aux classes d'intention sous-représentées. Cela a considérablement amélioré les performances. L'évaluation du modèle seq2seq, en utilisant les transcriptions d'hypothèse de sortie de RAP Kaldi, montre bien l'impact des erreurs de transcription et réduisait les performances, en particulier pour les prédictions de concepts (cf. tableau 7.8).

7.3 Conclusion

Ce chapitre décrit notre approche de SLU séquentielle de référence qui se compose du module de RAP Kaldi et d'un module de NLU dont le noyau est une architecture seq2seq. Pour le module de RAP, nous avons obtenu les meilleurs résultats, un WER de 22.92%, en utilisant un modèle HMM-DNN et un ML qui est une interpolation d'un ML générique et un ML spécifique au domaine.

Pour valider notre approche de NLU, nous avons comparé 3 modèles de NLU alignés, en utilisant RASA, un outil Tri-CRF et Att-RNN. Le décalage entre les performances de ces modèles sur le corpus Port-Media et le corpus VocADom@A4H, au détriment du dernier ensemble d'évaluation, démontrait le défi de la distance entre les données d'entraînement synthétiques et l'ensemble de test réaliste.

Les performances du modèle de NLU Att-RNN aligné montrent des performances compétitives avec le modèle seq2seq non-alignée. Pour réduire les taux d'erreurs introduits par le passage du module de RAP au module de NLU, nous avons choisi le modèle non-aligné. L'augmentation des données d'apprentissage du corpus artificiel VocADom@ARTIF de phrases sans intention provenant du corpus de parole spontanée ESLO2, introduit une sur-représentation de la classe intention *none*. La modification de la fonction de coût d'entropie croisée résout partiellement ce problème. Nous avons obtenu alors les meilleures performances sur l'ensemble de test avec une F-mesure de 84.21% pour la prédiction d'intentions et un CER de 36.24% pour la prédiction des concepts. Il s'agit des résultats de notre approche de référence que nous comparons avec les performances de notre approche cible de SLU E2E dans le prochain chapitre.

TABLE 7.8 – Performances de SLU séquentielle (%) obtenues sur le corpus de test VocADom@A4H

Modèle	**Intention F-mesure**	**Concept CER**
Non pondéré-réf.	76.95	42.67
Pondéré-réf. (Seq2seq2(6))	**85.51**	**33.78**
Pondéré-hyp. (Kaldi-Seq2seq-complet(7))	**84.21**	**36.24**

CHAPITRE 8

Compréhension de la parole de bout en bout (E2E)

À la différence d'une approche séquentielle dont le dernier étage de classification n'a pas accès au signal d'entrée, une approche SLU E2E a un accès direct aux informations *acoustiques* pour inférer les concepts et les intentions. Grâce à l'optimisation conjointe des processus de RAP et de NLU, les performances finales ne dépendent que partiellement de la qualité de la RAP. Au cours de ce chapitre et le chapitre suivant, nous essaierons de comprendre **quels avantages une approche SLU E2E peut offrir par rapport à une approche en pipeline classique**.

Étant donné que l'approche SLU E2E prend comme entrée un signal acoustique et non un texte, nous avons *étendu* le corpus artificiel VocADom@ARTIF [1] avec des énoncés de *synthèse vocale*. Afin de vérifier si cette approche est valide, nous avons tout d'abord évalué les performances de RAP sur le corpus VocADom@A4H en utilisant des données d'apprentissage *avec* et *sans* parole artificielle. Afin de résoudre le problème de la distance entre les données de test réelles et les données d'apprentissage artificielles, nous avons vérifié l'impact de différentes *proportions de parole artificielle* dans les données d'apprentissage. Nous avons ensuite utilisé une technique d'*apprentissage par transfert* pour tirer le meilleur parti des données artificielles tout en respectant la contrainte de ne pas inclure de données de test réelles dans l'apprentissage. Nous expliquerons ces étapes dans les sections qui suivent.

Au cours de ce chapitre, nous référerons de nouveau à l'annexe A, figure A.1, qui donne un aperçu résumé des ensembles d'apprentissages.

8.1 La tâche de SLU vue comme un problème de transcription de parole enrichie

Pour la SLU E2E, nous nous sommes appuyé sur un modèle CNN-biLSTM pyramidal tel qu'utilisé par la plupart des modèles de RAP E2E. En effet, ces modèles étant capables de reconstruire des transcriptions complètes, il était pertinent de tester s'il pouvait également apprendre la tâche de RAP en même temps que celle de NLU (ici, trouver des frontières sémantiques durant la transcription).

Pour effectuer cette tâche, nous avons inséré les 8 symboles d'intention et les 17 étiquettes de concept (chapitre 6, section 6.3.3) dans les transcriptions des énoncés des don-

1. Pour des raisons de simplicité de lecture, le corpus artificiel VocADom@ARTIF sera dénommé Artif.

nées d'entraînement (*ESPnet-complet*, figure A.1). De cette manière, les données artificielles des données d'apprentissage et les 7% des phrases de données de parole réelle (sans intention) ont été enrichies de symboles de concept et d'intention. Nous avons effectué un *bootstrap* des annotations d'étiquettes de concepts symboliques à partir des données artificielles vers les énoncés sans intention dans les données d'apprentissage. Les 2 phrases suivantes sont extraites du corpus. La dernière phrase est un exemple d'un énoncé hors domaine sur laquelle nous avons appliqué un bootstrap à base du corpus VocADom@ARTIF enrichi de symboles.

```
v o c a d o m <space> ^ b a i s s e r ^ <space> } l u m i è r e }

c o n t i n u e z <space> j u s q u ' <space> à <space> } l a <space>
p o r t e } <space> d e <space> l a <space> c h a p e l l e
```

Dans la phrase exemple *"Continuez jusqu'à la porte de la chapelle"*, l'étiquette de concept *"La porte"* est `device`, bien que cet énoncé ne contienne pas d'intention. Cette technique nous a permis de générer plus d'étiquettes de concept dans les données d'entraînement de parole *réelle*. Pour les phrases d'intention `none` sans commande vocale, aucune annotation d'intention symbolique n'a été insérée.

Pour cette tâche nous nous retrouvons donc avec un grand nombre de *données artificielles* du domaine domotique et un grand nombre de *données réelles* hors domaine. Cette situation permet de couvrir d'une part la tâche du domaine d'application et, d'autre part, la variabilité acoustique de la parole.

8.2 Choix et paramétrisation du modèle E2E

Nous avons choisi ESPnet (Watanabe et coll., 2018) comme modèle E2E, car cet outil intègre la préparation de données et l'extraction de paramètres acoustiques de la boite à outils Kaldi qui est également utilisée pour construire le module de RAP de notre approche de SLU séquentielle de référence (chapitre 7, section 7.1). Nous étions également motivés par l'exploration de cet outil pour effectuer de la SLU E2E, étant donné les performances intéressantes qu'il a montré pour la RAP (chapitre 3, section 3.1.4.2).

Le tableau 8.1 contient les paramètres de l'état de l'art d'ESPnet que nous avons utilisés pour les expérimentations de RAP et de SLU E2E. Comme nous l'avons décrit dans le chapitre 3, section 3.1.1.1, plus récemment, les paramètres de *fbank* sont de plus en plus utilisés par les systèmes de RAP de réseaux de neurones profonds, au lieu des paramètres MFCC. Il s'avère en effet que les approches de réseaux de neurones profonds sont moins sensibles aux entrées hautement corrélées (Fayek, 2016). Comme les paramètres acoustiques du module de RAP Kaldi sont MFCC, nous avons également entraîné des modèles de RAP et de SLU E2E en utilisant des paramètres MFCC dans le contexte de l'analyse acoustique des performances de RAP et de SLU dans le chapitre 9, section 9.2.4. Nous avions initialement varié le poids α pour optimiser l'équilibre CTC/attention, mais étant donné que les meilleures performances obtenues en utilisant le poids $\alpha = 0.5$, nous avons fait le choix de cette valeur.

TABLE 8.1 – SLU E2E : paramètres choisis pour l'outil ESPnet

Paramètre	**Description**
Encodeur	CNN, 2 couches de VGG (very deep convolutional network) BLSTM de 4 couches bidirectionnelles avec 320 unités.
Décodeur	Une seule couche LSTM avec 300 unités. Taille de faisceau de 20.
Optimiseur	Adadelta
Taille de lot	30
Paramètres acoustiques	80 log mel-fbank pitch
CTC/attention	poids d'équilibre est $\alpha = 0.5$ (pour l'équation 3.18)
ML	RNN basé sur des caractères poids de $\beta = 1$ (équation 3.19) mêmes données que celles utilisées pour la RAP de SLU
Apprentissage	20 époques *Early stopping* après 3 époques durée totale de 129 heures (5.3 jours) 1 GPU, GeForce GTX TITAN Black

8.3 Performance de l'outil ESPnet en RAP

Afin d'éclairer l'interprétation des performances de cet outil de SLU E2E, nous avons d'abord entraîné un modèle de RAP ESPnet *avec* et *sans* parole de synthèse vocale à base du corpus artificiel VocADom@ARTIF. Le MA de RAP *sans* parole de synthèse vocale contient les mêmes données de parole réelle que celles du modèle de RAP Kaldi de la SLU séquentielle, soit 472.65 heures de données d'apprentissage (annexe A, figure A.1, *Réal.*). Le MA de RAP *avec* parole artificielle a été appris sur les mêmes données auxquelles nous avons ajouté les énoncés de synthèse vocale du corpus artificiel (annexe A, figure A.1, *Artif.*). Ces modèles de RAP nous permettent d'évaluer l'impact de la parole synthétique intégrée dans les données d'apprentissage afin de déterminer si l'augmentation des données de parole *réelle* de données de parole *artificielle* est une approche valide. Cela nous permet également d'estimer les paramètres DNN pour une approche de SLU E2E.

TABLE 8.2 – Performances de la RAP ESPnet sur l'ensemble de test VocADom@A4H

Modèle	**WER** (%)
Réal.	53.50
Réal.+ML	50.60
Réal.+ML+Artif.	**46.50**

Le tableau 8.2 présente les résultats en utilisant comme ensemble de test VocADom@A4H. Les performances obtenues par ESPnet (Modèle *Réal.*) sont très inférieures à celles de Kaldi (WER = 22.92%, tableau 7.4, DNN) alors qu'il s'agit des mêmes données d'apprentissage. Comme le montre la table 9.2 du chapitre 9, les types d'erreurs les plus fréquents pour ESPnet par rapport à Kaldi sont les erreurs de substitution. Celles-ci se produisent en particulier pour les noms propres. Elles sont également fréquentes sous forme

d'erreurs morphologiques pour les noms et les verbes, dont quelques exemples sont présentés dans la section 9.1.1. Cela montre l'impact d'une approche de RAP E2E à la fois sans ML à base de mots et sans lexique phonétique.

L'ajout du même ML que celui utilisé dans le cas de Kaldi (chapitre 7, section 7.1.1.3), à base de *caractères* au lieu de mots, améliore *légèrement* le WER (Modèle *Réal.+ML* de la table 8.2). L'ajout des données générées par synthèse vocale (Modèle *Réal.+ML+Artif.*) fournit par contre une amélioration *significative* bien que les performances restent toujours inférieures à celles du système Kaldi. Cependant, cette étude permet de valider l'intérêt des données artificielles (synthèse) pour l'apprentissage. Par ailleurs, nous verrons plus bas que même si les performances RAP de ESPnet sont deux fois moins bonnes que celles de Kaldi, nous pouvons atteindre des performances de SLU dépassant celles de l'approche séquentielle.

8.4 Apprentissage du modèle SLU E2E : impact des données artificielles

Pour l'apprentissage du modèle SLU E2E de ESPnet, nous avons utilisé les mêmes conditions d'apprentissage que pour celles choisies pour la RAP et en utilisant les transcriptions enrichies. Les paramètres du modèle sont ceux décrits en section 8.2.

De la même manière que pour la RAP, nous avons voulu mesurer l'impact de l'utilisation de données artificielles et réelles sur les performances de la tâche de SLU. Ainsi, différentes proportions de parole générée par synthèse vocale ont été utilisées pour constituer le corpus d'apprentissage. L'ensemble de test est toujours le corpus VocADom@A4H. Cette technique est également utilisée dans l'étude de Li et coll. (2018). Dans la suite de cette section, nous présentons tout d'abord l'effet sur la prédiction d'intentions puis sur la prédiction de concepts.

8.4.1 Prédiction d'intentions

Les résultats des différents modèles appris sont résumés dans le tableau 8.3. Ce tableau décrit :

— La quantité de données d'apprentissage qui sont constituées d'une combinaison d'énoncés de parole réelle et artificielle (+*Artif.*) par expérimentation (*Ensemble d'apprentissage*), *sans* ou *avec* un décodage en utilisant un modèle de langage (+*ML*).

— Le ratio de données générées par synthèse vocale dans le corpus d'apprentissage.

— Les performances (*F-mesure*) sur le corpus de test VocADom@A4H.

— Le ratio de phrases *sans* commande vocale (intention*None*).

Nous comparons également les performances en effectuant un décodage sans, et avec ML. La première expérience a consisté à apprendre un modèle uniquement à partir de l'ensemble d'apprentissage **+Artif.** (553.9 heures de parole, modèle entraîné sur toutes les données réelles et artificielles, annexe A, figure A.1, *ESPnet-complet(Artif.+Réal.)*). Les résultats

montrent que les classes d'intention ne sont *pas* bien prédites pour l'ensemble de test VocADom@A4H. Ces résultats indiquent une distance trop large entre les caractéristiques acoustiques des données artificielles de synthèse vocale et les données de parole réelle VocADom@A4H.

La deuxième expérience a consisté à déplacer 1k phrases de l'ensemble de test vers l'ensemble d'apprentissage. Dans ce cas, la prédiction d'intention augmente ce qui signifie que la prédiction des classes d'intention bénéficie davantage des données réelles ajoutées(**+Artif.+VocADom@A4H_1k**).

Ces deux premières expériences montrent également un biais d'apprentissage dû à l'intention `none` très majoritaire. Pour prendre en compte ce biais, nous avons traité les données sur-représentées ou sous-représentées de la manière suivante.

— **+Artif.+VocADom@A4H_1k+dim.** : en *diminuant* les instances de classe d'intention `none` *sur-représentées*. Nous avons réduit l'impact des énoncés sans commande vocale en ne laissant que 11k énoncés avec une étiquette de classe `none` dans le corpus d'apprentissage. Cette manipulation permet d'améliorer légèrement les performances.

— **+Artif.+VocADom@A4H_1k+augm.** : en *augmentant* les instances des classes d'intention *sous-représentées* `set_device_property`, `set_room_property`, `check_device`, `get_world_property`, `get_room_property`, jusqu'à environ 20k instances par classe. Ceci a eu pour conséquence d'augmenter la F-mesure.

— **+Artif.+VocADom@A4H_1k+dim.+ML**, **+Artif.+VocADom@A4H_1k+augm.+ML** : un décodage *avec* ML est effectué. Les symboles des classes d'intentions ont été ajoutés aux phrases des corpus artificiel et SWEET-HOME, faisant partie des données de ML, également utilisées pour le décodage du module de RAP de l'approche séquentielle de SLU (chapitre 7, section 7.1.1.3). L'ajout de connaissance *a priori* sur la tâche via le ML (au niveau caractère), améliore considérablement la F-mesure des intentions que cela soit dans le cas *dim.* ou *augm.* La matrice de confusion (figure 8.1) montre par contre que l'impact des classes majoritaires d'intentions des commandes vocales `set_device` et `None` n'a pas disparu.

La performance SLU E2E maximale a été atteinte, en variant le poids α pour optimiser l'équilibre pour l'apprentissage multi-tâche attention-CTC, et nous avons obtenu les meilleurs performances en utilisant le poids $\alpha = 0.5$. Pour le poids de β, nous avons obtenu les meilleurs résultats de 1, avec une influence maximale du ML.

8.4.2 Prédiction de concepts

Pour évaluer les performances d'inférence de concepts de l'approche SLU E2E, nous avons appris des modèles à partir de l'ensemble de données *complet* (tableau 8.3, + *Artif.*) que nous avons renommé *ESPnet-complet (9)* dans le tableau 8.4 ainsi que sur les données du corpus VocADom@ARTIF *uniquement* (*ESPnet-Artif.-uniq. (10)*). Le lecteur peut se référer au plan des corpus en annexe A, figure A.1 pour un meilleur suivi des expérimentations.

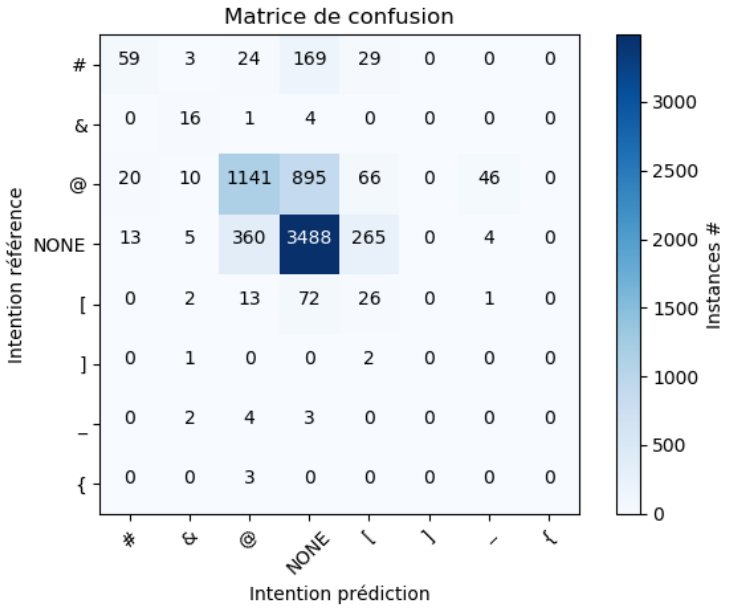


FIGURE 8.1 – Matrice de confusion de la prédiction d'intentions avec le modèle réduit de ESPnet

TABLE 8.3 – Évaluation de la prédiction d'intentions par ESPnet sur le corpus de test VocADom@A4H (F-mesure)

Ensemble d'apprentissage + (décodage sans/avec ML)	**Quantité (heures)**	**Ratio (%) synth. voc.**	**F-mesure test (%)**	**Ratio (%) None**
+Artif.	553.90	14.67	47.31	86.58
+Artif.+VocADom@A4H_1k	554.50	14.41	50.99	86.41
+Artif.+VocADom@A4H_1k+augm.	669.66	29.13	53.15	70.01
+Artif.+VocADom@A4H_1k+augm.+ML	669.66	29.13	67.95	70.01
+Artif.+VocADom@A4H_1k+dim.	84.69	94.39	53.92	13.70
+Artif.+VocADom@A4H_1k+dim.+ML	84.69	94.39	**70.21**	13.70

Le tableau 8.4 résume les performances obtenues sur le corpus de test. On peut voir que les performances de SLU E2E (*ESPnet-complet)*), pour la prédiction des concepts ne surpassent pas les performances de concepts de l'approche de SLU séquentielle de référence *Kaldi-Seq2seq-complet(7)* qui a été reportée dans ce tableau. Les pires performances sont affichées pour le modèle constitué uniquement de données artificielles (*ESPnet-Artif.-uniq.(10)*).

Enfin, lorsque l'apprentissage est effectué sur le même ensemble de données réduit que celui qui avait donné les meilleures performances pour la prédiction d'intentions (cf. tableau 8.3), alors on obtient les meilleures performances de prédiction de concepts avec un CER de 26,17% bien inférieur au CER de l'approche séquentielle (36,24 %). Ces résultats étant cohérents avec les résultats précédents, ceci valide l'approche SLU E2E comme une alternative crédible à l'approche séquentielle.

Afin de mieux comprendre la différence de capacité de généralisation des deux approches, nous avons comparé l'effet du corpus réduit sur Kaldi et ESPnet sur la tâche de RAP. L'apprentissage de Kaldi sur cet ensemble a conduit à un WER supérieur à 90% alors

que ESPnet affiche un WER de 60,6% sur l'ensemble de test. Il semble donc que ESPnet soit moins impacté par une faible quantité de données que les modèles DNN de Kaldi. En reconsidérant l'approche séquentielle avec le meilleur des deux modules RAP (ici ESPnet) le module NLU ne permet pas de surpasser les performances du module de SLU E2E (*ESPnet-Seq2seq-réduit*). C'est donc bien la tâche conjointe de bout-en-bout qui permet d'obtenir les meilleures performances de prédiction de concepts.

TABLE 8.4 – Évaluation de la prédiction de concepts par ESPnet sur le corpus de test VocADom@A4H, CER (%)

Ensemble d'apprentissage	**Quantité (heures)**	**Ratio (%) de synthèse vocale**	**CER Concept**	**F1 Intention**
Kaldi-Seq2seq-complet	472.65	0.00	36.24	84.21
ESPnet-complet	553.90	14.67	51.87	47.31
ESPnet-Artif.-uniq.	81.25	100.00	56.00	35.94
ESPnet-réduit	84.69	94.39	**26.17**	70.21
ESPnet-Seq2seq-réduit	84.69	94.39	35.62	61.35

8.4.3 Bilan

Les performances obtenues avec ESPnet pour la prédiction de concepts par SLU sont bien meilleures que dans le cas de la RAP E2E. Cela semble donc confirmer que la SLU E2E ne semble dépendre que *partiellement* des performances de la RAP. Le choix du corpus d'apprentissage s'est avéré extrêmement important. En effet, nous avons obtenu les meilleures performances pour notre approche de SLU E2E sur un ensemble de données d'apprentissage bien plus réduit que le corpus complet. Ce corpus réduit se compose d'énoncés du corpus artificiel, d'un minimum de données réalistes incluant le corpus SWEET-HOME et également 1000 énoncés qui ont été déplacés du corpus de test VocADom@A4H vers l'ensemble de données d'apprentissage.

Avec un modèle ESPnet qui n'est entraîné *que* sur des énoncés synthétiques (tableau 8.4, *ESPnet-Artif.-uniq.*), il s'avère que la prédiction des concepts est possible mais très erronée. Par contre, en combinant ce corpus artificiel avec des énoncés réalistes, nous obtenons de bonnes performances globales qui permettent de tirer un bon parti du corpus artificiel pour l'adéquation à la tâche et des données réalistes. Cependant, pour atteindre ces performances, il a fallu déplacer une partie des énoncés de parole *réelle* de l'ensemble de test VocADom@A4H vers les données d'apprentissage, ce qui viole nos conditions de départ (apprentissage sans exemples du corpus de test). Cependant, ces expérimentations nous fournissent les informations suivantes sur l'approche E2E.

— Il est faisable d'apprendre un modèle crédible en utilisant de *petits* ensembles de données spécifiques au domaine.

— Un modèle peut être appris à l'aide de données d'apprentissage *artificielles*

— L'augmentation de la proportion de données d'apprentissage *réelles*, spécifiques au domaine augmente également les performances.

— Une *combinaison* naïve des données réelles et artificielles n'apporte *pas* nécessairement une hausse de performances.

8.5 Apprentissage par transfert

Une autre façon de tirer parti d'un grand ensemble de données hors domaine et d'un petit ensemble de données du domaine est d'utiliser une approche d'apprentissage par transfert, approche que nous avons choisie au vu de ses performances prometteuses (chapitre 4, section 4.2.3). Dans notre cas, il s'agit d'initialiser un modèle avec la grande quantité de données de parole sur une tâche de RAP afin que le modèle apprenne les représentations d'entrée du signal acoustique. Lorsque le modèle est appris, l'apprentissage est relancé sur la tâche de SLU avec d'autres jeux de données conçues spécifiquement pour apprendre les concepts et les intentions. Nous effectuerons ce type d'apprentissage en 4 étapes. Il s'agit d'une première étape pour la RAP, de 2 étapes pour la prédiction des concepts et d'une quatrième étape pour la prédiction des intentions. La figure 8.2a illustre la façon dont se déroulent ces 4 étapes, la figure 8.2b schématise la façon dont les corpus sont utilisés :

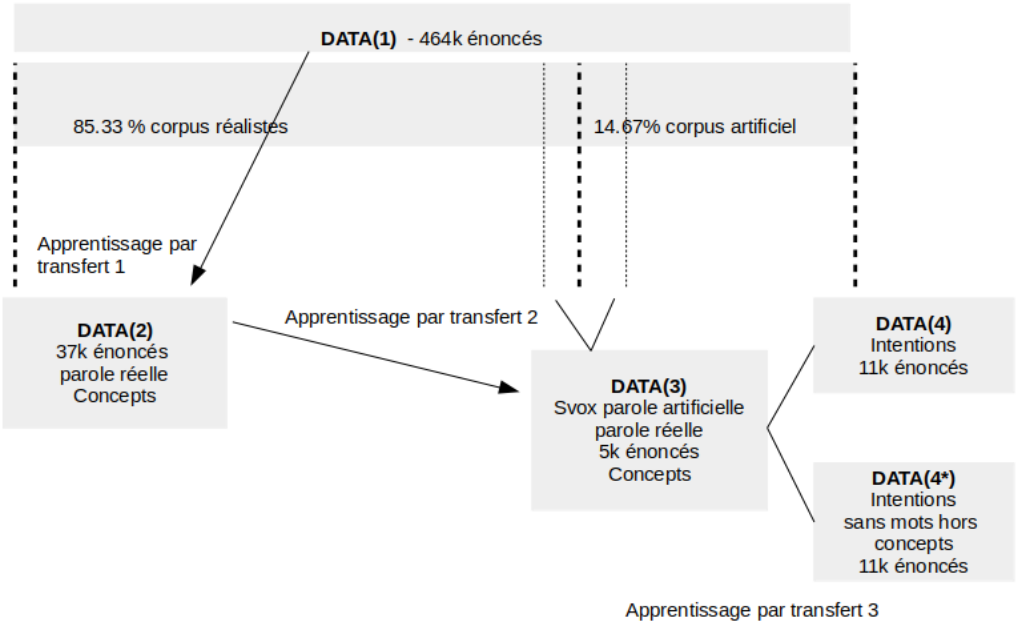
1. La première étape est l'apprentissage du modèle de RAP qui est appris sur l'ensemble des énoncés de la parole réelle et artificielle (553.9 heures, section 8.3), que nous avons appelé *data(1)* (figure 8.2).

2. Pour la deuxième étape, il y a un apprentissage sur les données de parole *réelle*, faisant partie des 553.9 heures de *data(1)*, qui contiennent des concepts symboliques spécifiques au domaine domotique. Nous avons appelé l'ensemble de ces 37k phrases *data(2)*

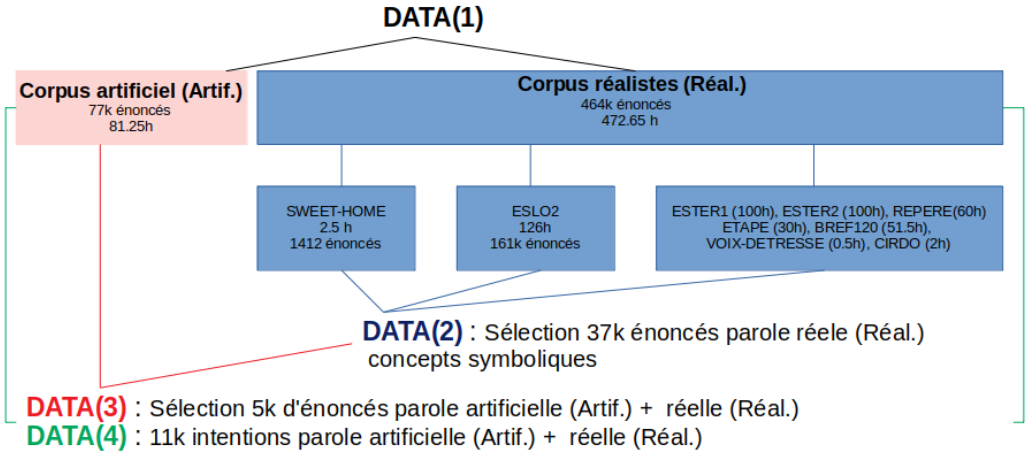
3. La troisième étape est un apprentissage sur une partie des énoncés du corpus VocADom@ARTIF, dont les concepts *manquent* ou sont *sous-représentés* dans la partie des énoncés de la parole *réelle*. Il s'agit d'une sélection de 3800 phrases extraites du corpus artificiel, desquelles nous décrivons la procédure de sélection dans la section 8.5.1. Nous avons également sélectionné 1651 énoncés, extraits des 37k phrases de *data(2)* contenant des concepts sous-représentés. D'une telle façon, nous combinons un apprentissage par *transfert* et une technique de *duplication* de données. Nous appelons ces 5451 énoncés sélectionnés du troisième ensemble d'apprentissage *data(3)*.

4. L'étape finale est un apprentissage des énoncés contenant une intention. Il s'agit de *data(4)* qui contient 11K énoncés de la parole réelle et artificielle contenant les 8 intentions.

Ainsi l'apprentissage par transfert permet non seulement de tirer le meilleur parti de données hors domaine mais aussi d'équilibrer la couverture des classes à prédire. La figure 8.2 résume les corpus utilisés, le lecteur est invité à s'y référer au fil de cette section.



(a) Méthode d'apprentissage



(b) Utilisation des des corpus pour l'apprentissage par transfert

FIGURE 8.2 – Méthode et corpus utilisés pour l'apprentissage par transfert

8.5.1 Sélection des énoncés du corpus artificiel pour apprentissage par transfert

La figure 6.7 du chapitre 6 montre que les proportions d'intentions dans le corpus VocADom@ARTIF sont similaires à celles de l'ensemble de test VocaDom@A4H. Par conséquent, nous avons estimé que la combinaison totale de données *d'apprentissage* réalistes et artificielles devrait refléter les proportions d'intentions du *corpus artificiel* (l'intention `None` non incluse).

Les expérimentations des sections précédentes nous ont appris que nous devons essayer de conserver *autant de données réalistes* domotiques que possible, dans les données d'apprentissage. En revanche, ces expérimentations ont aussi montré que les données artificielles contribuaient également à de meilleures performances de SLU E2E. Pour cette raison, nous avons composé *data(3)* en utilisant les données artificielles uniquement pour *combler*

TABLE 8.5 – Fréquences d'apparition de valeurs de concepts sous-représentées dans data(2)

Concept	**Valeur**	**# Artif.**	**# Réal./data(2)**	**# Sélection/data(3)**
device-setting	éteinte	590	10	46
device	aspirateur	298	8	90
device	réfrigérateur	266	3	14
device	hotte	561	6	27
device	bouilloire	88	1	61

les lacunes des valeurs de concepts sous-représentées dans les données réalistes de *data(2)*. Nous avons également repris les énoncés réalistes de *data(2)* avec des valeurs de concepts peu fréquentes.

À cette fin, nous avons calculé les fréquences des valeurs liées à des concepts dans le corpus artificiel et les 37k énoncés de *data(2)*. Le corpus *data(3)* est composé de 5451 énoncés dont 3800 sont issus du corpus artificiel et 1651 de data(2). La sélection des énoncés artificiels s'est faite de telle manière :

— Concernant le corpus des données réalistes *data(2)*, contenant des étiquettes sémantiques symboliques, nous avons sélectionné toutes les valeurs des concepts avec une fréquence de moins que 20 dans *data(2)*.

— Ensuite les énoncés du corpus artificiel qui contiennent ces valeurs sous-représentées du *data(2)* ont été sélectionnés, à condition qu'ils ne dépassent pas la longueur moyenne des énoncés du corpus artificiel de 15 mots.

La table 8.5 présente pour chaque *concept* un exemple de *valeur*, sa fréquence dans le corpus artificiel (*# Artif.*) et leur fréquence basse dans *data(2)* (en colonne *# Réal/data(2)*). La colonne *# Sélection/data(3)* montre la fréquence obtenue pour ces différentes valeurs de concepts après sélection pour l'ensemble de données *data(3)*.

8.5.2 Résultats de l'apprentissage par transfert de l'approche SLU E2E

Pour les phases d'apprentissage par transfert (figure 8.2), nous avons utilisé les mêmes paramètres que ceux spécifiés dans la section 8.2. Par contre, le nombre d'époques en appliquant un *early stopping* après 3 époques où le modèle n'apprend plus, est moins de 20 pour les 2 aprentissages par transfert :

— Data(1) → Data(2) : 16 époques

— Data(2) → Data(3) : 12 époques

— Data(2) → Data(3) → Data(4) : 9 époques

— Data(2) → Data(3) → Data(4*) : 11 époques

Les résultats sont présentés sur le tableau 8.6 qui contient les résultats des performances pour toutes les phases d'apprentissage. Les deux résultats de référence avec Kaldi (*Kaldi-Seq2seq-complet*) et *ESPnet-reduit* sont affichés dans le tableau. Pour rappel, *ESPnet-réduit* inclut des données de tests dans son MA. Les résultats du transfert de la tâche de RAP

(*Data(1)*) vers la tâche de SLU (*Data(2)*), (*Data(1)* → *Data(2)*), montre des performances inférieures à l'approche avec des données réduites. Afin de vérifier que ces résultats sont basés sur un *vrai* apprentissage par transfert, nous avons comparé des performances basées sur un modèle *uniquement* appris sur les 5k énoncés de *Data(3)*. Un apprentissage uniquement sur *data(3)* montre bien ses limites (CER montant à 69,11%). Par contre, un apprentissage sur la tâche de SLU *data(2)* puis transféré sur la tâche de SLU *data(3)* (*Data(2)* → *Data(3)*) montre son efficacité en obtenant un CER plus performant (32,12%) que l'approche séquentielle *Kaldi-Seq2seq-complet* (CER = 36,24%). Ce résultat reste inférieur à l'approche *ESPnet-réduit* mais cette approche n'est pas comparable car ce dernier modèle contient 1k énoncés de l'ensemble de test. On peut donc conclure que l'apprentissage par transfert est pertinent pour la SLU E2E.

Pour la prédiction d'intentions, nous avons continué le principe d'apprentissage par transfert en utilisant les données d'intention (*data(4)*). Pour apprendre cette tâche d'apprentissage d'intention, les transcriptions enrichies, contenant des symboles de concepts, ont été complétées par des symboles d'intention. Le tableau 8.7 donne des exemples de ces transcriptions enrichies pour chaque tâche, pour une commande vocale avec l'intention `set_device`. Comme comparaison avec Ghannay et coll. (2018), nous avons également créé une autre version du corpus (*data(4)*) dans lequel les termes non liés à un concept ont été remplacés par des astérisques (*data(4*)*).

La table 8.6 montre qu'en utilisant un apprentissage par transfert, nous n'avons pas réussi à surpasser l'approche SLU séquentielle de référence, quant à la prédiction d'intentions. Nous avons obtenu le meilleur résultat en utilisant le modèle *Data(2)* → *Data(3)* → *Data(4*)* où les termes "hors concepts" ont été remplacés par des astérisques. Néanmoins, les performances quant à la prédiction d'intentions avec un apprentissage par *transfert* surpassent celles du modèle réduit (*ESpnet-réduit*). La matrice de confusion pour la prédiction d'intentions avec un apprentissage par *transfert* (figure 8.3) montre de meilleures performances pour les classes majoritaires d'intentions `set_device` et `none`, par rapport à l'approche avec le modèle réduit (*ESpnet-réduit*, figure 8.1), non pour les classes minoritaires.

TABLE 8.6 – Évaluation des performances en SLU de ESPnet avec et sans apprentissage par transfert testées sur le corpus VocADom@A4H

Modèle	**Intention (%)** **F-mesure**	**Concept (%)** **CER**
Apprentissage sans transfert :		
Kaldi-Seq2seq-complet	**84.21**	36.24
ESPnet-réduit	70.21	**26.17**
Data(3)	-	69.11
Apprentissage par transfert :		
Data(1) → Data(2)	-	42.19
Data(2) → Data(3)	-	**32.12**
Data(2) → Data(3) → Data(4)	68.13	-
Data(2) → Data(3) → Data(4*)	**74.57**	-

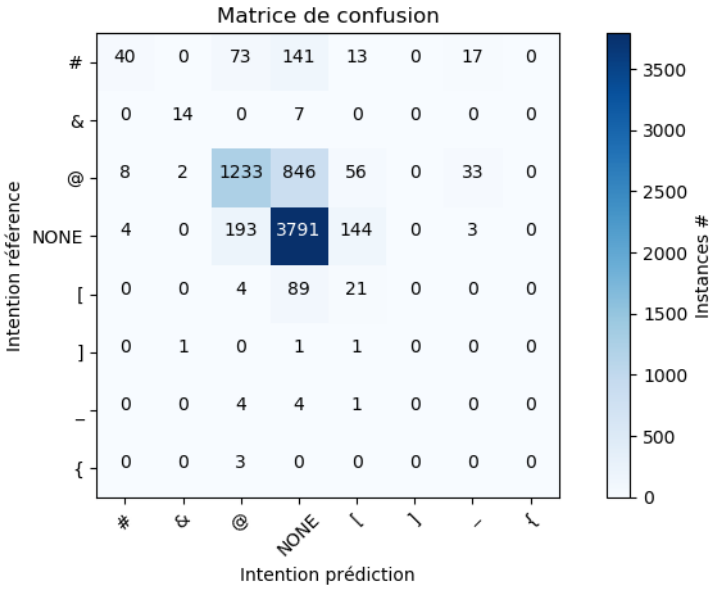


FIGURE 8.3 – Matrice de confusion de la prédiction d'intentions par la SLU ESPnet avec apprentissage par transfert

TABLE 8.7 – Symboles d'intentions et de concepts utilisés pour la SLU E2E

Concept (data(3))
hestia s'il vous plaît ^baisser^ }la lampe} >de la chambre> `action device location-room`
Intention + Concept (data(4))
@@ hestia s'il vous plaît ^baisser^ }la lampe} >de la chambre> @@ `set_device action device location-room`
Intention + Concept - sans mots hors des concepts (data(4*))
@@ hestia * * * ^baisser^ }la lampe} >de la chambre> @@ `set_device action device location-room`

8.6 Conclusion

Ce chapitre a présenté l'approche SLU E2E proposée par cette thèse. Nous avons posé le problème de SLU E2E comme un problème de transcriptions de parole enrichies avec des symboles de classes d'intentions et d'étiquettes de concepts. Ce cadre posé, nous avons choisi une architecture de modèle SLU E2E basée sur les modèles de RAP E2E. Dans notre étude, nous utilisons le modèle pyramidal implanté dans ESPNet. Pour constituer les données de cette tâche de SLU, nous avons automatiquement inséré ces symboles dans les phrases des corpus réalistes et artificiels des données d'apprentissage. Ce sont les données générées artificiellement qui nous ont permis de traiter le cas de données manquantes en équilibrant la couverture des concepts dans les données d'apprentissage.

Nous avons mesuré l'impact de l'intégration de la parole artificielle (synthèse de parole) aux données d'apprentissage au départ constituées de parole réelle sur une tâche de RAP de bout en bout. Cette expérimentation a montré que cet ajout améliore les performances de la RAP. Ceci nous indique que l'augmentation des données de parole réelle de données de

parole artificielle constitue une approche valide pour effectuer de la SLU de bout en bout.

Nous avons ensuite abordé la tâche d'apprentissage de modèle SLU de bout en bout en envisageant différentes compositions du corpus de données. En utilisant un ensemble d'apprentissage réduit, avec un ajout minimal des données de tests, nous avons obtenu les meilleures performances et surpassé les résultats de l'approche SLU séquentielle pour la prédiction des concepts. Ces expériences ont aussi montré qu'il n'est pas nécessaire d'obtenir des performances de RAP parfaites pour obtenir de bonnes performances de SLU E2E.

Comme notre objectif final est bien évidemment de développer un outil de SLU de bout en bout *sans* inclure des données de test dans les données d'apprentissage, nous avons alors conçu une approche *d'apprentissage par transfert* pour laquelle les paroles réelle et artificielle sont mises à profit durant plusieurs étapes d'apprentissages. En appliquant cette technique, les résultats de prédiction de concepts surpassent les performances du modèle de SLU séquentielle. Les résultats de la prédiction d'intentions restent par contre inférieurs à ceux de l'approche séquentielle. Ce chapitre montre que l'approche SLU de bout en bout est une approche prometteuse et qu'elle permet bien de contourner le problème de cascade d'erreurs de l'approche séquentielle. La difficulté d'accès à des données d'apprentissage, qui sont de taille très réduite pour un nouveau domaine, reste un défi de taille. Si les modèles profonds sont réputés pour être très dépendants des données massives, nous avons cependant pu montrer qu'une sélection attentive des données en un corpus restreint permet d'obtenir des résultats acceptables.

À la différence d'une approche SLU séquentielle, une approche SLU de bout en bout a accès aux informations *acoustiques* du signal d'entrée. Il est donc pertinent de se poser la question de savoir si le modèle exploite réellement les données para-linguistiques pour inférer des informations sémantiques. Dans le chapitre suivant, nous étudions donc dans quelle mesure ces informations acoustiques contribuent aux performances de SLU E2E.

CHAPITRE 9

Analyse des performances de SLU

Ce dernier chapitre présente une analyse des performances de RAP et de SLU du modèle séquentiel de référence et du modèle cible de bout en bout. Cette analyse est effectuée aux niveaux *acoustique* et *symbolique*. Le tableau 9.1 donne une vision synthétique des niveaux d'analyse que nous allons maintenant considérer au cours des sections suivantes. Cette évaluation comprend également des réponses à deux questions de recherche que nous avons posées dans l'introduction du chapitre 5 :

1. En appliquant la technique d'un apprentissage par transfert, la prédiction des concepts du modèle SLU E2E, inférés directement du signal, surpasse l'approche SLU séquentielle comme on l'a montré au cours du chapitre 8. Est-ce qu'on peut supposer que **le modèle E2E exploite les informations au niveau prosodique et acoustique de telle manière que les performances de prédiction d'intentions et de concepts soient améliorées** ?
2. Un des défis de recherche de cette thèse est le développement d'un modèle SLU qui soit robuste à la *variation linguistique*, notamment au niveau *lexical* et *syntaxique*, introduite par ses différents utilisateurs. **Est-ce qu'un modèle SLU E2E est robuste aux variations grammaticales** ? Peut-il modéliser les mots hors vocabulaire et les syntaxes inusuelles, mieux qu'un système de SLU séquentielle ?

TABLE 9.1 – Niveaux d'analyse des performances des systèmes de RAP et de SLU séquentiels ou de bout en bout sur le corpus de test VocADom@A4H

Niveau d'analyse :	**Acoustique**				**Symbolique**	
	Source		**Prosodie**		**Lexical**	**Syntaxique**
Analyse :	**Bruit**	**Genre**	**F0 moyenne**	**Délexicalisation**	**OOV**	**Variation**
Hypothèse :	(1) Le modèle E2E exploite-t-il l'information **prosodique, acoustique ?** (2) Est-ce qu'un modèle SLU E2E est plus robuste au niveau **symbolique**, aux variations du **vocabulaire** et de la **syntaxe** ?					

Nous essayons également de démontrer que le modèle de SLU E2E peut bien fonctionner *sans* nécessairement être très performant au niveau de RAP. À cette fin, nous calculons les corrélations entre les performances de RAP et de SLU. À la différence du modèle séquentiel, le modèle E2E a accès aux informations acoustiques et prosodiques, c'est pourquoi nous avons analysé leur impact sur les performances de la RAP et de la SLU. Pour ce faire, nous avons estimé la corrélation entre des paramètres acoustiques tels que le *pitch* et l'énergie et des mesures de performances telles que le WER dans le cas de la RAP et le CER dans celui

de la SLU. Une partie des énoncés du corpus VocADom@A4H sont affectés par un bruit de fond. La case correspondante est notée *Bruit* dans la table 9.1. Par conséquent, nous allons déterminer dans quelle mesure ces données bruitées ont un impact sur les performances des modèles séquentiels et de bout en bout. À cette fin, nous appliquons la technique de *délexicalisation*. Nous évaluons également les performances en synthétisant, à partir des données réelles, de la parole fixant pour *F0* la valeur de *F0 moyenne* du locuteur.

Cette analyse évalue également l'impact des mots hors vocabulaire (OOV – *Out Of Vocabulary*) et de la variation syntaxique, car nos utilisateurs cibles sont des personnes âgées qui ont tendance à s'écarter d'un ensemble prédéfini de commandes vocales. Pour effectuer cette analyse, au niveau *symbolique*, nous avons remplacé le vocabulaire faisant partie des concepts dans l'ensemble de test par un vocabulaire qui ne fait pas partie des données d'apprentissage. Au niveau *syntaxique*, des verbes ont été substitués par des structures syntaxiques plus complexes et des disfluences y ont été ajoutées. Finalement, nous avons analysé les erreurs spécifiques à certains locuteurs et aux locuteurs masculins et féminins (*Genre*).

9.1 Analyse acoustique des performances de la RAP et de la SLU

En tant que étape préparatoire avant d'effectuer la tâche de SLU de bout en bout, en utilisant l'outil de RAP ESPnet, un modèle de RAP a d'abord été entraîné avec de la parole de synthèse vocale à base du corpus artificiel VocADom@ARTIF, (chapitre 8, section 8.3). Les meilleures performances de ce modèle (46.50% WER) n'ont pas surpassé celles du module de RAP Kaldi (22.92 % WER), faisant partie de l'approche SLU séquentielle (chapitre 7, section 7.1).

Bien que les performances de RAP de l'outil ESPnet ne surpassent pas les performances de Kaldi, nous avons montré dans le chapitre précédent qu'il peut quand même mieux prédire les concepts dans une tâche de SLU E2E par rapport à une approche SLU séquentielle, en appliquant un apprentissage par transfert. L'analyse des résultats de RAP et de SLU E2E dans les sections suivantes donne des pistes pour comprendre comment cela est possible.

9.1.1 Performances de RAP

La table 9.2 donne un aperçu des nombres des mots corrects (C), supprimés (D), substitués (S) et insérés (I) pour les systèmes de RAP de KALDI et d'ESPnet. Le plus grand nombre d'erreurs se produit au niveau des *substitutions*, notamment pour ESPnet. La table 9.3 montre les erreurs de RAP pour les mots-clés (Chapitre 6, section 6.1.1), dont la classe de substitutions est également la plus grande. Il s'avère que 19,93% des erreurs de substitution d'ESPnet concernent les noms propres ou des mots-clés, ce qui n'empêche pas nécessairement une bonne performance de SLU comme le montre l'exemple suivant :

```
REF:  ** Hé CIRRUS ^éteignez^ }la télé}
```

```
HYP:  ET SI  VOUS  ^éteignez^ }la télé}
Eval: I  S   S
```

Le mot-clé 'Cirrus' de l'énoncé de référence (*REF :*) est mal prédit dans la transcription d'hypothèse (*HYP :*). En revanche, cette erreur n'empêche pas la bonne prédiction des concepts `action` (^éteignez^) et `device` (}la télé}).

En outre la SLU peut parfaitement être effectuée sur des mots avec une prédiction de RAP erronée de substitution. Le tableau 9.4 montre que *toutes* les prédictions de RAP d'ESPnet pour le mot 'baissez' (concept `action`) sont fausses. La transcription d'hypothèse pour ce mot cible avec une fréquence de 45 dans l'ensemble de test VocADom@A4H, est 24 fois 'baisser', et 4 fois 'baisse'. Néanmoins, le concept a été bien prédit, comme le montre l'exemple suivant :

```
REF:  vocadom ^BAISSEZ^ }le store}
HYP:  vocadom ^BAISSER^ }le store}
Eval:             S
```

'baissez' est substitué par 'baisser', ce qui n'empêche pas la bonne prédiction du concept `action`.

L'omission de mots extérieur aux concepts n'empêche par la détection des concepts. L'exemple suivant montre la bonne prédiction des concepts `action` (^éteindre^) et `device` (}les lumières}), bien que le mot-clé et l'adjectif 'toutes' soient omis de la transcription d'hypothèse (*HYP :*) :

```
REF:  ULYSSE ^éTEINS^    TOUTES    }les lumières}
```

TABLE 9.2 – Performances globales de la RAP des systèmes Kaldi et ESPnet

	Kaldi		**ESPnet**	
Type de sortie	Mots #	Mots #(%)	Mots #	Mots #(%)
Mots corrects (C)	26193	77.08	18180	53.50
Mots supprimés (D)	2739	8.06	2080	6.12
Mots substitués (S)	*3541*	*10.42*	*10942*	*32.20*
Mots insérés (I)	1509	4.44	2780	8.18
Nombre total de mots	33982			
WER (%)	22.92		46.50	

TABLE 9.3 – Performances des systèmes de RAP Kaldi et ESPnet sur la reconnaissance des mots-clés

	Kaldi		**ESPnet**	
Type d'erreur	Mots #	Mots #(%)	Mots #	Mots #(%)
Mots corrects (C)	2398	90.66	415	15.68
Mots supprimés (D)	55	2.07	42	1.59
Mots substitués (S)	177	6.69	2181	82.46
Mots Insérés (I)	15	0.56	7	0.26
Nombre total de mots-clés	2645			

```
HYP:  PUISSE ^éTEINDRE^  ******    }les lumières}
Eval:    S        S         D
```

L'exemple final semble indiquer que le mécanisme d'attention, en interaction avec la fonction de CTC, comprend qu'il faut se concentrer sur la bonne prédiction des concepts en forme de symboles. La prédiction correcte des caractères qui l'entourent peut être considérée comme moins importante. Autrement dit, malgré une transcription de parole erronée d'un ou plusieurs mots à l'intérieur d'un concept, le modèle E2E arrive quand même à inférer les étiquettes correctes. Le concept `device` est correctement prédit, bien que la transcription de sa valeur, 'le ventilateur', soit erronée.

```
REF:  minouche ^allumer^  }LE  ***   VENTILATEUR}
HYP:  minouche ^allumer^  }LA PORTE} NATEUR
Eval:                       S   I       S
```

9.1.2 Performances de SLU

Après cette analyse des performances de la RAP, nous analysons les performances de la SLU pour ensuite calculer la corrélation entre les deux.

9.1.2.1 Prédictions des intentions

L'approche SLU E2E ne surpasse pas l'approche SLU séquentielle de référence pour la prédiction d'intentions. Pour les 2 approches, les performances de prédiction d'intentions sont les plus faibles pour les classes minoritaires dans les données d'apprentissage et de test, et les plus élevées pour les classes majoritaires dans les données d'apprentissage et de test, comme indiqué dans les tableaux 9.5 et 9.6. L'apprentissage du modèle seq2seq du module de NLU de l'approche SLU séquentielle (chapitre 7, section 7.2.2.2), s'est effectuée en pondérant les classes pour éviter les biais dus aux classes majoritaires/minoritaires. Cependant l'influence des classes majoritaires n'a pas disparu.

De la même manière, l'équilibrage des instances de classes d'intention dans le corpus d'apprentissage ont amélioré les performances pour le SLU E2E (chapitre 8), mais l'influence des intentions des classes majoritaires continuent à avoir un fort impact.

TABLE 9.4 – Performances de la RAP des systèmes Kaldi et ESPnet sur la reconnaissance du mot « baissez »

Mot de référence	**Absence ou type d'erreur**	**Kaldi**	**ESPnet**
baissez	Correct (C)	7	0
	Supprimé (D)	3	5
	Substitué (S)	35	40
baissez substitué par			
	baisser	33	24
	baisse	1	4
	téraphim	1	-
	est	-	6
	à, et, les, mais, si…	-	6

TABLE 9.5 – Performances de la prédiction d'intention par la SLU séquentielle testée sur le corpus VocADom@A4H

Intention	Précision (%)	Rappel (%)	F-mesure (%)	#Intentions réf.
check_device	47	36	41	284
contact	0	0	0	114
get_room_property	0	0	0	3
get_world_property	0	0	0	3
none	88	97	92	4135
set_device	89	76	82	2178
set_device_property	0	0	0	9
set_room_property	100	5	10	21
Moyenne	84.87	84.74	84.21	6747

TABLE 9.6 – Performances de la prédiction d'intention par la SLU de bout en bout testée sur le corpus VocADom@A4H

Intention	Précision (%)	Rappel (%)	F-mesure (%)	#Intentions réf.
check_device #	77	14	24	284
contact [	9	19	12	114
get_room_property {	0	0	0	3
get_world_property]	0	0	0	3
none	77	93	84	4135
set_device @	85	57	68	2178
set_device_property _	0	0	0	9
set_room_property &	81	65	74	21
Moyenne	78.31	75.61	74.57	6747

9.1.2.2 Prédictions des concepts

En utilisant une combinaison de SLU E2E et un apprentissage par transfert, où la parole synthétique générée sur les données du corpus artificiel n'était que partiellement utilisée, pour compenser les concepts peu représentés dans les données d'apprentissage, nous avons pu surpasser la prédiction de concepts dans l'approche SLU séquentielle.

Le détail des performances de la prédiction de concepts des deux approches est présenté sur le tableau 9.7. Similaire à la prédiction des intentions, l'influence des étiquettes de concept majoritaires a un fort impact sur la prédiction de concepts en général, malgré la modification de la fonction de coût de l'approche NLU du modèle de SLU séquentielle et l'équilibrage du corpus d'apprentissage dans le cas E2E. Les performances pour les étiquettes minoritaires comme `device-component`, `location-house`, `location-inroom`, sont significativement pires pour l'approche séquentielle de SLU ainsi que pour l'approche E2E.

Ces erreurs se produisent partiellement à cause des mots hors vocabulaire. Il se produit que certains concepts qui apparaissent fréquemment dans les données d'apprentissage ont une valeur, ou terme *synonyme* qui n'apparaît pas dans les données de test. D'une telle façon `location-house` est un concept avec une haute fréquence dans le corpus d'apprentissage VocADom@ARTIF (47643 instances) avec "maison" comme valeur associée. Néanmoins les quelques occurrences de ce concept dans les commandes vocales des données du corpus de test sont dans la plupart des cas associées à la valeur "appartement". De la même façon, la valeur associée au concept `device-component` est "station" dans le corpus d'apprentissage ("minouche changez la *station* de la radio dans la chambre"). Par contre la valeur

associée au concept `device-component` dans le corpus VocADom@A4H est également "chaîne" ("ulysse change de *chaîne* de radio").

Nous avions composé les données d'apprentissage du corpus artificiel pour que les proportions des concepts et les intentions soient similaires à celles du corpus de test VocADom@A4H. En conséquence, des confusions se produisent entre les étiquettes de concepts minoritaires telles que `location-inroom` d'une part, et les étiquettes de concepts plus fréquemment représentées, telles que `location-room` d'autre part. Pour l'énoncé de test "allume la lumière de chevet", "de chevet" est prédit comme `location-room` au lieu de `location-inroom`. Ce biais peut s'expliquer par des énoncés fréquents dans les données d'apprentissage où le concept `location-inroom` est précédé par le terme "lumière" et suivi d'une pièce avec la préposition "de" ("de la cuisine", "du bureau"," du salon", etc.)

Pour tous les autres concepts le modèle E2E montre une performance équivalente ou une diminution nette.

9.1.3 Corrélations entre les performances de RAP et de SLU

Dans la section 9.1.1 nous avons montré qu'un modèle de SLU E2E n'a pas besoin de transcriptions de RAP parfaites pour extraire des concepts. Cela signifie que nous nous attendrions à une corrélation faible entre les valeurs de WER des performances de RAP et les valeurs de CER de la prédiction de concepts. Afin de vérifier cela, nous avons calculé les coefficients de corrélation de Pearson et de Spearman (chapitre 5, section 5.6) entre les valeurs de WER moyennes par énoncé, du modèle de RAP E2E (chapitre 8, section 8.3, tableau 8.2, *RAP E2E Réal.+ML+Artif.*) et les valeurs moyennes de CER par énoncé du modèle de SLU E2E qui est le résultat d'un apprentissage par transfert (chapitre 8, section 8.5, tableau 8.6, *SLU E2E*

TABLE 9.7 – Comparaison des performances de la prédiction de concept par les SLU séquentielle et de bout en bout testées sur le corpus VocADom@A4H

Concept	**Symbole**	**SLU séq. CER (%)**	**SLU E2E CER (%)**	**#Concept réf.**
action	^	19.90	24.33	2211
device	}	19.45	25.7	2473
device-component	*	60.00	100.00	5
device-setting	,	50.70	51.05	284
location-floor	;	96.14	88.81	805
location-house	!	100.00	100.00	9
location-room	>	21.42	16.95	1055
location-inroom	?	100.00	100.00	34
organization	§	100.00	100.00	4
person-occupation	=	100.00	100.00	2
room-property	/	75	33.33	24
value-numeric	%	100.00	100.00	1
value-qualitative	|	100.00	100.00	2
world-property	∘	100.00	100.00	3
Moyenne		36.24	32.12	6912

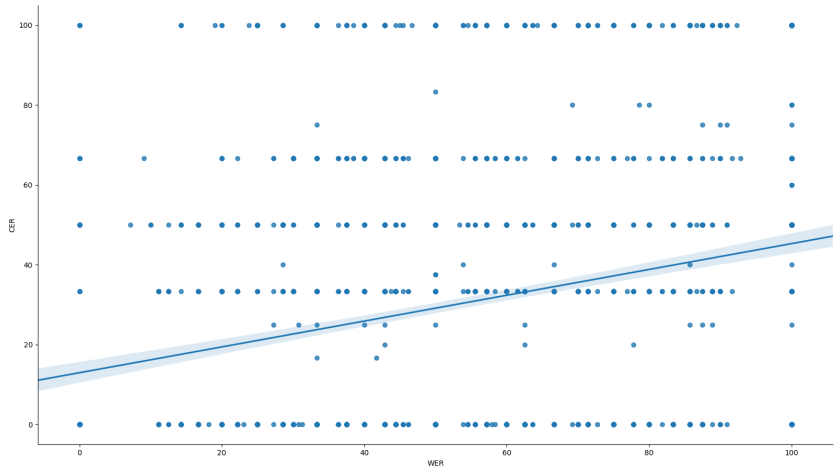


FIGURE 9.1 – Diagramme de dispersion montrant la corrélation entre les valeurs de WER et CER en sortie de ESPnet

Data(2) → Data(3)). Le tableau 9.8 et les figures 9.1 et 9.2, montrent des corrélations significatives qui ne sont pas négligeables mais qui ne font pas ressortir un motif clair. Il semble donc qu'une amélioration sur une tâche de RAP peut provoquer un gain sur la tâche de SLU mais que les performances de RAP ne sont pas suffisamment prédictives des performances SLU.

À l'inverse de l'approche SLU séquentielle, l'approche E2E a accès à des informations acoustiques et prosodiques pour la prédiction d'intentions et de concepts. Dans les sections suivantes, nous montrons quels éléments acoustiques influencent les performances de SLU E2E.

9.2 Impact des caractéristiques acoustiques et prosodiques sur la SLU E2E

Dans cette section nous montrons quels aspects prosodiques et acoustiques jouent un rôle dans le processus de SLU E2E en se concentrant sur la prédiction des concepts car leurs performances sont compétitives pour les modèles séquentiels et E2E de SLU.

TABLE 9.8 – Corrélations de Pearson et de Spearman entre les résultats de la RAP (WER) et de la SLU E2E (CER)

Modèle	**WER** (%)	**CER** (%)
RAP E2E Réal.+ML+Artif.	**46.50**	-
SLU E2E Data(2) → Data(3)	-	**32.12**
Coefficients de corrélation entre les 2 modèles :		
Pearson (r)	0.26^{**}	
Spearman (r_s)	0.25^{**}	

* signifie $p < 0.05$; ** signifie $p < 0.01$

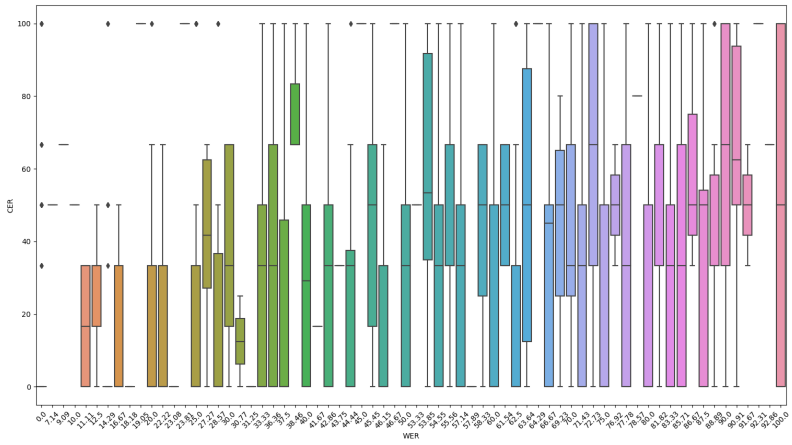


FIGURE 9.2 – Corrélations entre les valeurs de WER et de CER en sortie de ESPnet (boîte à moustache)

Goldwater et coll. (2010) étudient les caractéristiques prosodiques qui sont liées à une augmentation du taux d'erreur de RAP. Le minimum, le maximum, la moyenne et la plage de valeurs de *pitch* (hauteur en Français) et d'énergie par mot ont été calculées. Ils ont conclu que les moyennes de *pitch* et d'intensité ont relativement peu d'effet, sauf à des valeurs extrêmes. Le pitch est lié à la fréquence, mais les deux ne sont pas équivalents. La fréquence est un attribut objectif qui peut être mesuré. Le pitch par contre est la perception subjective de chaque personne d'une onde sonore, qui ne peut pas être mesurée directement. Néanmoins, deux tons ont généralement le même pitch s'ils partagent le même F0. Concernant la parole, les contours de pitch montant et descendant, aident à définir la prosodie (Plack et Oxenham, 2005).

Selon l'étude de Goldwater et coll. (2010), une plus grande plage de valeurs de *pitch* ou d'intensité conduit également à un WER plus faible. Selon les études de Stehwien et Vu (2016) et Su et Tseng (2018), la plupart des termes liés à des concepts portent également un accent de *pitch*. De telles informations prosodiques peuvent pointer vers les informations sémantiques les plus importantes ou tout du moins sur les frontières de l'information prégnante. Afin de vérifier cela, ils étudient la corrélation entre les variations de *pitch* et les mots avec des étiquettes de concept. Ce sont ces études qui ont inspiré notre analyse des relations entre les valeurs de *pitch* et d'énergie d'une part, et les performances de SLU d'autre part.

9.2.1 Corrélations entre RAP, pitch et énergie

Dans un premier temps, nous avons calculé le WER moyen par transcription d'hypothèse de sortie des modèles entraînés Kaldi (Chapitre 7, *DNN*, tableau 7.4) et ESPnet (Chapitre 8, *Réal.+ML+Artif.*, tableau 8.2), sur le corpus de test VocADom@A4H. Pour les mêmes données nous avons également calculé la moyenne des valeurs de F0 et de l'énergie par énoncé en utilisant l'outil *Praat*[1]. Nous avons calculé les valeurs F0 avec les paramètres de Praat par

1. https://www.fon.hum.uva.nl/praat/

défaut, entre 75 et 600Hz, contenant typiquement les fréquences de locuteurs masculins et féminins. Également nous avons calculé les valeurs F0 sans filtres. D'une telle façon, les valeurs de bruit de fond sont également incluses, comme par exemple les hautes fréquences de l'aspirateur.

Par la suite, nous avons calculé les corrélations entre les valeurs de F0 et d'énergie moyenne par énoncé d'une part, et le WER par énoncé de l'ensemble de test VocADom@A4H d'autre part. Plus précisément, cela signifie le calcul des corrélations entre :

- Valeurs de WER KALDI - valeurs d'énergie/pitch sur :
  - ensemble de test complet
  - commandes vocales de l'ensemble de test
- Valeurs de WER ESPnet - valeurs d'énergie/pitch sur :
  - ensemble de test complet
  - commandes vocales de l'ensemble de test

La cible de cette étude est l'extraction de concepts et d'intentions. Par conséquent, nous avons calculé les corrélations sur l'ensemble de test complet, ainsi que pour 2612 énoncés contenant une commande vocale.

La table 9.9 nous montre que plus *l'énergie* est élevée meilleures sont les performances de RAP pour KALDI et ESPnet sur l'ensemble de test VocADom@A4H complet (*Complet*), ce qui est moins le cas lorsqu'on considère uniquement les énoncés de commandes vocales (*Concept*). Par ailleurs, cette même table montre que la RAP et le *pitch* sont plus corrélés (Pearson) pour les énoncés avec des commandes vocales, notamment pour les transcriptions d'hypothèse de Kaldi par rapport à ESPnet. C'est surtout le cas pour les valeurs de pitch sans filtre. L'annexe B contient des diagrammes de dispersion et des boîtes à moustache pour les résultats du tableau 9.9 (*Complet, Concept : M&F*). Les figures B.4 et B.10 de cette annexe, montrent clairement que les plus hautes valeurs de WER correspondent aux plus hautes valeurs de pitch pour les performances de RAP Kaldi, ce qui est moins le cas pour les performances de ESPnet.

9.2.2 Impact du pitch sur la prédiction de concepts

La section précédente a montré que les corrélations Pearson entre les valeurs de WER et F0 sont légèrement plus élevées pour Kaldi par rapport à ESPnet pour les énoncés *comprenant* des commandes vocales. Cela semble indiquer qu'ESPnet est moins dépendant des variations de pitch que Kaldi. Cela peut également suggérer, comme les études sur les informations prosodiques Stehwien et Vu (2016) et Su et Tseng (2018) le mentionnent, que les informations de pitch peuvent pointer vers les informations sémantiques les plus importantes d'un énoncé. Afin de vérifier cela, nous avons effectué les étapes suivantes :

1. Pour les énoncés du corpus VocADom@A4H, nous avons calculé avec Praat les valeurs de F0 toutes les 0.01 secondes en les bornant entre 75 et 600Hz.

TABLE 9.9 – Corrélations entre WER et Pitch/Énergie pour les systèmes Kaldi et ESPnet

Corrélation pitch/énergie-RAP	**Pitch sans filtre**		**Pitch - 75-600Hz filtre**		**Énergie**	
	Kaldi	**ESPnet**	**Kaldi**	**ESPnet**	**Kaldi**	**ESPnet**
Complet :						
Pearson (r)	-0.02	0.003	0.05**	0.007	-0.22**	-0.07**
Spearman (r_s)	-0.07**	-0.03*	0.01	0.002	-0.14**	-0.08**
Concept : **M&F**						
Pearson (r)	0.23**	0.18**	0.09**	0.06*	0.04*	0.04*
Spearman (r_s)	0.10**	0.19**	0.06**	0.08**	0.05*	0.06**
M						
Pearson (r)	0.18**	0.20**	0.18**	0.16**	-0.07**	-0.03
Spearman (r_s)	0.10**	0.23**	0.02**	0.17**	-0.04	0.0002
F						
Pearson (r)	0.33**	0.14**	0.01	-0.08	0.26**	0.19**
Spearman (r_s)	0.11**	0.12**	-0.06	-0.06*	0.22**	0.18**

* signifie $p < 0.05$; ** signifie $p < 0.01$

2. Les horodatages pour les frontières de mots des transcriptions de référence et d'hypothèse sont générés avec des étiquettes de concepts symboliques. Ils sont calculés à l'aide de scripts Kaldi qui appliquent un *alignement forcé* (Forced Alignment) et génèrent un fichier CTM (Time-Marked Conversation file).
3. Les horodatages de valeurs F0 d'une part et les horodatages des frontières de mots sont alignés (figure 9.3). Les mots et les concepts avec les valeurs de F0 les plus élevées par énoncé sont sélectionnés.
4. Nous vérifions également si un concept associé à une forte valeur de F0 est plus facilement compris correctement.

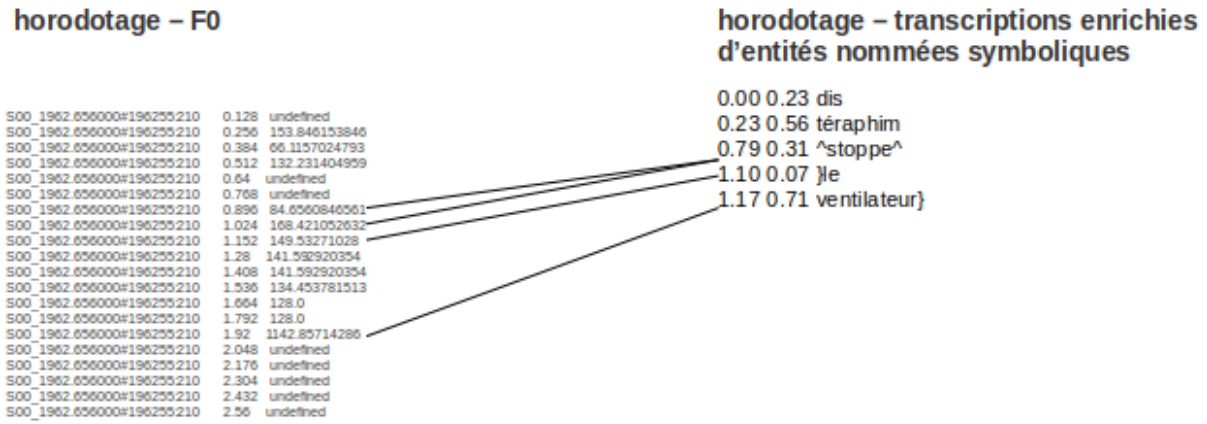


FIGURE 9.3 – Alignement entre horodotages, valeurs F0 et entitités nommées

La liste de fréquences du tableau 9.10 montre que 3 étiquettes de concept (`device`, `action`, `location-room`) figurent parmi les 10 termes les plus fréquents avec la valeur F0 la plus élevée par énoncé. La figure 9.4 montre que les valeurs de F0 (ligne colorée en bleu) sont les plus élevées pour les concepts `action` (^allume^), `device` ({bouilloire}) et

le mot-clé (vocadom). 47.79% de toutes les commandes vocales du corpus VocADom@A4H (1222 énoncés des 2557 commandes vocales) contiennent un concept contenant des mots avec les plus hautes valeurs F0 de l'énoncé. Il s'avère alors que les locuteurs, en faisant plus d'effort en prononçant les concepts et les mots-clés, parlent avec une intonation élevée, ce qui résulte en plus grandes valeurs de F0 pour les mots faisant partie des concepts et des mot-clés.

TABLE 9.10 – Fréquence de termes et concepts associés ayant la F0 la plus élevée par énoncé

Fréquence	**Mot**	**Fréquence**	**Mot**
538	**} (device)**	62	vocadom
509	**^ (action)**	59	est-ce
163	cirrus	51	hé
160	dis	47	hestia
131	ulysse	43	chanticou
131	**> (location-room)**	39	allô
105	téraphim	37	que
84	ichéfix	35	messire
72	minouche	32	, (device-setting)

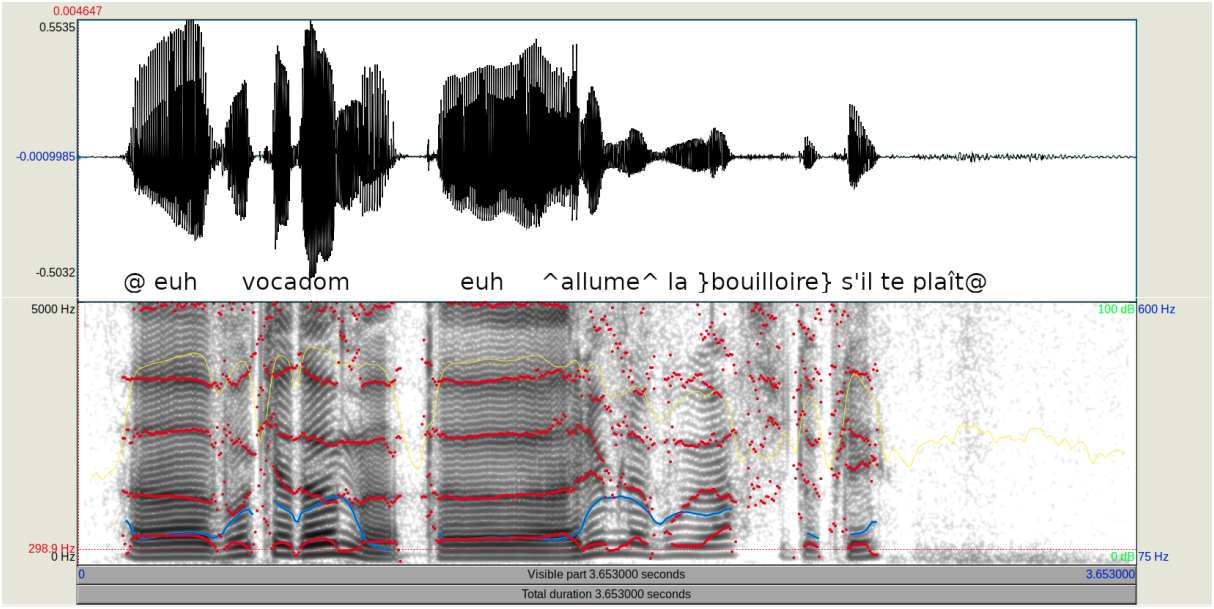


FIGURE 9.4 – Affichage en bleu des contours de pitch sur le spectrogramme de la commande vocale « vocadom allume la bouilloire s'il te plaît »

Nous n'avons pas seulement vérifié si le modèle E2E SLU est plus robuste aux variations de pitch que le modèle de référence, mais aussi si la prise en compte des informations contenues dans le pitch pouvaient conduire à une amélioration des performances. Pour les 1222 énoncés (du corpus de référence VocADom@4H) ayant pour les concepts la valeur de pitch la plus élevée, nous avons déterminé si ce concept était bien celui retenu dans l'hypothèse de compréhension , ceci pour chacun des 2 systèmes étudiés. Comme le montre le tableau 9.11 (*concept réf. dans hyp.*), lorsque les concepts sont prononcés avec un pitch plus élevé, le modèle de bout en bout (E2E) est plus performant que le modèle séquentiel.

Cela indique que la SLU E2E est légèrement moins impactée par les effets de pitch que la SLU séquentielle, conformément aux résultats de corrélation du tableau 9.9. Les résultats du tableau 9.11 indiquent également que les concepts avec des valeurs de F0 élevées, contribuent à de meilleures prédictions de SLU dans une approche SLU E2E.

TABLE 9.11 – Proportion de bonne compréhension des concepts lorsque ils sont prononcés avec une valeur de F0 plus élevé dans le corpus VocADom@4H.

Modèle de SLU	**concept réf. dans hyp.(%)**
Séquentiel	74.22
E2E	**75.00**

9.2.3 Impact du bruit de fond sur la prédiction de concepts

L'ensemble de test VocADom@A4H contient également des énoncés affectés par du *bruit de fond*, surtout ceux de la phase 3 (chapitre 6, section 6.2.1). Les commandes vocales lues sont enregistrées en présence d'un bruit de fond tel que l'aspirateur, la radio, la télévision, etc. Certains de ces bruits ont une fréquence très élevée, comme par exemple l'aspirateur. Cependant, comme au moment de l'étude ces énoncés n'avaient pas été annotés, nous avons tiré des énoncés au hasard que nous avons annotés d'étiquettes de bruit de fond jusqu'à ce que nous disposions d'environ 10% d'énoncés correspondant à ce cas, soit 204 énoncés. Pour ces énoncés, la table 9.12 montre les corrélations existantes entre le WER de la RAP d'une part, et l'énergie et le pitch d'autre part. Ces corrélations sont significativement plus élevées pour Kaldi que pour ESPnet, comme cela apparaît à l'annexe B en figures B.5, B.6, B.11, B.12, B.17 et B.18.

TABLE 9.12 – Corrélations entre WER et Pitch/Énergie pour les systèmes Kaldi et ESPnet pour les commandes vocales affectées par un bruit de fond

Corrélation	**Pitch sans filtre**		**Pitch - 75-600Hz filtre**		**Énergie**	
Pitch/Énergie-RAP	**Kaldi**	**ESPnet**	**Kaldi**	**ESPnet**	**Kaldi**	**ESPnet**
concept - bruit de fond : M&F						
Pearson (r)	0.51^{**}	0.28^{**}	0.39^{**}	0.22^{**}	0.21^{**}	-0.02
Spearman (r_s)	0.50^{**}	0.23^{**}	0.41^{**}	0.23^{**}	0.16^{*}	-0.01
M						
Pearson (r)	0.51^{**}	0.35^{**}	0.54^{**}	0.18^{*}	0.21^{**}	-0.05
Spearman (r_s)	0.52^{**}	0.32^{**}	0.36^{**}	0.19^{*}	0.14	-0.04
F						
Pearson (r)	0.53^{**}	-0.04	-0.21	-0.24^{*}	0.08	-0.09
Spearman (r_s)	0.56^{**}	0.04	-0.26^{*}	-0.25	0.03	-0.11

* signifie $p < 0.05$; ** signifie $p < 0.01$

La table 9.13 montre que les performances de SLU E2E (*Bruits de fond - Tous*, CER) surpassent celles de la SLU séquentielle. Elles sont notamment meilleures pour les énoncés avec le bruit de fond d'un aspirateur (*Aspirateur*) qui a également un *pitch élevé*. Bien que

les performances de RAP Kaldi soient en général meilleures par rapport à ESPnet, les performances des deux modèles sont plus proches pour les énoncés prononcés en présence du bruit de fond d'un aspirateur. Ces résultats confirment des corrélations ESPnet plus faibles entre le WER d'une part et le pitch et l'énergie d'autre part, par rapport à Kaldi, comme nous l'avons décrit dans les sections précédentes. Ils valent pour les locuteurs féminins (*F*) ainsi que pour les locuteurs masculins (*M*). Ces résultats correspondent également aux résultats de l'étude de (Qian et coll., 2016) selon lesquels un système de RAP E2E affiche de bonnes performances et une meilleure robustesse en traitant de la parole affectée par du bruit, en intégrant des réseaux de neurones convolutifs (CNN - *Convolutional Neural Network*) dans leur architecture (chapitre 3, section 3.1.4.2.3).

TABLE 9.13 – Performances des systèmes de RAP et de SLU pour des commandes vocales du corpus VocADom@A4H prononcées en présence de bruit de fond (test sur 204 commandes)

Type de bruit de fond (M : homme - F : femme)	**SLU séquentielle**		**SLU E2E**		**# Fréq. énoncés**
	WER (%)	**CER (%)**	**WER (%)**	**CER (%)**	
Bruits de fond – Tous :					
M&F	38.58	57.80	57.53	**39.73**	204
M	30.78	54.98	52.74	34.05	152
F	58.72	65.06	69.87	54.38	52
Aspirateur :					
M&F	57.00	77.62	59.00	53.79	108
M	46.64	75.46	54.23	47.82	72
F	77.75	81.94	71.31	65.74	36
Radio& TV :					
M&F	20.31	35.77	56.31	25.94	75
M	18.08	36.64	53.30	20.90	58
F	27.96	32.84	66.57	43.13	17
Ventilateur :					
M&F	14.74	32.69	65.36	26.92	21
M	13.18	24.99	62.84	38.88	17
F	18.27	49.99	71.03	0	4

9.2.4 Délexicalisation et suppression des variations de F0

Le but de la délexicalisation est de rendre le contenu lexical d'un énoncé inintelligible. Ceci est réalisé en supprimant les caractéristiques segmentales des phonèmes et les mots du signal de parole, tout en préservant les caractéristiques suprasegmentales telles que le pitch (Kain et Santen, 2010). La délexicalisation est une autre méthode pour vérifier l'impact du pitch que l'on a appliqué à l'ensemble de test VocaDom@A4H, en utilisant des filtres passe-bande, de 2, 3 et 4 kHz. L'application des filtres a considérablement détérioré les résultats de RAP pour les modèles Kaldi et ESPnet. C'est pourquoi nous n'avons pas poursuivi l'expérience. On peut cependant renvoyer le lecteur au tableau de résultats synthétiques 9.23 qui indique que ESPNet est bien plus sensible au filtrage. Ceci tendrait a suggérer que ESPnet utilise des paramètres large bande pour la tâche de RAP alors que Kaldi modélise plus

étroitement les bandes de fréquences de chaque phonème.

Une autre façon de vérifier l'impact du pitch sur le décodage de RAP et les performances de SLU consiste à supprimer la variation de pitch de l'ensemble de test. À cette fin, nous avons calculé la valeur de F0 moyenne par locuteur. Ces moyennes sont incluses dans le tableau 9.14 pour les 4 locuteurs féminins (F) et les 7 locuteurs masculins (M). En utilisant Praat, tous les énoncés de test de chaque locuteur ont été synthétisés en choisissant comme valeur de F0 la moyenne des valeurs de F0 produites par chaque locuteur.

TABLE 9.14 – Valeur moyenne de F0 pour chacun des 11 locuteurs du corpus VocADom@4H

Locuteurs	**F0 Moyenne**
S00(M)	152.62
S01(M)	167.40
S02(M)	165.08
S03(M)	156.34
S04(F)	232.36
S05(F)	249.89
S06(M)	173.22
S07(M)	168.07
S08(M)	155.72
S09(F)	243.64
S10(F)	277.09

Un modèle de RAP E2E (ESPnet) a également été entraîné, en utilisant des paramètres *MFCC*, au lieu des paramètres *fbank*, pour comparer les performances d'ESPnet avec les mêmes paramètres acoustiques que ceux utilisés pour Kaldi. Les mêmes données d'apprentissage ont été utilisées que pour le modèle de RAP ESPnet *Réal.+ML+Artif* de la table 8.2.

La table 9.15 montre que :

— au niveau du module de RAP de la SLU séquentielle (**RAP SLU séq**), les performances de Kaldi (MFCC) (*Kaldi RAP DNN*, tableau 7.4) sur des données *sans* variation de pitch sont supérieures aux performances sur des données *avec* variation de pitch ;

— en revanche, pour la **RAP E2E**, les performances d'ESPnet sur des données *avec* variation de pitch (*ESPnet RAP Réal+ML+Artif.*, tableau 8.2) sont supérieures aux données *sans* variation de pitch, notamment avec les paramètres fbank ;

— Également, pour la **SLU E2E**, au niveau de la prédiction de concepts, les performances d'ESPnet sur des données *avec* variation de pitch (*ESPnetData(2) → Data(3)*, tableau 8.6) sont supérieures aux données *sans* variation de pitch.

1. Ensuite les énoncés des locuteurs masculins (**M(1)**) et féminins (**F(1)**) ont été évalués séparément. Ces énoncés ont été comparés avec ceux des locuteurs masculins (**M(2)**) et féminins (**F(2)**), mais *seulement* avec un pitch *au-dessus* de la moyenne par locuteur.

2. Ces résultats montrent que les performances (**M(2)**) et (**F(2)**) avec *suppression* des variations de pitch sont nettement pires qu'avec variations de pitch. Ce décalage

(**Pitch Diff.**) entre le CER d'un modèle *sans* variation de pitch et *avec* variation de pitch est le plus élevé pour (*M(2)*) (10.92).

TABLE 9.15 – Performances de la RAP et de la SLU après suppression des variations de pitch dans le corpus VocADom@A4H

Modèle	**Param. acoust.**	**Sans var. pitch**		**Avec var. pitch**		**Pitch Diff.**
		WER (%)	**CER (%)**	**WER (%)**	**CER (%)**	
RAP SLU séq :	MFCC	21.48	-	22.92	-	-1.44
RAP E2E :	fbank	50.20	-	46.50	-	3.7
	MFCC	49.90	-	47.60	-	2.3
SLU E2E :	fbank	-	40.02	-	32.12	7.9
M(1).	fbank	-	41.94	-	32.90	9.04
F(1).	fbank	-	36.58	-	30.74	5.84
F0 > F0 moyen loc.						
M(2).	fbank	-	53.32	-	42.40	**10.92**
F(2).	fbank	-	37.89	-	32.36	5.53

De ces résultats, il s'avère que les paramètres *MFCC* sont plus robustes aux énoncés *sans* variation de pitch que les paramètres *fbank*. Ceci est davantage le cas pour les énoncés avec un pitch *au-dessus* de la moyenne par locuteur. Cela indique que par rapport aux paramètres fbank, les paramètres MFCC utilisés avec un modèle de réseaux de neurones profonds E2E réduisent les performances pour les énoncés avec un pitch élevé dû à leur représentation compressée.

Cette hypothèse est en résonance avec les résultats de l'étude de (Abdel-Hamid et coll., 2012) selon laquelle l'intégration de CNN dans un système de RAP E2E, en combinaison avec des paramètres fbank, contribue à l'amélioration de ses performances (chapitre 3, section 3.1.4.2.3). Contrairement aux paramètres MFCC, aucune transformée en cosinus discrète (DCT - Discrete Cosinus Transform) n'a été appliquée aux paramètres fbank de sorte que ceux-ci ne sont pas décorrélés et que les réseaux de neurones exploitent cette redondance.

9.3 Analyse des performances de la RAP et de la SLU au niveau symbolique

Après avoir effectué des analyses au niveau *acoustique*, nous présentons ensuite une analyse au niveau *symbolique*. Nous mesurons l'impact des mots hors vocabulaire (OOV – *out of vocabulary*) de l'ensemble de test VocADom@A4H, sur les performances de SLU séquentielle et de SLU E2E. Nous mesurons également la robustesse de ces modèles sur des données de test contenant une *variabilité syntaxique* augmentée.

9.3.1 Mots hors vocabulaire (OOV)

Pour mesurer l'impact d'un taux de mots hors vocabulaire augmenté, nous avons progressivement remplacé des mots par des synonymes n'apparaissant pas dans le vocabulaire

des données d'apprentissage comme cela est présenté en annexe B à la table B.1. Ceci a été fait en 4 étapes, selon le concept auquel les mots se rattachent :

— Étape 1 : `action` et `device-setting`

— Étape 2 : Étape 1 et `device`

— Étape 3 : Étape 2 et `location`

— Étape 4 : Étape 3 et `mots-clés`

L'exemple ci-dessous montre une commande vocale extraite du corpus de test VocADom@A4H contenant une intention et des concepts symboliques *avant* (1) et *après* (2) substitution par des mots hors vocabulaire (étape 4) :

(1) @ ah vocadom euh ^allume^ }la bouilloire} @
(2) @ ah ursule euh ^enclenche^ }la bouillotte} @

La table 9.16 montre que les mots substitués à l'étape 4 représentent 26,15% du nombre total de mots et 3,48 % du nombre total de types de mots.

TABLE 9.16 – Modification du corpus Vocadom@A4H pour l'analyse des performances au niveau symbolique : nombre de mots hors vocabulaire par rapport au nombre total de mots

Substitutions	**#Type mots**	**#Mots**	**(%) Type Mots**	**(%) Total Mots**
Étape 1	22	1785	1.50	5.72
Étape 2	34	4276	2.32	13.70
Étape 3	41	5516	2.80	17.68
Étape 4	51	8160	3.48	26.15
Total	1462	31k	-	-

Les phrases générées de chaque étape ont alimenté un synthetiseur vocal en utilisant le même outil de synthèse vocale que celui que nous avons utilisé pour la génération du corpus artificiel (chapitre 6, section 6.3.4). Les données de test artificielles qui en résultent sont les énoncés d'entrée des outils de SLU séquentielle et E2E. Cependant, la partie acoustique des données d'apprentissage du modèle de SLU E2E, contient de la parole artificielle, ce qui n'est pas le cas pour le module de RAP de l'approche SLU séquentielle. Par conséquent, nous avons utilisé les transcriptions d'hypothèse de RAP E2E (ESPnet) comme transcriptions d'entrée de l'outil de SLU séquentielle pour effectuer une comparaison équitable des 2 approches SLU.

Pour les 2 approches SLU, les tables 9.17 et 9.18 montrent que les performances de prédiction de concepts (*concept CER*) et d'intention (*Intent. F-mesure*) se détériorent en cas d'augmentation des taux d'OOV. D'une manière générale, les décalages (*Diff.*) entre les performances de prédiction de concepts (*concept*) et d'intentions (*Intent.*) pour *Complet synth.* d'une part et *Étape 4* d'autre part, sont plus petits pour le modèle E2E que pour le modèle de SLU séquentielle. Cela indique une plus grande robustesse du modèle E2E pour faire face à une augmentation du taux de mots hors vocabulaire.

TABLE 9.17 – Impact des mots hors vocabulaire et de la variation syntaxique sur les performances de la SLU séquentielle testée sur le corpus VocADom@4H

Modèle	**Réf. (=NLU)**		**Hyp. (=RAP+NLU)**	
	Concept CER (%)	**Intent. F-mesure (%)**	**Concept CER (%)**	**Intent. F-mesure (%)**
Complet réel	33.78	85.51	36.24	84.21
Complet synth.	-	-	37.07	83.34
Hors vocab :				
Étape 1	37.75	81.50	45.43	79.56
Étape 2	53.77	72.39	62.03	72.48
Étape 3	63.01	69.58	68.07	70.29
Étape 4	90.45	63.66	86.44	65.03
Diff.	56.67	21.85	49.37	18.31
Var. syntax. :				
Étape 1	38.41	81.06	50.40	77.45
Étape 2	38.34	81.19	52.75	76.36
Diff2.	4.56	4.32	15.68	6.98

Pour la SLU séquentielle, dans la plupart des cas, les performances sur des données textuelles manuellement annotées (tableau 9.17, *Réf.*) dépassent largement les performances sur les transcriptions de RAP d'hypothèse (*Hyp.*). Cependant, plus le taux d'OOV est élevé, plus ce décalage diminue.

Pour le modèle E2E, les performances de prédiction d'intentions diminuent considérablement pour *Complet synth.* par rapport à *Complet réel*. Ceci est particulièrement dû à un taux d'erreurs augmenté pour les intentions `None` qui consistent en parole réelle dans les données d'entraînement, alors que nous avons utilisé des données d'évaluation synthétiques pour l'évaluation de l'impact des mots hors vocabulaire.

TABLE 9.18 – Impact des mots hors vocabulaire et de la variation syntaxique sur performances de la SLU E2E testée sur le corpus VocADom@A4H

Tâche	**RAP WER (%)**	**Concept CER (%)**	**Intention F-mesure (%)**
Complet réel	46.50	32.12	74.57
Complet synthétique	39.30	25.00	53.70
Hors vocabulaire :			
Étape 1	44.00	30.75	50.39
Étape 2	53.20	46.75	50.26
Étape 3	52.50	50.89	51.59
Étape 4	55.90	58.80	51.43
Diff.	16.6	33.8	2.27
Var. syntaxique :			
Étape 1	44.40	16.29	52.59
Étape 2	50.90	22.07	49.09
Diff2.	11.60	2.93	4.61

9.3.2 Variation syntaxique

Nos utilisateurs cibles sont des personnes âgées qui ont tendance à s'écarter d'un ensemble prédéfini de commandes vocales. Nous avons pris en compte la variabilité syntaxique de la parole de nos utilisateurs cibles. Dans cette section, nous mesurons la robustesse des modèles de SLU séquentielle et E2E, en prédisant des concepts et des intentions sur des données d'évaluation avec une variabilité syntaxique progressive en deux étapes,

— Étape 1 : nous avons substitué 32 verbes faisant partie des concepts `action` par des constructions syntaxiques plus complexes (Annexe B, tableau B.2)

— Étape 2 : Les substitutions d'Étape 1 ont été augmentées de disfluences qui entourent les termes de 18 concepts étiquetés de `device` (Annexe B, tableau B.3)

L'exemple suivant montre une commande vocale contenant une intention et des concepts symboliques du corpus de test VocADom@A4H *avant* (1) et *après* (2) insertion de constructions syntaxiques plus complexes et de disfluences (étape 2) :

(1) @ ah vocadom euh ^allume^ }la bouilloire} @
(2) @ ah vocadom euh est–ce que tu pourrais ^allumer^ bou la }bouilloire} @

Nous avons également généré de la synthèse vocale basée sur les ensembles de test modifiés qui en résultent, et nous les avons évalués de la même façon que pour les mots hors vocabulaire comme expliqué dans la section précédente.

Pour les deux approches SLU, les tableaux 9.17 et 9.18 montrent que les décalages (*Diff2.*) entre les performances de prédiction de concepts et d'intentions (Intent.) pour *Complet synth.* d'une part et *Var. syntax., Étape 2* d'autre part, sont plus petits pour le modèle E2E que pour le modèle de SLU séquentielle. Cela indique de nouveau une plus grande robustesse du modèle E2E pour faire face à une variation syntaxique augmentée.

Le tableau 9.18 montre également, que les performances du modèle E2E pour la prédiction des concepts, s'améliorent avec une syntaxe plus complexe. Ceci peut être dû à une longueur de phrase moyenne de 15 mots pour les énoncés du corpus artificiel, tandis que la longueur de phrase moyenne pour les énoncés d'évaluation (d'origine) n'est que de 5. La variation syntaxique accrue, augmente également la longueur des énoncés d'évaluation qui s'approche par conséquent de la longueur moyenne des énoncés du corpus artificiel.

9.4 Analyse d'erreurs de RAP et de SLU spécifiques aux locuteurs

Pour compléter nos analyses, nous nous intéressons maintenant à l'effet du locuteur sur les performances de la SLU. Nous comparons également les performances entre les locuteurs féminins et masculins. Le tableau 9.19 présente les résultats de la RAP et de la SLU pour les approches séquentielle et de bout en bout en fonction des locuteurs. Nous commenterons différents aspects de ces résultats par la suite.

9.4.1 Phrases monosyllabiques

Lorsque l'on compare les prédictions de RAP Kaldi, ESPnet, les prédictions de SLU séquentielle (concepts) et E2E par locuteur, il s'avère que le participant S06(M) appartient aux locuteurs montrant des performances aberrantes et entre les plus faibles pour presque tous les modèles, à l'exception de la SLU séquentielle (tableau 9.19). Il s'avère que la fréquence moyenne des énoncés monosyllabiques tels que 'ah', 'bon', 'euh', 'cool', 'hum', 'non', 'ok', 'oui', 'ouais' etc., est plus élevée pour le locuteur S06(M) que pour la plupart des autres locuteurs. Le calcul du WER (Kaldi) sur tous les énoncés monosyllabiques par locuteur montre que ceux-ci sont les plus élevés pour le locuteur S06(M) (tableau 9.20).

TABLE 9.19 – Erreurs de la RAP et de la SLU spécifiques pour les locuteurs du corpus VocADom@A4H

Locuteurs	**RAP Kaldi WER (%)**	**RAP ESPnet WER (%)**	**SLU séquentielle CER (%)**	**SLU E2E CER (%)**
S00(M)	12.24	33.04	**55.76**	16.25
S01(M)	23.07	51.16	32.72	43.77
S02(M)	17.06	41.67	45.20	37.29
S03(M)	18.30	38.11	29.13	19.66
S04(F)	24.13	43.37	31.61	23.50
S05(F)	35.06	52.40	28.34	38.07
S06(M)	**34.31**	**59.00**	30.11	**47.27**
S07(M)	18.43	47.40	35.65	39.32
S08(M)	39.78	48.23	31.20	38.22
S09(F)	27.02	44.42	30.29	31.85
S10(F)	18.33	36.01	31.67	29.88

TABLE 9.20 – Reconnaissance des énoncés monosyllabiques pour les différents locuteurs du corpus VocADom@A4H par le système Kaldi

Locuteurs	**#Énoncés**	**#Énoncés monosyll.**	**Kaldi WER (%) #Énoncés monosyll.**
S00(M)	558	22	18.18
S01(M)	546	45	13.8
S02(M)	734	48	12.5
S03(M)	425	11	63.63
S04(F)	510	42	26.19
S05(F)	594	68	55.88
S06(M)	813	72	**75**
S07(M)	498	16	25
S08(M)	798	125	51.2
S09(F)	647	45	49.25
S10(F)	624	60	24.44

9.4.2 Articles définis et indéfinis

Le locuteur S00(M) fait partie des locuteurs montrant les pires performances pour la SLU séquentielle (tableau 9.19). L'analyse de ce locuteur montre un style de parole moins grammatical. Les articles définis et indéfinis sont souvent omis, comme par exemple dans la commande vocale "cirrus fermer porte" où l'article défini "la" a été omis. La moyenne de la fréquence d'articles définis/indéfinis par commande vocale pour ce locuteur dans le tableau 9.21, est inférieur à celle des autres locuteurs. Néanmoins, les performances de SLU E2E pour ce locuteur sont les meilleures. Il semble donc que le modèle E2E est plus robuste aux énoncés d'un style de parole moins grammatical.

TABLE 9.21 – Présence d'articles définis et indéfinis par commande vocale dans le corpus VocADom@A4H

Locuteurs	**#Art.**	**#Commande**	**#Art. par commande**
S00(M)	148	242	**0.61**
S01(M)	200	202	0.99
S02(M)	236	236	1
S03(M)	220	220	1
S04(F)	222	222	1
S05(F)	202	206	0.98
S06(M)	226	239	0.94
S07(M)	233	238	0.97
S08(M)	314	319	0.98
S09(F)	248	250	0.99
S10(F)	231	238	0.97

9.4.3 Locuteurs masculins et féminins

Dans le chapitre 6, section 6.3.4.2 nous avons montré que les distances entre les échantillons de la synthèse vocale féminine et les échantillons réels feminins sont plus faibles que les distances entre les échantillons de la parole réelle masculine et la synthèse vocale (féminine), comme la parole artificielle est générée pour une voix française *féminine*. Le tableau 9.22 montre que, par conséquent, les performances de RAP E2E et de SLU E2E, pour les locuteurs féminins surpassent celles des locuteurs masculins comme les modèles acoustiques de RAP et de SLU E2E contiennent de la parole artificielle féminine. Par contre le module de RAP de l'approche SLU séquentielle ne contient *que* de la parole réelle. Les différences entre les performances des locuteurs masculins et féminins sont également moindres par rapport aux performances de RAP et de NLU du modèle de SLU séquentielle.

9.5 Conclusion

La table 9.23 en page 181 résume les résultats obtenus par la SLU discutés au cours de ce chapitre. La table 9.24 explique les abréviations utilisées dans le tableau 9.23. Nous avons

montré que les corrélations entre les performances de RAP et de SLU E2E sont plutôt faibles et que des performances de SLU E2E élevées ne supposent pas nécessairement des performances de RAP élevées. Des cas spécifiques qui confirment cette conclusion, montrent que les étiquettes de concept sont correctement prédites malgré les erreurs de prédiction de mots-clés. De plus, les étiquettes de concept peuvent également être correctement prédites même si la prédiction de leur valeur par la RAP est erronée.

Nous pouvons confirmer l'une des conclusions de l'étude de Stehwien et Vu (2016) et Su et Tseng (2018) à savoir que l'information prosodique peut pointer vers l'information la plus importante du point de vue sémantique. Les corrélations entre le WER et les valeurs de pitch pour les données contenant des concepts sont plus élevées pour le module de RAP de l'approche SLU de référence (Kaldi), que pour la RAP de bout en bout (ESPnet). Les valeurs de F0 les plus élevées se produisent particulièrement pour les concepts, car les locuteurs semblent faire plus d'efforts pour prononcer des informations importantes au niveau sémantique importantes pour la commande vocale. En comparant les concepts d'un pitch élevé des phrases de test de référence et d'hypothèse issues par les modèles de SLU séquentielle et E2E, nous avons également montré que ces valeurs de pitch plus élevées contribuaient à améliorer notamment les performances de SLU E2E.

ESPnet semble être en mesure de bien traiter la parole affectée par du bruit et fonctionne mieux que Kaldi dans ce cas. Les corrélations entre le WER d'une part et le pitch et l'énergie d'autre part sont significativement plus élevées avec Kaldi que pour ESPnet pour 10% de données de test annotées avec des étiquettes de parole bruitée. C'est notamment le cas pour les bruits de fond avec pitch élevé, comme celui d'un aspirateur. Nous avons comparé les performances en utilisant des paramètres MFCC et fbank avec ESPnet. Nous avons constaté que les paramètres fbank, utilisés avec ESPnet en intégrant des CNN, semblent plus adaptés au traitement de la parole bruitée que les paramètres MFCC utilisés par défaut avec Kaldi. Les paramètres MFCC ont contribué à de meilleures performances de RAP pour les données d'évaluation avec une variation de pitch supprimée (en particulier pour Kaldi), que les paramètres fbank. En revanche les performances d'ESPnet pour les données avec variations de pitch sont meilleures en utilisant les paramètres fbank. C'est notamment le cas pour les énoncés dont le pitch est plus élevé que la moyenne de pitch par locuteur. Ces résultats in-

TABLE 9.22 – Performances de la RAP et de la SLU selon le genre du locuteur (%)

Mesure	**SLU Séquentielle**	**SLU E2E**	**# Énoncés**
WER–complet	22.92	46.50	6747
Masculin	22.25	46.90	4372
Féminin	24.12	**44.00**	2375
F-mesure Intention–complet	84.21	74.57	6747
Masculin	83.64	74.30	4372
Féminin	86.48	**75.04**	2375
CER–complet	36.24	32.12	6747
Masculin	37.56	32.90	4372
Féminin	33.77	**30.71**	2375

diquent que **le modèle E2E exploite l'information prosodique ce qui favorise ses performances de SLU**.

L'impact du niveau acoustique se voit également à travers l'évaluation des performances de la RAP et de la SLU sur les énoncés des locuteurs masculins et féminins. Dans le chapitre 6, section 6.3.4.2 nous avions montré que les distances entre les échantillons de la synthèse vocale féminine et les échantillons réels féminins sont plus basses que les distances entre les échantillons de la parole réelle masculine et la synthèse vocale (féminine), ce qui est cohérent étant donné que la parole artificielle est générée pour une voix française féminine. Par conséquent, les performances de la RAP E2E et de la SLU E2E sur les locuteurs féminins surpassent celles des locuteurs masculins étant donné que les modèles acoustiques de RAP et de SLU E2E contiennent de la parole artificielle féminine.

Comme nos utilisateurs cibles sont des personnes âgées qui ont tendance à s'écarter d'un ensemble prédéfini de commandes, nous avons vérifié l'impact des mots hors vocabulaire et d'une variabilité syntaxique accrue. À cette fin, nous avons supprimé le vocabulaire des concepts et l'avons remplacé par des mots hors vocabulaire, absents des données d'apprentissage. Cela a été fait progressivement en 4 étapes jusqu'à ce que environ un quart du total de mots des données de test VocADom@A4H soit hors vocabulaire, ceci a montré que le modèle SLU E2E est plus robuste. Afin de mesurer l'impact de la variabilité syntaxique, nous avons substitué les verbes du concept `action` par des structures syntaxiques plus complexes. Dans un deuxième temps, nous avons inséré des disfluences autour du concept `device`. Le modèle SLU E2E montre là aussi plus de robustesse. Cependant, nous devons souligner que les données d'évaluation avec des mots hors vocabulaire et une complexité syntaxique accrue ont été générées par synthèse vocale. La parole synthétique fait également partie des données d'apprentissage de SLU E2E et non des données d'apprentissage de SLU séquentielle. Pour une comparaison équitable, nous avons utilisé des transcriptions d'hypothèse de RAP comme sorties de l'outil de RAP ESPnet comme phrases d'entrée du module de NLU de l'approche SLU séquentielle. En outre, le modèle SLU E2E montre des performances nettement meilleures par rapport au modèle de SLU séquentielle pour le traitement des énoncés peu grammaticaux. Cette évaluation suggère **qu'un modèle SLU E2E est plus robuste aux variations du vocabulaire et de la syntaxe que l'approche séquentielle**.

TABLE 9.23 – Résumé des résultats de l'évaluation des SLU séquentielle et de bout en bout sur le corpus VocADom@A4H

Niveau d'analyse :	**Acoustique**												**Symbolique**						
	Source							**Prosodie**					**Lexical**					**Syntaxique**	
Analyse :	**Bruit**				**Genre**			**F0 moyenne**		**Délexicalisation**			**OOV**					**Variation**	
Résultats :	**A**	**RT**	**V**	**T**	**M**	**F**	**M&F**	**MFCC**	**fbank**	**4kHz**	**3kHz**	**2kHz**	**0**	**1**	**2**	**3**	**4**	**1**	**2**
SLU séquentielle :																			
WER	57.00	20.31	14.74	38.58	22.25	24.12	22.92	21.48	-	24.64	28.11	44.28	-	-	-	-	-	-	-
Concept Error Rate	77.62	35.77	32.69	57.80	37.56	33.77	36.24	-	-	-	-	-	37.07	45.43	62.03	68.07	86.44	50.40	52.75
Intent. F-mesure	-	-	-	-	83.64	86.48	84.21	-	-	-	-	-	83.34	79.56	72.48	70.29	65.03	77.45	76.36
SLU E2E :																			
WER	59.00	56.31	65.36	57.53	46.90	44.00	46.50	49.90	50.20	62.50	73.10	94.10	39.30	44.00	53.20	52.50	55.90	44.40	50.90
Concept Error Rate	53.79	25.94	26.92	39.73	32.90	30.71	32.12	-	40.02	-	-	-	25.00	30.75	46.75	50.89	58.80	16.29	22.07
Intent. F-mesure	-	-	-	-	74.30	75.04	74.57	-	-	-	-	-	53.70	50.39	50.26	51.59	51.43	52.59	49.09

TABLE 9.24 – Signification des termes utilisés dans la table 9.23 résumant les performances de la SLU

Analyse		**Explication**
Bruit	A	Aspirateur
	RT	Radio-Télévision
	V	Ventilateur
	T	Tous les bruits de fond
Genre	M	Masculin
	F	Féminin
	M&F	Masculin et Féminin
F0 moyenne	mfcc, fbank	Moyenne de valeurs de F0 par locuteur, mfcc, fbank
Délexicalisation	4kHz, 3kHz, 2kHz	Filtre passe-bande 4kHz, 3kHz, 2kHz
Lexical OOV	0	Synthèse vocale, vocabulaire complet
	1, 2, 3, 4	Hors vocabulaire, Étapes 1, 2, 3, 4
Variation syntaxique	1, 2	Étapes 1, 2

CHAPITRE 10

Conclusion et perspectives

10.1 Conclusion

La compréhension automatique de la parole (SLU) est une tâche qui consiste à extraire automatiquement un sens d'un énoncé oral. C'est une fonction que l'on retrouve dans plusieurs objets du quotidien comme les assistants vocaux des smart phones ou les enceintes intelligentes (smart speakers) lorsqu'on leur demande les horaires d'un train ou de vols. Cependant, il s'avère que la longueur des énoncés des commandes vocales traitées par des enceintes intelligentes ne dépasse souvent pas une longueur de 5 mots (Bentley et coll., 2018).

En outre, dans le cadre du projet VocADom, dont l'objectif est de concevoir un système de commande vocale robuste, les utilisateurs cibles sont des personnes âgées. Il a été montré que cette population a tendance à s'écarter facilement d'une grammaire figée de commandes vocales (Möller et coll., 2008; Takahashi et coll., 2003; Vacher et coll., 2015). C'est cette incapacité des maisons intelligentes et des *smart speakers* à traiter des commandes plus complexes au niveau linguistique qui a motivé le développement d'un système de reconnaissance des *commandes vocales* spécifiques au domaine *domotique* qui tient compte de la *variation syntaxique et sémantique* de ses utilisateurs cibles.

Comme vu dans les chapitres de l'état de l'art, la compréhension automatique de la parole (SLU) a souvent été abordée en considérant deux sous-tâches séparées. Les transcriptions d'hypothèse, issues d'un module de reconnaissance automatique de la parole (RAP), sont les entrées d'un module de compréhension automatique du langage (NLU). Cette approche effectue alors la SLU en cascadant ces deux tâches dans un pipeline ce qui conduit à une approche séquentielle. Dans cette thèse nous avons cherché à **comprendre quels avantages une approche SLU de bout-en-bout (E2E) peut offrir par rapport à une approche en pipeline classique**. En effet, le problème principal des systèmes séquentiels de SLU est la dépendance à la qualité des transcriptions sorties du module de RAP. Les décalages entre les performances de leur modules de NLU et du système complet de SLU séquentielle restent souvent élevés, malgré les stratégies employées pour les réduire (chapitre 4). Cela nous a naturellement conduit à la question de savoir **si on peut éviter la cascade d'erreurs de l'approche SLU séquentielle ?**

Notre réponse est un modèle d'inférence qui extrait les intentions et concepts directement du signal par une approche de bout-en-bout (*End-to-End, E2E*). Cette approche basée sur des réseaux de neurones profonds nous a permis d'éviter la *cascade d'erreurs* grâce à l'ap-

prentissage conjoint de ces deux tâches dans un seul modèle. En comparant notre approche SLU E2E avec une approche séquentielle composée d'un système de RAP état de l'art et d'un module de NLU, appris sur les données spécifiques au domaine domotique, nous avons pu montrer (chapitre 8, section 8.5.2) que l'approche E2E donne de meilleures performances au niveau des concepts.

La reconnaissance de la parole est une tâche pivot pour la SLU et l'état actuel des connaissances indique que pour atteindre de bonnes performances SLU il est nécessaire d'obtenir de bonnes performances de RAP. Nous avons également comparé nos deux approches SLU en pipeline et E2E, sur une tâche unique de RAP apprise sur des données équivalentes. Nos expériences montrent que l'approche de RAP E2E obtient un taux d'erreurs de mot bien plus élevé que le module de RAP de l'approche séquentielle. Pourtant l'approche E2E montre de meilleures performances SLU pour la prédiction de concepts, en utilisant le même outil (ESPnet) que pour la RAP E2E. **On peut donc en conclure que le modèle E2E permet bien d'éviter la *cascade d'erreurs***. En outre les performances de RAP parfaites ne sont pas nécessaires pour obtenir de bonnes performances SLU E2E, alors qu'elles sont essentielles dans le cas d'une approche séquentielle, ce que nous avons démontré dans Desot et coll. (2019b) au niveau de la prédiction d'intentions et dans Desot et coll. (2019a) au niveau des concepts.

Par ailleurs, étant donné que l'approche E2E infère des concepts et des intentions transportés par un énoncé directement à partir du signal acoustique, on peut se poser la question de savoir **si le modèle E2E exploite l'information prosodique qui favorise ses performances de prédiction d'intentions et de concepts**. Pour le savoir, les expériences du chapitre 9 révèlent que l'information prosodique permet au modèle de pointer vers l'information sémantique la plus importante. Il s'avère que les valeurs de pitch plus élevées contribuent à améliorer les performances de l'approche SLU E2E (section 9.2.2) et que l'approche SLU E2E est plus robuste à la parole bruitée (section 9.2.3).

Un autre avantage d'une approche E2E est le pouvoir d'abstraction. Dans une approche séquentielle, la transcription du signal d'entrée en mots est une étape fondamentale. Or les mots hors vocabulaires, les syntaxes inusuelles ne sont pas bien modélisées par les systèmes de RAP courants. **Est-ce qu'un modèle SLU E2E serait capable d'être plus robuste aux variations de vocabulaire et de syntaxe ?** Cette question a été étudiée à deux niveaux (section 9.3) : en augmentant le taux des mots hors vocabulaire et en faisant varier la structure syntaxique des énoncés. Dans tous les cas, le modèle de SLU E2E s'est montré plus robuste que l'approche séquentielle. L'approche SLU E2E montre des performances nettement meilleures par rapport au modèle de SLU séquentielle pour le traitement d'énoncés qui s'écartent des canons grammaticaux du corpus d'apprentissage (section 9.4.1).

La dernière question de cette thèse était liée aux contraintes des approches à base d'apprentissage de réseaux de neurones profonds qui nécessitent des corpus d'apprentissage de grande taille qui sont également adaptés aux besoins de l'application visée. Cependant, comme indiqué au chapitre 2, il y a un manque de corpus dans le domaine domotique en français. On peut donc se demander **comment nous pouvons apprendre des modèles de**

réseaux de neurones profonds avec une faible quantité de données initiales. Pour pouvoir mener nos recherches, nous avons tout d'abord défini l'espace sémantique spécifique au domaine domotique (chapitre 5, section 5.2), et enregistré le corpus VocADom@A4H qui comprend 6747 énoncés de parole réelle (chapitre 6, section 6.2). Ce corpus a été annoté avec 17 étiquettes de concepts différents et avec 8 intentions différentes de manière semi-automatique (chapitre 6, section 6.2.2). Cependant, ce corpus était trop petit pour entraîner des modèles de neurones profonds. C'est pourquoi nous avons généré le corpus artificiel VocADom@ARTIF (chapitre 6, section 6.3), dont les 77k énoncés ont été automatiquement générés et étiquetés de concepts et d'intentions. En utilisant ces énoncés nous avons pu créer un corpus de parole en utilisant de la synthèse vocale. Pour vérifier la pertinence de cette approche nous avons estimé la distance entre le corpus de test de parole réelle et le corpus d'apprentissage de parole artificielle. Cette estimation montrait une distance moyenne interlocuteurs plus petite que la distance moyenne entre les locuteurs et la synthèse vocale. Pour combler cette plus grande distance, et pour disposer d'énoncés de parole spontanée *sans* intention, nous avons augmenté le corpus d'apprentissage grâce au corpus ESLO2 (chapitre 6, section 6.4).

Durant toutes nos expériences, le corpus d'apprentissage a été composé du corpus VocADom@ARTIF ainsi que des extraits de corpus de parole dont les énoncés étaient cohérents avec notre domaine d'application. Le corpus de test était le corpus réel VocADom@A4H. Aucun énoncé de VocADom@A4H n'a été utilisé dans le corpus d'apprentissage afin de respecter le cas réaliste d'un habitat intelligent inconnu avant installation du système. En outre, la génération contrôlée de données nous a notamment permis d'équilibrer les classes d'intention (chapitre 7, section 7.2.2.2, et chapitre 8 , section 8.4) pour éviter les biais d'apprentissage de classes majoritaires. Par ailleurs, la manipulation des corpus a permis de montrer que l'approche SLU E2E était capable d'atteindre les meilleures performances de SLU avec un ensemble d'apprentissage plus petit que l'approche séquentielle. Les activités et les résultats de cette étude ont été publiés dans Desot et coll. (2020).

Cette thèse permet donc de conclure que l'approche SLU E2E est effectivement une approche valide. Cette approche permet de tirer parti du signal acoustique pour prendre ses décisions ce que ne permettaient pas les approches séquentielles classiques. Ces travaux ouvrent également de nombreuses perspectives de recherche.

10.2 Perspectives

Nous avons abordé le manque de données d'apprentissage spécifiques au domaine en combinant la *génération automatique de textes* (NLG - Natural Language Generation) et la synthèse vocale (TTS) pour générer des données d'apprentissage en français. À cette fin, une approche de NLG (Gatt et Krahmer, 2018) à partir d'expertises a été effectuée, approche qui était basée sur des règles, pour créer le corpus artificiel. Cependant, pour générer des données d'entraînement avec plus de variation lexicale et syntaxique, avec moins de distance entre les données artificielles et les données réelles de l'ensemble de test VocADom@A4H, la

technique de *réseaux adverses génératifs* (GANs – *Generative Adversarial Networks*) (Goodfellow et coll., 2014) pourrait être explorée et étudiée. Les GANs sont une classe de réseaux d'apprentissage automatique, où un *générateur* essaie de «tromper» un *discriminateur*. Étant donné un ensemble de données d'apprentissage, le générateur crée de nouvelles données, avec les mêmes statistiques que l'ensemble d'apprentissage. La tâche du discriminateur est opposée dans le sens où son rôle est de distinguer les données *authentiques* des données *artificiellement* créées par le générateur. Les fausses données qui sont quand même considérées comme des données authentiques par le discriminateur, peuvent être ajoutées à l'ensemble d'apprentissage donné. Dans le contexte de ce travail, nos données d'apprentissage pourraient être augmentées par une technique de GAN en utilisant les données spécifiques au domaine de notre ensemble d'entraînement. Une telle approche pourrait contribuer à réduire la distance au niveau lexical et syntaxique entre les données d'apprentissage et de test.

Afin de réduire davantage la distance *acoustique* entre la parole artificielle des données d'apprentissage et la parole réelle de l'ensemble de test, on pourrait générer de nouvelles données par synthèse vocale à base d'un plus grand nombre de voix et les ajouter aux données d'apprentissage. Une autre possibilité, consisterait à entraîner un modèle de synthèse vocale neuronale tel que *Tacotron* (Wang et coll., 2017; Li et coll., 2018) pour un ou plusieurs locuteurs du corpus réel SWEET-HOME. Le modèle appris pourrait être utilisé pour générer de nouveaux signaux acoustiques qui seraient plus proches du corpus de référence.

En utilisant une technique d'apprentissage par transfert, le modèle E2E surpasse le modèle de SLU séquentielle pour la prédiction des *concepts*. Par contre la prédiction *d'intentions* n'a pas suffisamment bénéficié d'un apprentissage par transfert. Une explication possible pourrait provenir du fait que ESPNet étant un moteur de RAP, les décisions sont prises à un niveau trop local pour une abstraction aussi globale que l'intention. Pour une meilleure prédiction des intentions, une approche pourrait être de modifier l'architecture de ESPnet, en ajoutant un décodeur pour les intentions en parallèle de la transcription. Les tâches de reconnaissance de concepts et de classification d'intentions pourraient alors être apprise conjointement. Un apprentissage multi-tâche similaire a déjà été proposé dans le cas de la NLU dans l'étude de Liu et Lane (2016).

Un des atouts de l'approche SLU E2E est son traitement de données bruitées, ce qui rend cette approche très appropriée dans une situation d'habitat intelligent réaliste où les commandes vocales doivent êtres extraites d'énoncés prononcés dans des bruits de fond divers. En plus des bruits de fond, les habitants se trouvent à distance des microphones. Ce qui entraine des signaux acoustiques déformés par la reverberation selon l'acoustique de la pièce. Pour mieux évaluer et comprendre les performances du système de SLU E2E dans une telle situation réaliste, la version en conditions de parole distante du corpus VocADom@A4H devrait être testée dans une prochaine étape. Il s'agit des enregistrements effectués par les quatre antennes de 4 microphones intégrées dans le plafond de l'habitat intelligent Amiqual4Home (chapitre 2 , section 2.2.8). Ces enregistrements pourraient être combinés avec des corpus spécifiques au domaine domotique VoiceHome et VoiceHome-2 (chapitre 2, sec-

tion 2.7.6) qui contiennent environ 7.5 heures de parole bruitée en conditions distantes. Cependant, la durée totale de parole de ces deux corpus et celle du corpus VocADom@A4H (< 20 heures) sont insuffisantes pour permettre d'entraîner un modèle SLU E2E sur des données de parole bruitée. Une solution viable serait l'augmentation de nos données d'entraînement avec des données de *room impulse response* (RIR). Cependant, l'acquisition de données RIR réelles n'est pas anodine. Ko et coll. (2017) montrent que les modèles acoustiques entraînés sur des données de RIR simulées sont compétitifs avec des données de RIR réelles. Par conséquent, une technique d'augmentation de nos données d'apprentissage, des données de RIR simulées, serait une piste de recherche à explorer.

Les activités de recherche de cette thèse ne répondent pas à la question comment résoudre des ambiguïtés dans le cas où il y a des objects identiques dans des pièces différentes. De la commande vocale « allume la lampe », nos systèmes de SLU E2E ainsi que l'approche de référence séquentielle de SLU peuvent extraire l'intention `set_device` et les concepts `action` (allume) et `device` (la lampe). Par contre le système ne peut pas inférer dans quelle pièce de la maison le locuteur se trouve, et ne sait donc pas s'il s'agit de la lampe de la cuisine, ou la lampe du salon, ou d'une autre pièce de la maison. Dans Vacher et coll. (2015) l'interface vocale du système SWEET-HOME (chapitre 2, section 2.3) est combinée avec des modules effectuant une reconnaissance de la situation d'énonciation dans l'habitat. L'emplacement et l'activité des habitants de la maison intelligente, sont inférées à partir des capteurs domotiques. Un classificateur, basé sur les réseaux logiques de Markov (MLN – *Markov Logic Networks*) effectue la reconnaissance et la localisation des activités des habitants. Grâce à ces informations, des règles de decision , également apprises par MLN, déduisent les objets concernés par la commande vocale. La multimodalité du corpus VocADom@A4H (Vacher et coll., 2018; Portet et coll., 2019) pourrait également être exploitée pour intégrer ce type de raisonnement contextuel dans notre système. Comme le corpus a également été conçu pour la *localisation humaine* (HL - *Human Localization*) et pour la *reconnaissance des activités humaines* (HAR - *Human Activity Recognition*), il a été annoté avec l'emplacement et les activités des participants. Outre les annotations des intentions et des concepts du corpus VocADom@A4H, l'annotation d'emplacements, d'activités, et ses horodatages nous permettraient de développer un suivi de l'état de la compréhension au fil des énoncés (*dialogue state tracking* – DST). Les énoncés, les commandes vocales, liées aux activités et aux lieux précédents d'un locuteur pourraient contribuer à une meilleure prédiction des concepts et des intentions. Si l'habitant d'une maison intelligente a l'habitude de se réveiller à 7h00 dans sa chambre à coucher, prend sa douche dans la salle de bain à 7h15, émet une commande vocale dans la cuisine à 7h30 pour que la maison intelligente prépare son café, la prise en compte de l'historique de ces informations de contexte, autrement dit l'apprentissage des habitudes de l'habitant, pourrait améliorer les performances de SLU. La prédiction d'intentions et de concepts dépendrait alors de plusieurs niveaux d'informations comme la commande vocale précédente, l'énoncé précédent sans commande vocale, l'activité précédente, l'emplacement précédent etc. Similaire à l'étude de (Su et coll., 2018) un mécanisme d'attention (chapitre 3, section 3.1.4.2.2) pourrait être exploité en utilisant des

réseaux de neurones profonds, pour sélectionner les niveaux d'information les plus pertinents. Une telle piste de recherche pourrait mener vers une SLU de bout en bout *contextuelle* (Contextual, End-to-end Spoken Language Understanding).

Bibliographie

ABDEL-HAMID, O., MOHAMED, A.-r., JIANG, H. et PENN, G. (2012). Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4277–4280. IEEE.

ABOWD, G. D., BOBICK, A. F., ESSA, I. A., MYNATT, E. D. et ROGERS, W. A. (2002). The aware home : A living laboratory for technologies for successful aging. Dans *Proceedings of the AAAI-02 Workshop "Automation as Caregiver*, pages 1–7.

AMAN, F. (2014). *Automatic speech recognition for ageing voices in the context of assisted living*. Thèse de doctorat, Université de Grenoble.

AMAN, F., VACHER, M., ROSSATO, S. et PORTET, F. (2013). Analysing the performance of automatic speech recognition for ageing voice : Does it correlate with dependency level? Dans *Speech and Language Processing for Assistive Technologies, Satellite workshop of Interspeech2013*, pages 1–7, Grenoble, France.

AMODEI, D., ANANTHANARAYANAN, S., ANUBHAI, R., BAI, J., BATTENBERG, E., CASE, C., CASPER, J., CATANZARO, B., CHENG, Q., CHEN, G. et coll. (2016). Deep speech 2 : End-to-end speech recognition in english and mandarin. Dans *International conference on machine learning (ICML)*, pages 173–182.

ANASTASAKOS, T., MCDONOUGH, J., SCHWARTZ, R. et MAKHOUL, J. (1996). A compact model for speaker-adaptive training. Dans *International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 1137–1140. IEEE.

ANGELINI, B., BRUGNARA, F., FALAVIGNA, D., GIULIANI, D., GRETTER, R. et OMOLOGO, M. (1993). Automatic segmentation and labeling of english and italian speech databases. Dans *European Conference on Speech Communication and Technology (EUROSPEECH)*.

BAHDANAU, D., CHO, K. et BENGIO, Y. (2015). Neural machine translation by jointly learning to align and translate. Dans *International Conference on Learning Representations (ICLR)*.

BAHDANAU, D., CHOROWSKI, J., SERDYUK, D., BRAKEL, P. et BENGIO, Y. (2016). End-to-end attention-based large vocabulary speech recognition. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4945–4949. IEEE.

BAPNA, A., TUR, G., HAKKANI-TUR, D. et HECK, L. (2017). Sequential dialogue context modeling for spoken language understanding. Dans *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*.

BARKER, J., MARXER, R., VINCENT, E. et WATANABE, S. (2017). The CHiME challenges : Robust speech recognition in everyday environments. Dans *New Era for Robust Speech Recognition - Exploiting Deep Learning*, pages 327–344. Springer.

BARKER, J., VINCENT, E., MA, N., CHRISTENSEN, H. et GREEN, P. (2013). The pascal chime speech separation and recognition challenge. *Computer Speech & Language*, 27(3):621–633.

BARKER, J., WATANABE, S., VINCENT, E. et TRMAL, J. (2018). The fifth'chime'speech separation and recognition challenge : Dataset, task and baselines. Dans *Interspeech*, pages 1561–1565.

BÉCHET, F. et CHARTON, E. (2010). Unsupervised knowledge acquisition for extracting named entities from speech. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5338–5341. IEEE.

BENTLEY, F., LUVOGT, C., SILVERMAN, M., WIRASINGHE, R., WHITE, B. et LOTTRIDGE, D. (2018). Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–24.

BERTIN, N., CAMBERLEIN, E., LEBARBENCHON, R., VINCENT, E., SIVASANKARAN, S., ILLINA, I. et BIMBOT, F. (2019). Voicehome-2, an extended corpus for multichannel speech processing in real homes. *Speech Communication*, 106:68–78.

BERTIN, N., CAMBERLEIN, E., VINCENT, E., LEBARBENCHON, R., PEILLON, S., LAMANDÉ, E., SIVASANKARAN, S., BIMBOT, F., ILLINA, I., TOM, A., FLEURY, S. et JAMET, E. (2016). A French corpus for distant-microphone speech processing in real homes. Dans *Interspeech*, pages 2781–2785.

BISHOP, C. M. (2006). *Pattern recognition and machine learning*. springer.

BOBILLIER-CHAUMON, M.-E., CUVILLIER, B., BOUAKAZ, S. et VACHER, M. (2012). Démarche de développement de technologies ambiantes pour le maintien à domicile des personnes dépendantes : vers une triangulation des méthodes et des approches. Dans *Actes du 1er Congrès Européen de Stimulation Cognitive*, pages 121–122, Dijon, France.

BONNEAU-MAYNARD, H., ROSSET, S., AYACHE, C., KUHN, A. et MOSTEFA, D. (2005). Semantic annotation of the french media dialog corpus. Dans *European Conference on Speech Communication and Technology (EUROSPEECH)*.

BRAUN, D., HERNANDEZ-MENDEZ, A., MATTHES, F. et LANGEN, M. (2017). Evaluating natural language understanding services for conversational question answering systems. Dans *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*.

BRITZ, D., GOLDIE, A., LUONG, T. et LE, Q. (2017). Massive Exploration of Neural Machine Translation Architectures. *ArXiv e-prints*.

BRUMITT, B., MEYERS, B., KRUMM, J., KERN, A. et SHAFER, S. (2000). Easyliving : Technologies for intelligent environments. Dans *International Symposium on Handheld and Ubiquitous Computing*, pages 12–29. Springer.

BRUMITT, B. et SHAFER, S. (2001). Better living through geometry. *Personal and ubiquitous computing*, 5(1):42–45.

BRUTTI, A., CRISTOFORETTI, L., KELLERMANN, W., MARQUARDT, L. et OMOLOGO, M. (2008). WOZ acoustic data collection for interactive TV. Dans *International Conference on Language Resources and Evaluation (LREC)*, pages 2330–2334.

CANCELLIERI, A. (1992). *L'habitat du futur : défis et prospective pour le prochain quart de siècle*. Documentation française.

CAUBRIÈRE, A., TOMASHENKO, N., LAURENT, A., MORIN, E., CAMELIN, N. et ESTÈVE, Y. (2019). Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability. Dans *interspeech*, pages 1198–1202.

CETTOLO, M., GIRARDI, C. et FEDERICO, M. (2012). Wit3 : Web inventory of transcribed and translated talks. Dans *Conference of european association for machine translation*, pages 261–268.

CHAN, M., ESTÈVE, D., ESCRIBA, C. et CAMPO, E. (2008). A review of smart homes- present state and future challenges. *Computer Methods and Programs in Biomedicine*, 91(1):55–81.

CHAN, W., JAITLY, N., LE, Q. et VINYALS, O. (2016). Listen, attend and spell : A neural network for large vocabulary conversational speech recognition. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4960–4964. IEEE.

CHO, J., BASKAR, M. K., LI, R., WIESNER, M., MALLIDI, S. H., YALTA, N., KARAFIAT, M., WATANABE, S. et HORI, T. (2018). Multilingual sequence-to-sequence speech recognition : architecture, transfer learning, and language modeling. Dans *IEEE Spoken Language Technology Workshop (SLT)*, pages 521–527. IEEE.

CHUNG, J., GÜLÇEHRE, Ç., CHO, K. et BENGIO, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

CIERI, C., MILLER, D. et WALKER, K. (2004). The fisher corpus : a resource for the next generations of speech-to-text. Dans *International Conference on Language Resources and Evaluation (LREC)*, volume 4, pages 69–71.

COOK, D. et DAS, S. K. (2004). *Smart environments : technology, protocols, and applications*, volume 43. John Wiley & Sons.

CORTES, C. et VAPNIK, V. (1995). Support-vector networks. Dans *Machine Learning*, pages 273–297.

CRISTOFORETTI, L., RAVANELLI, M., OMOLOGO, M., SOSI, A., ABAD, A., HAGMÜLLER, M. et MARAGOS, P. (2014). The dirha simulated corpus. Dans *International Conference on Language Resources and Evaluation (LREC)*, pages 2629–2634.

CRYSTAL, D. (2011). *A Dictionary of Linguistics and Phonetics*. The Language Library. Wiley.

DAS, S. K., COOK, D. J., BATTACHARYA, A., HEIERMAN, E. O. et LIN, T.-Y. (2002). The role of prediction algorithms in the mavhome smart home architecture. *IEEE wireless communications*, 9(6):77–84.

DAVIS, S. et MERMELSTEIN, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.

DELÉGLISE, P., ESTÈVE, Y., MEIGNIER, S. et MERLIN, T. (2005). The lium speech transcription system : a cmu sphinx iii-based system for french broadcast news. Dans *Interspeech 2005*, pages 1653–1656.

DELÉGLISE, P., ESTEVE, Y., MEIGNIER, S. et MERLIN, T. (2009). Improvements to the lium french asr system based on cmu sphinx : what helps to significantly reduce the word error rate? Dans *Tenth Annual Conference of the International Speech Communication Association*.

DESOT, T., PORTET, F. et VACHER, M. (2019a). Slu for voice command in smart home : comparison of pipeline and end-to-end approaches. Dans *Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 822–829. IEEE.

DESOT, T., PORTET, F. et VACHER, M. (2019b). Towards end-to-end spoken intent recognition in smart home. Dans *Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–8.

DESOT, T., PORTET, F. et VACHER, M. (2020). Corpus generation for voice command in smart home and the effect of speech synthesis on end-to-end slu. Dans *International Conference on Language Resources and Evaluation (LREC)*, pages 6395–6404.

DESOT, T., RAIMONDO, S., MISHAKOVA, A., PORTET, F. et VACHER, M. (2018). Towards a french smart-home voice command corpus : Design and nlu experiments. Dans *International Conference on Text, Speech, and Dialogue (TSD)*, pages 509–517. Springer.

DEVLIN, J., CHANG, M.-W., LEE, K. et TOUTANOVA, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. Dans *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

DHINGRA, S. D., NIJHAWAN, G. et PANDIT, P. (2013). Isolated speech recognition using mfcc and dtw. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (IJAREEIE)*, 2(8):4085–4092.

DIGALAKIS, V. V. et NEUMEYER, L. G. (1996). Speaker adaptation using combined transformation and bayesian methods. *IEEE transactions on speech and audio processing*, 4(4):294–300.

DOWDING, J., GAWRON, J. M., APPELT, D., BEAR, J., CHERNY, L., MOORE, R. et MORAN, D. (1993). Gemini : A natural language system for spoken-language understanding. Dans *International Conference on Computational Linguistics)*, pages 54–61. Association for Computational Linguistics.

DUÉE, M. et REBILLARD, C. (2006). La dépendance des personnes âgées : une projection en 2040. *Données sociales - La société française*, pages 613–619.

ELLOUMI, Z. (2019). *Prédiction de performances des systèmes de Reconnaissance Automatique de la Parole*. Thèse de doctorat.

ELLOUMI, Z., BESACIER, L., GALIBERT, O., KAHN, J. et LECOUTEUX, B. (2018). Asr performance prediction on unseen broadcast programs using convolutional neural networks. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5894–5898. IEEE.

ESTEVE, Y., BAZILLON, T., ANTOINE, J.-Y., BÉCHET, F. et FARINAS, J. (2010). The epac corpus : Manual and automatic annotations of conversational speech in french broadcast news. Dans *International Conference on Language Resources and Evaluation (LREC)*, pages 1686–1689. Citeseer.

FAYEK, H. M. (2016). Speech processing for machine learning : Filter banks, mel-frequency cepstral coefficients (mfccs) and what's in-between.

FLEURY, A., NOURY, N. et VACHER, M. (2010a). Introducing knowledge in the process of supervised classification of activities of daily living in health smart homes. Dans *International Conference on e-Health Networking, Applications and Services*, pages 322–329. IEEE.

FLEURY, A., VACHER, M., PORTET, F., CHAHUARA, P. et NOURY, N. (2010b). A multimodal corpus recorded in a health smart home. Dans *International Conference on Language Resources and Evaluation (LREC), Workshop Multimodal Corpora and Evaluation*, pages 99–105.

GALLIANO, S., GEOFFROIS, E., MOSTEFA, D., CHOUKRI, K., BONASTRE, J.-F. et GRAVIER, G. (2005). The ester phase ii evaluation campaign for the rich transcription of french broadcast news. Dans *European Conference on Speech Communication and Technology (EUROSPEECH)*.

GALLIANO, S., GRAVIER, G. et CHAUBARD, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. Dans *Tenth Annual Conference of the International Speech Communication Association*.

GALLISSOT, M., CAELEN, J., JAMBON, F. et MEILLON, B. (2013). Une plate-forme usage pour l'intégration de l'informatique ambiante dans l'habitat : Domus. *Technique et Science Informatiques (TSI)*, 32(5):547–574.

GATT, A. et KRAHMER, E. (2018). Survey of the state of the art in natural language generation : Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61(1).

GAUVAIN, J.-L., LAMEL, L. F. et ESKÉNAZI, M. (1990). Design considerations and text selection for bref, a large french read-speech corpus. Dans *International Conference on Spoken Language Processing (ICSLP)*, pages 1097–1100.

GHANNAY, S., CAUBRIÈRE, A., ESTÈVE, Y., LAURENT, A. et MORIN, E. (2018). End-to-end named entity extraction from speech. *arXiv preprint arXiv :1805.12045*.

GHANNAY, S., ESTÈVE, Y., CAMELIN, N. et DELÉGLISE, P. (2016). Acoustic word embeddings for asr error detection. Dans *Interspeech*, pages 1330–1334.

GIRAUDEL, A., CARRÉ, M., MAPELLI, V., KAHN, J., GALIBERT, O. et QUINTARD, L. (2012). The repere corpus : a multimodal corpus for person recognition. Dans *International Conference on Language Resources and Evaluation (LREC)*, pages 1102–1107.

GODFREY, J. J., HOLLIMAN, E. C. et MCDANIEL, J. (1992). Switchboard : Telephone speech corpus for research and development. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 517–520. IEEE.

GOLDWATER, S., JURAFSKY, D. et MANNING, C. D. (2010). Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

GOODFELLOW, I. J., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. et BENGIO, Y. (2014). Generative adversarial nets. Dans *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, pages 2672–2680.

GRAVES, A., FERNÁNDEZ, S., GOMEZ, F. et SCHMIDHUBER, J. (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. Dans *International conference on Machine learning (ICML)*, pages 369–376. ACM.

GRAVES, A., MOHAMED, A.-r. et HINTON, G. (2013). Speech recognition with deep recurrent neural networks. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 6645–6649. IEEE.

GRAVIER, G., ADDA, G., PAULSON, N., CARRÉ, M., GIRAUDEL, A. et GALIBERT, O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. Dans *International Conference on Language Resources and Evaluation (LREC)*.

GUPTA, N., TUR, G., HAKKANI-TUR, D., BANGALORE, S., RICCARDI, G. et GILBERT, M. (2005). The at&t spoken language understanding system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):213–222.

HAHN, S., LEHNEN, P., RAYMOND, C. et NEY, H. (2008). A comparison of various methods for concept tagging for spoken language understanding. Dans *International Conference on Language Resources and Evaluation (LREC)*.

HAKKANI-TÜR, D., BÉCHET, F., RICCARDI, G. et TUR, G. (2006). Beyond asr 1-best : Using word confusion networks in spoken language understanding. *Computer Speech & Language*, 20(4):495–514.

HANNUN, A., CASE, C., CASPER, J., CATANZARO, B., DIAMOS, G., ELSEN, E., PRENGER, R., SATHEESH, S., SENGUPTA, S., COATES, A. et coll. (2014). Deep speech : Scaling up end-to-end speech recognition. *arXiv preprint arXiv :1412.5567*.

HATMI, M., JACQUIN, C., MORIN, E. et MEIGNER, S. (2013). Incorporating named entity recognition into the speech transcription process. Dans *Interspeech 2013*, pages 3732–3736.

HATON, J.-P. (1999). Neural networks for automatic speech recognition : a review. Dans *Speech Processing, Recognition and Artificial Neural Networks*, pages 259–280. Springer.

HATON, J.-P., CERISARA, C., FOHR, D., LAPRIE, Y. et SMAÏLI, K. (2006). *Reconnaissance automatique de la parole : Du signal à son interprétation*. Dunod.

HE, Y. et YOUNG, S. (2003). A data-driven spoken language understanding system. Dans *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 583–588. IEEE.

HEMPHILL, C. T., GODFREY, J. J. et DODDINGTON, G. R. (1990). The ATIS spoken language systems pilot corpus. Dans *Proceedings of the Workshop on Speech and Natural Language*, pages 96–101. Association for Computational Linguistics.

HENDERSON, M., THOMSON, B. et WILLIAMS, J. D. (2014). The second dialog state tracking challenge. Dans *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.

HERMANSKY, H. et COX JR, L. A. (1991). Perceptual linear predictive (plp) analysis-resynthesis technique. Dans *European Conference on Speech Communication and Technology (EUROSPEECH)*.

HUANG, L., SIL, A., JI, H. et FLORIAN, R. (2017). Improving slot filling performance with attentive neural networks on dependency structures. Dans *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2588–2597.

INTILLE, S. S. (2006). The goal : smart people, not smart homes. Dans *The International Conference on Smart Homes and Health Telematics (ICOST)*, pages 3–6.

ISTRATE, D. (2003). *Détection et Reconnaissance des Sons pour la Surveillance Médicale*. Thèse de doctorat, INP Grenoble, École Doctorale « Électronique, Électrotechnique, Automatique, Télécommunications, Signal ».

ISTRATE, D., CASTELLI, E., VACHER, M., BESACIER, L. et SERIGNAT, J.-F. (2006). Information extraction from sound for medical telemonitoring. *IEEE Transactions on Information Technology in Biomedicine*, 10:264–274.

JELINEK, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556.

JELINEK, F., MERCER, R. L., BAHL, L. R. et BAKER, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

JEONG, M. et LEE, G. G. (2008). Triangular-chain conditional random fields. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 16(7):1287–1302.

JEONG, M. et LEE, G. G. (2009). Multi-domain spoken language understanding with transfer learning. *Speech Communication*, 51(5):412–424.

KAIN, A. et SANTEN, J. P. v. (2010). Frequency-domain delexicalization using surrogate vowels. Dans *Interspeech*, pages 474–477.

KIPYATKOVA, I. et KARPOV, A. (2016). Dnn-based acoustic modeling for russian speech recognition using kaldi. Dans *International Conference on Speech and Computer (SPECOM)*, pages 246–253. Springer.

KNESER, R. et NEY, H. (1995). Improved backing-off for m-gram language modeling. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 181–184. IEEE.

KO, T., PEDDINTI, V., POVEY, D., SELTZER, M. L. et KHUDANPUR, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5220–5224. IEEE.

KOEHN, P. (2005). Europarl : A parallel corpus for statistical machine translation. Dans *MT summit*, volume 5, pages 79–86. Citeseer.

KRUEGER, K. A. et DAYAN, P. (2009). Flexible shaping : How learning in small steps helps. *Cognition*, 110(3):380–394.

KUHN, R. et DE MORI, R. (1995). The application of semantic classification trees to natural language understanding. *IEEE transactions on pattern analysis and machine intelligence*, 17(5):449–460.

LEE, A., KAWAHARA, T. et SHIKANO, K. (2001). Julius—an open source real-time large vocabulary recognition engine. Dans *Interspeech*, pages 1219–1222.

LEFÈVRE, F., MOSTEFA, D., BESACIER, L., ESTÈVE, Y., QUIGNARD, M., CAMELIN, N., FAVRE, B., JABAIAN, B. et ROJAS-BARAHONA, L. M. (2012). Leveraging study of robustness and portability of spoken language understanding systems across languages and domains : the PORTMEDIA corpora. Dans *International Conference on Language Resources and Evaluation (LREC)*, pages 1436–1442.

LEGGETTER, C. et WOODLAND, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2):171 – 185.

LEVIN, E. et PIERACCINI, R. (1995). Concept-based spontaneous speech understanding system. Dans *European Conference on Speech Communication and Technology (EUROSPEECH)*.

LI, J., GADDE, R., GINSBURG, B. et LAVRUKHIN, V. (2018). Training neural speech recognition systems with synthetic speech augmentation. *arXiv preprint arXiv :1811.00707*.

LISON, P. et TIEDEMANN, J. (2016). Opensubtitles2016 : Extracting large parallel corpora from movie and tv subtitles. Dans *International Conference on Language Resources and Evaluation (LREC)*, pages 923–929.

LIU, B. et LANE, I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. Dans *Interspeech*, pages 685–689.

LIU, C., ZHU, S., ZHAO, Z., CAO, R., CHEN, L. et YU, K. (2020). Jointly encoding word confusion network and dialogue context with bert for spoken language understanding. Dans *Interspeech*, pages 871–875.

LIU, Y., FUNG, P., YANG, Y., CIERI, C., HUANG, S. et GRAFF, D. (2006). Hkust/mts : A very large scale mandarin telephone speech corpus. Dans *International Symposium on Chinese Spoken Language Processing*, pages 724–735. Springer.

LÓPEZ, G., QUESADA, L. et GUERRERO, L. A. (2017). Alexa vs. siri vs. cortana vs. google assistant : a comparison of speech-based natural user interfaces. Dans *International Conference on Applied Human Factors and Ergonomics*, pages 241–250. Springer.

LUGOSCH, L., MEYER, B., NOWROUZEZAHRAI, D. et RAVANELLI, M. (2020). Using speech synthesis to train end-to-end spoken language understanding models. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE.

LUGOSCH, L., RAVANELLI, M., IGNOTO, P., TOMAR, V. S. et BENGIO, Y. (2019). Speech model pre-training for end-to-end spoken language understanding. Dans *Interspeech*, pages 814–818.

MAEKAWA, K. (2003). Corpus of spontaneous japanese : Its design and evaluation. Dans *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, pages 7–12.

MANGU, L., BRILL, E. et STOLCKE, A. (2000). Finding consensus in speech recognition : word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4):373 – 400.

MARKEL, J. D. et GRAY, A. J. (2013). *Linear prediction of speech*, volume 12. Springer Science & Business Media.

MCCULLOCH, W. S. et PITTS, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

MESNIL, G., DAUPHIN, Y., YAO, K., BENGIO, Y., DENG, L., HAKKANI-TUR, D., HE, X., HECK, L., TUR, G., YU, D. et OTHERS (2015). Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(3):530–539.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S. et DEAN, J. (2013). Distributed representations of words and phrases and their compositionality. Dans *Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119.

MISHAKOVA, A., PORTET, F., DESOT, T. et VACHER, M. (2019). Learning natural language understanding systems from unaligned labels for voice command in smart homes. Dans *IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 832–837.

MÖLLER, S., GÖDDE, F. et WOLTERS, M. (2008). Corpus analysis of spoken smart-home interactions with older users. Dans *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

MUDA, L., BEGAM, M. et ELAMVAZUTHI, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *Journal Of Computing*, 2(3):138–143.

NÉVÉOL, A., GROUIN, C., LEIXA, J., ROSSET, S. et ZWEIGENBAUM, P. (2014). The quaero french medical corpus : A ressource for medical entity recognition and normalization. Dans *In Proc BioTextM, Reykjavik*. Citeseer.

PALLETT, D. S. (1989). Benchmark tests for darpa resource management database performance evaluations. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 536–539. IEEE.

PASZKE, A., GROSS, S., CHINTALA, S., CHANAN, G., YANG, E., DEVITO, Z., LIN, Z., DESMAISON, A., ANTIGA, L. et LERER, A. (2017). Automatic differentiation in pytorch. Dans *Advances in Neural Information Processing Systems (NIPS) workshop*.

PAUL, D. B. et BAKER, J. M. (1992). The design for the wall street journal-based csr corpus. Dans *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics.

PERENNOU, G. (1986). B.D.L.E.X. : A data and cognition base of spoken French. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 11, pages 325–328. IEEE.

PLACK, C. J. et OXENHAM, A. J. (2005). Overview : The present and future of pitch. Dans *Pitch*, pages 1–6. Springer.

POMPONIO, L., LE GOC, M., ANFOSSO, A. et PASCUAL, E. (2012). Levels of abstraction for behavior modeling in the gerhome project. *International Journal of E-Health and Medical Communications (IJEHMC)*, 3(3):12–28.

PORTET, F., CAFFIAU, S., RINGEVAL, F., VACHER, M., BONNEFOND, N., ROSSATO, S., LECOUTEUX, B. et DESOT, T. (2019). Context-aware voice-based interaction in smart home-vocadom@ a4h corpus collection and empirical assessment of its usefulness. Dans *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 811–818. IEEE.

POVEY, D., BURGET, L., AGARWAL, M., AKYAZI, P., KAI, F., GHOSHAL, A., GLEMBEK, O., GOEL, N., KARAFIÁT, M., RASTROW, A. et coll. (2011a). The subspace gaussian mixture model—a structured model for speech recognition. *Computer Speech & Language*, 25(2):404–439.

POVEY, D., GHOSHAL, A., BOULIANNE, G., BURGET, L., GLEMBEK, O., GOEL, N., HANNEMANN, M., MOTLICEK, P., QIAN, Y., SCHWARZ, P. et coll. (2011b). The kaldi speech recognition toolkit. Dans *Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

POVEY, D., ZHANG, X. et KHUDANPUR, S. (2015). Parallel training of dnns with natural gradient and parameter averaging. Dans *International Conference on Learning Representations (ICLR)*.

PRICE, P., FISHER, W. M., BERNSTEIN, J. et PALLETT, D. S. (1988). The darpa 1000-word resource management database for continuous speech recognition. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 651–654. IEEE.

PRINCIPI, E., SQUARTINI, S., PIAZZA, F., FUSELLI, D. et BONIFAZI, M. (2013). A distributed system for recognizing home automation commands and distress calls in the italian language. Dans *Interspeech 2013*, pages 2049–2053.

PURINGTON, A., TAFT, J. G., SANNON, S., BAZAROVA, N. N. et TAYLOR, S. H. (2017). " alexa is my new bff" social roles, user satisfaction, and personification of the amazon echo. Dans *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA)*, pages 2853–2859.

QIAN, Y., BI, M., TAN, T. et YU, K. (2016). Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(12):2263–2276.

RABINER, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

RASHIDI, P. et COOK, D. J. (2009). Keeping the resident in the loop : Adapting the smart home to the user. *IEEE Transactions on systems, man, and cybernetics-part A : systems and humans*, 39(5):949–959.

RATANAMAHATANA, C. A. et KEOGH, E. (2004). Everything you know about dynamic time warping is wrong. Dans *Third workshop on mining temporal and sequential data (SIGKDD)*, volume 32. Citeseer.

RAVANELLI, M., CRISTOFORETTI, L., GRETTER, R., PELLIN, M., SOSI, A. et OMOLOGO, M. (2015). The DIRHA-English corpus and related tasks for distant-speech recognition in domestic environments. Dans *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 275–282.

RENALS, S. et HAIN, T. (2010). Speech recognition. Dans CLARK, A., FOX, C. et LAPPIN, S., éditeurs : *Handbook of Computational Linguistics and Natural Language Processing*. Wiley Blackwell.

RENALS, S., MORGAN, N., BOURLARD, H., COHEN, M. et FRANCO, H. (1994). Connectionist probability estimators in hmm speech recognition. *IEEE transactions on speech and audio processing*, 2(1):161–174.

RIALLE, V., NOURY, N. et HERVÉ, T. (2001). An experimental health smart home and its distributed internet-based information and communication system : first steps of a research project. *Studies in health technology and informatics*, (2):1479–1483.

ROBINSON, T., HOCHBERG, M. et RENALS, S. (1996). The use of recurrent neural networks in continuous speech recognition. Dans *Automatic speech and speaker recognition*, pages 233–258. Springer.

ROUSSEAU, A., BOULIANNE, G., DELÉGLISE, P., ESTÈVE, Y., GUPTA, V. et MEIGNIER, S. (2014). Lium and crim asr system combination for the repere evaluation campaign. Dans *International Conference on Text, Speech, and Dialogue (TSD)*, pages 441–448. Springer.

RYBACH, D., GOLLAN, C., HEIGOLD, G., HOFFMEISTER, B., LÖÖF, J., SCHLÜTER, R. et NEY, H. (2009). The rwth aachen university open source speech recognition system. Dans *Interspeech*, pages 2111–2114.

SAADE, A., COUCKE, A., CAULIER, A., DUREAU, J., BALL, A., BLUCHE, T., LEROY, D., DOUMOURO, C., GISSELBRECHT, T., CALTAGIRONE, F. et coll. (2018). Spoken language understanding on the edge. *arXiv preprint arXiv :1810.12735*.

SAINATH, T. N., KINGSBURY, B., MOHAMED, A.-r., DAHL, G. E., SAON, G., SOLTAU, H., BERAN, T., ARAVKIN, A. Y. et RAMABHADRAN, B. (2013a). Improvements to deep convolutional neural networks for lvcsr. Dans *Workshop on automatic speech recognition and understanding (ASRU)*, pages 315–320. IEEE.

SAINATH, T. N., MOHAMED, A.-r., KINGSBURY, B. et RAMABHADRAN, B. (2013b). Deep convolutional neural networks for lvcsr. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 8614–8618. IEEE.

SAKAMURA, K. (1996). Bibliography of the tron project (1984-1996). Dans *Proceedings 13th TRON Project International Symposium /TEPS '96*, pages 144–175.

SAKOE, H. et CHIBA, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

SCHWARTZ, R., BARRY, C., CHOW, Y.-L., DERR, A., FENG, M.-W., KIMBALL, O., KUBALA, F., MAKHOUL, J. et VANDEGRIFT, J. (1989). The bbn byblos continuous speech recognition system. Dans *Proceedings of the workshop on Speech and Natural Language*, pages 94–99. Association for Computational Linguistics.

SCHWARTZ, R., MILLER, S., STALLARD, D. et MAKHOUL, J. (1996). Language understanding using hidden understanding models. Dans *International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 997–1000. IEEE.

SENEFF, S. (1992). Tina : A natural language system for spoken language applications. *Comput. Linguist.*, 18(1):61–86.

SERBAN, I. V., LOWE, R., HENDERSON, P., CHARLIN, L. et PINEAU, J. (2015). A survey of available corpora for building data-driven dialogue systems. *CoRR*, abs/1512.05742.

SERDYUK, D., WANG, Y., FUEGEN, C., KUMAR, A., LIU, B. et BENGIO, Y. (2018). Towards end-to-end spoken language understanding. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5754–5758. IEEE.

SERPOLLET, N., BERGOUNIOUX, G., CHESNEAU, A. et WALTER, R. (2007). A large reference corpus for spoken french : Eslo 1 and 2 and its variations. Dans *Proceedings from Corpus Linguistics Conference Series, University of Birmingham*.

SHAFER, S., KRUMM, J., BRUMITT, B., MEYERS, B., CZERWINSKI, M. et ROBBINS, D. (1998). The new easyliving project at microsoft research. Dans *Proceedings of the 1998 DARPA/NIST smart spaces workshop*, pages 127–130.

SHAWAR, B. A. et ATWELL, E. (2002). *A comparison between Alice and Elizabeth chatbot systems.* University of Leeds, School of Computing research report 2002.19.

SIMONNET, E., GHANNAY, S., CAMELIN, N., ESTÈVE, Y. et DE MORI, R. (2017). Asr error management for improving spoken language understanding. Dans *Interspeech*, pages 3329–3333.

STEHWIEN, S. et VU, N. T. (2016). Exploring the correlation of pitch accents and semantic slots for spoken language understanding. Dans *Interspeech*, pages 730–734.

STOLCKE, A. (2002). Srilm – an extensible language modeling toolkit. Dans *International Conference on Spoken Language Processing (ICSLP)*, pages 901–904.

SU, C.-y. et TSENG, C.-y. (2018). Perceivable information structure in discourse prosody-detecting prominent prosodic words in spoken discourse using f0 contour. Dans *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 424–428. IEEE.

SU, H., DZODZO, B., WU, X., LIU, X. et MENG, H. (2019). Unsupervised methods for audio classification from lecture discussion recordings. Dans *Interspeech 2019*, pages 3347–3351.

SU, S.-Y., YUAN, P.-C. et CHEN, Y.-N. (2018). How time matters : Learning time-decay attention for contextual spoken language understanding in dialogues. Dans *Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (HLT-NAACL)*, pages 2133–2142.

SUDOH, K., TSUKADA, H. et ISOZAKI, H. (2006). Incorporating speech recognition confidence into discriminative named entity recognition of speech data. Dans *International Conference on Computational Linguistics*, pages 617–624. Association for Computational Linguistics.

TAIGMAN, Y., WOLF, L., POLYAK, A. et NACHMANI, E. (2018). Voiceloop : Voice fitting and synthesis via a phonological loop. Dans *International Conference on Learning Representations (ICLR)*.

TAKAHASHI, S.-y., MORIMOTO, T., MAEDA, S. et TSURUTA, N. (2003). Dialogue experiment for elderly people in home health care system. Dans *International Conference on Text, Speech, and Dialogue (TSD)*, pages 418–423.

TAN, T.-P. et BESACIER, L. (2006). A french non-native corpus for automatic speech recognition. Dans *International Conference on Language Resources and Evaluation (LREC)*, volume 6, pages 1610–1613.

TESSEMA, N. M., ONS, B., van de LOO, J., GEMMEKE, J., DE PAUW, G., DAELEMANS, W. et coll. (2013). Metadata for corpora patcor and domotica-2. *Technical report KUL/ESAT/PSI/1303, KU Leuven, ESAT, Leuven, Belgium.*

TOKUI, S., OONO, K., HIDO, S. et CLAYTON, J. (2015). Chainer : a next-generation open source framework for deep learning. Dans *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on Neural Information Processing Systems (NIPS)*, volume 5, pages 1–6.

TUR, G. et DE MORI, R. (2011). *Spoken Language Understanding Systems for Extracting Semantic Information from Speech.* Wiley.

UENO, S., INAGUMA, H., MIMURA, M. et KAWAHARA, T. (2018). Acoustic-to-word attention-based model complemented with character-level ctc-based model. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5804–5808. IEEE.

VACHER, M., AMAN, F., ROSSATO, S., PORTET, F. et LECOUTEUX, B. (2019). Making emergency calls more accessible to older adults through a hands-free speech interface in the house. *ACM Transactions on Accessible Computing (TACCESS)*, 12(2):1–25.

VACHER, M., BOUAKAZ, S., CHAUMON, M.-E. B., AMAN, F., KHAN, R. A., BEKKADJA, S., PORTET, F., GUILLOU, E., ROSSATO, S. et LECOUTEUX, B. (2016). The cirdo corpus : comprehensive audio/video database of domestic falls of elderly people. Dans *International Conference on Language Resources and Evaluation (LREC)*, pages 1389–1396.

VACHER, M., CAFFIAU, S., PORTET, F., MEILLON, B., ROUX, C., ELIAS, E., LECOUTEUX, B. et CHAHUARA, P. (2015). Evaluation of a context-aware voice interface for Ambient Assisted Living : qualitative user study vs. quantitative system evaluation. *ACM Transactions on Accessible Computing (TACCESS)*, 7(issue 2):5 :1–5 :36.

VACHER, M., FLEURY, A., SERIGNAT, J.-F., NOURY, N. et GLASSON, H. (2008). Preliminary evaluation of speech/sound recognition for telemedicine application in a real environment. Dans *Interspeech)*, pages 496–499.

VACHER, M., LECOUTEUX, B., CHAHUARA, P., PORTET, F., MEILLON, B. et BONNEFOND, N. (2014). The Sweet-Home speech and multimodal corpus for home automation interaction. Dans *International Conference on Language Resources and Evaluation (LREC)*, pages 4499–4506.

VACHER, M., PORTET, F., FLEURY, A. et NOURY, N. (2011). Development of audio sensing technology for ambient assisted living : Applications and challenges. *International Journal of E-Health and Medical Communications*, 2(1):35 – 54.

VACHER, M., SERIGNAT, J.-F., CHAILLOL, S., ISTRATE, D. et POPESCU, V. (2006). Speech and sound use in a remote monitoring system for health care. Dans *International Conference on Text, Speech and Dialogue (TSD)*, pages 711–718. Springer.

VACHER, M., VINCENT, E., BOBILLIER CHAUMON, M.-E., JOUBERT, T., PORTET, F., FOHR, D., CAFFIAU, S. et DESOT, T. (2018). The VocADom Project : Speech Interaction for Well-being and Reliance Improvement. Dans *MobileHCI 2018 - 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*.

VANDEWIELE, F. et MOTAMED, C. (2011). Phd forum : A data mining approach for human activity learning in a multi-modal sensor system. Dans *2011 Fifth ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1–2. IEEE.

VAUFREYDAZ, D., BERGAMINI, C., SERIGNAT, J.-F., BESACIER, L., AKBAR, M. et coll. (2000). A new methodology for speech corpora definition from internet documents. Dans *International Conference on Language Resources and Evaluation (LREC)*.

VIRONE, G., NOURY, N. et DEMONGEOT, J. (2002). A system for automatic measurement of circadian activity deviations in telemedicine. *IEEE Transactions on Biomedical Engineering*, 49(12):1463–1469.

VIRTANEN, T., SINGH, R. et RAJ, B. (2012). *Techniques for noise robustness in automatic speech recognition*. John Wiley & Sons.

WALKER, W., LAMERE, P., KWOK, P., RAJ, B., SINGH, R., GOUVEA, E., WOLF, P. et WOELFEL, J. (2004). Sphinx-4 : A flexible open source framework for speech recognition. *Sun Microsystems Inc., Mountain View, CA, USA,Tech. Rep. SMLI TR-2004-139*.

WALLACE, R. S. (2009). *The Anatomy of A.L.I.C.E.*, pages 181–210. Springer Netherlands, Dordrecht.

WANG, Y., SKERRY-RYAN, R., STANTON, D., WU, Y., WEISS, R. J., JAITLY, N., YANG, Z., XIAO, Y., CHEN, Z., BENGIO, S. et coll. (2017). Tacotron : Towards end-to-end speech synthesis. Dans *Interspeech*, pages 4006–4010.

WANG, Y.-Y., DENG, L. et ACERO, A. (2011). Semantic frame based spoken language understanding. Dans *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, pages 35–80. Wiley.

WANG1, Y.-Y., ACERO, A., CHELBA, C., FREY, B. et WONG, L. (2002). Combination of statistical and rule-based approaches for spoken language understanding. Dans *International Conference on Spoken Language Processing (ICSLP)*, pages 609–612.

WARD, W. (1991). Understanding spontaneous speech : the phoenix system. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 365–367.

WARDEN, P. (2018). Speech commands : A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv :1804.03209*.

WATANABE, S., HORI, T., KARITA, S., HAYASHI, T., NISHITOBA, J., UNNO, Y., SOPLIN, N., HEYMANN, J., WIESNER, M., CHEN, N., RENDUCHINTALA, A. et OCHIAI, T. (2018). Espnet : End-to-end speech processing toolkit. Dans *Interspeech*, pages 2207–2211.

WATANABE, S., HORI, T., KIM, S., HERSHEY, J. R. et HAYASHI, T. (2017). Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

WEISER, M. (2002). The computer for the 21st century. *IEEE pervasive computing*, 1(1):19–25.

WEIZENBAUM, J. (1966). Eliza - a computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery (ACM)*, 9(1):36–45.

WILLIAMS, J. D., HENDERSON, M., RAUX, A., THOMSON, B., BLACK, A. et RAMACHANDRAN, D. (2014). The dialog state tracking challenge series. *AI Magazine*, 35(4):121–124.

WITTEN, I. H. et BELL, T. C. (1991). The zero-frequency problem : Estimating the probabilities of novel events in adaptive text compression. *Ieee transactions on information theory*, 37(4):1085–1094.

YOUNG, S., EVERMANN, G., GALES, M., HAIN, T., KERSHAW, D., LIU, X., MOORE, G., ODELL, J., OLLASON, D., POVEY, D. et coll. (2002). The htk book. *Cambridge university engineering department*, 3(175):12.

YU, D. et DENG, L. (2016). *Automatic Speech Recognition : A Deep Learning Approach*. Springer.

ZHAI, L., FUNG, P., SCHWARTZ, R., CARPUAT, M. et WU, D. (2004). Using n-best lists for named entity recognition from chinese speech. Dans *Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (HLT-NAACL)*, pages 37–40. Association for Computational Linguistics.

ZHANG, Y., PEZESHKI, M., BRAKEL, P., ZHANG, S., BENGIO, C. L. Y. et COURVILLE, A. (2016). Towards end-to-end speech recognition with deep convolutional neural networks. Dans *Interspeech*.

ZIEMSKI, M., JUNCZYS-DOWMUNT, M. et POULIQUEN, B. (2016). The united nations parallel corpus v1. 0. Dans *International Conference on Language Resources and Evaluation (LREC)*, pages 3530–3534.

ZOUBA, N., BREMOND, F., THONNAT, M., ANFOSSO, A., PASCUAL, E., MALLEA, P., MAILLAND, V. et GUERIN, O. (2009). A computer system to monitor older adults at home : Preliminary results. *Gerontechnology Journal*, 8(3):129–139.

ZUE, V., SENEFF, S. et GLASS, J. (1990). Speech database development at mit : Timit and beyond. *Speech communication*, 9(4):351–356.

Bibliographie personnelle

Conférences internationales avec comité de lecture

[1] DESOT T., PORTET, F. ET VACHER M. (2020). Corpus generation for voice command in smart home and the effect of speech synthesis on End-to-End SLU. Dans *Conference on Language Resources and Evaluation (LREC)*, pages 6395–6404.

[2] DESOT T., PORTET, F. ET VACHER M. (2019). SLU for voice command in smart home : comparison of pipeline and End-to-End approaches. Dans *Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 822–829.

[3] DESOT T., PORTET, F. ET VACHER M. (2019). Towards End-to-End spoken intent recognition in smart home. Dans *Conference on Speech Technology and Human-Computer Dialogue (SPED)*, pages 1–8.

[4] DESOT T., RAIMONDO, S., MISHAKOVA, A., PORTET, F. ET VACHER M. (2018). Towards a French Smart-Home Voice Command Corpus : Design and NLU Experiments. Dans *International Conference on Text, Speech, and Dialogue (TSD)*, pages 509–517.

[5] PORTET F., CAFFIAU S., RINGEVAL F., VACHER M., BONNEFOND N., ROSSATO S., LECOUTEUX B., ET DESOT T. (2019). Context-Aware Voice-based Interaction in Smart Home-VocADom@A4H Corpus Collection and Empirical Assessment of its Usefulness. Dans *International Conference on Pervasive Intelligence and Computing (PiCom)*, pages 811–818.

[6] MISHAKOVA, A., PORTET F., DESOT T. ET VACHER M. (2019). Learning Natural Language Understanding Systems from Unaligned Labels for Voice Command in Smart Homes. Dans *International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 832–837.

[7] VACHER M., VINCENT E., BOBILIER-CHAUMON, M., JOUBERT T., PORTET F., FOHR D., CAFFIAU S. ET DESOT T. (2018). The VocADom Project : Speech Interaction for Well-being and Reliance Improvement. Dans *MobileHCI 2018 - 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*.

Liste d'acronymes

AAL Assistance à l'autonomie à domicile (Ambient Assisted Living).

ADM Prise de décision automatique (Automatic Decision Making).

AD80 Anodin Détresse 80.

ANN Réseaux de neurones artificiels (Artificial Neural Networks).

Att-RNN RNN du type encodeur-décodeur bidirectionnel basé sur l'attention.

AVQ Activités de la Vie Quotidienne (ADL - Activities of Daily Living).

BDLEX Base de Données Lexicales.

BLSTM Bidirecional Long Short Term Memory (bi-LSTM).

CER Taux d'erreur de concept (Concept Error Rate) dans un contexte de NLU et SLU, ou Character Error Rate, dans un contexte de RAP.

CFG Grammaire libre de contexte (Context Free Grammar).

CNN Réseaux de neurones convolutifs (Convolutional Neural Networks).

CPU Central Processing Unit.

CRF Champs Conditionnels Aléatoires (CRF - Conditionnal Random Fields).

CTC Classification Temporelle Connexionniste (Connectionist Temporal Classification - CTC).

CTM Time-Marked Conversation file.

CVER Taux d'erreur de valeurs de concept (Concept Value Error Rate).

Darpa Defense Advanced Research Projects Agency.

DCT Transformée en cosinus discrète (Discrete CosineTransform).

DNN Réseaux de neurones profonds (Deep Neural Network).

DST Suivi de l'état de la compréhension au fil des énoncés (Dialogue State Tracking).

DTW Déformation temporelle dynamique (DTW - Dynamic Time Warping).

E2E Bout en bout (End-to-End).

ERES38 Entretiens RESidences 38.

F0 Fréquence fondamentale.

Fbank Filter bank.

fMLLR Constrained Maximum Likelihood Linear Regression.

FST Transducteur à états finis (Finite State Transducer).

GANs Réseaux adverses génératifs (Generative Adversarial Networks).

GETALP Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole.

GMM Modèle de mélange gaussien (Gaussian Mixture Model).

GRU Réseau récurrent à portes (Gated Recurrent Unit).

GPU Graphics Processing Unit.

HAR Reconnaissance de l'activité humaine (Human Activity Recognition).

HIS Habitat Intelligent pour la Santé.

HL Localisation humaine (Human Localization).

HMM Modèle de Markov Caché (Hidden Markov Model).

HVS États vectoriels cachés (Hidden Vector State model).

IOB Intérieur, extérieur, début (Inside-Outside-Begin).

JSON JavaScript Object Notation.

LIG Laboratoire d'Informatique de Grenoble.

LIUM Laboratoire d'Informatique de l'Université du Mans.

LPCC Linear Prediction Ceptral Coefficients.

LSTM Long Short Term Memory.

MA Modèle acoustique (Acoustic Model - AM).

MACI Maison de la Création et de l'Innovation.

MAD Maintien à Domicile.

ME Maximum d'entropie (Maximum Entropy).

MFCC Coefficients cepstraux de fréquence en échelle Mel (Mel-frequency cepstral coefficients).

ML Modèle de langage (Language Model - LM).

MLLR Maximum Likelihood Linear Regression.

NLG Génération automatique de textes (Natural Language Generation).

NLTK Natural Language Toolkit.

NLU Compréhension du Langage Naturel (Natural Language Understanding).

NMT Traduction automatique neuronale (Neural Machine Translation).

RAP Reconnaissance Automatique de la Parole (Automatic Speech Recognition - ASR).

ReLU Rectified linear unit.

RFID Radio-Frequency Identification.

RIR Room Impulse Response.

RNN Réseaux de Neurones Récurrents (Recurrent Neural Networks).

RSB Rapport Signal sur Bruit (SNR - Signal to Noise Ratio).

SAT Speaker Adaptive training.

SCFG Grammaire hors contexte probabiliste (Stochastic Context Free Grammar).

SE Rehaussement de la parole (Speech Enhancement).

seq2seq Séquence à séquence (seq2seq - sequence-to-sequence).

SGD Gradient stochastique descendant (SGD - Stochastic Gradient Descent).

SNR Rapport signal sur bruit (Signal to Noise Ratio).

SLU Compréhension de la parole (Spoken Language Understanding).

SVM Machines à vecteurs de support (Support Vector Machines).

TIMC-IMAG Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques, Applications, Grenoble.

Tf TensorFlow.

Tri-CRF Champs aléatoires conditionnels triangulaires (Triangular Conditional Random Field).

VAD Détection d'activité vocale (Voice Activity Detection).

WER Word Error Rate.

Index

ANNEXE A

Aperçu d'ensembles d'apprentissage

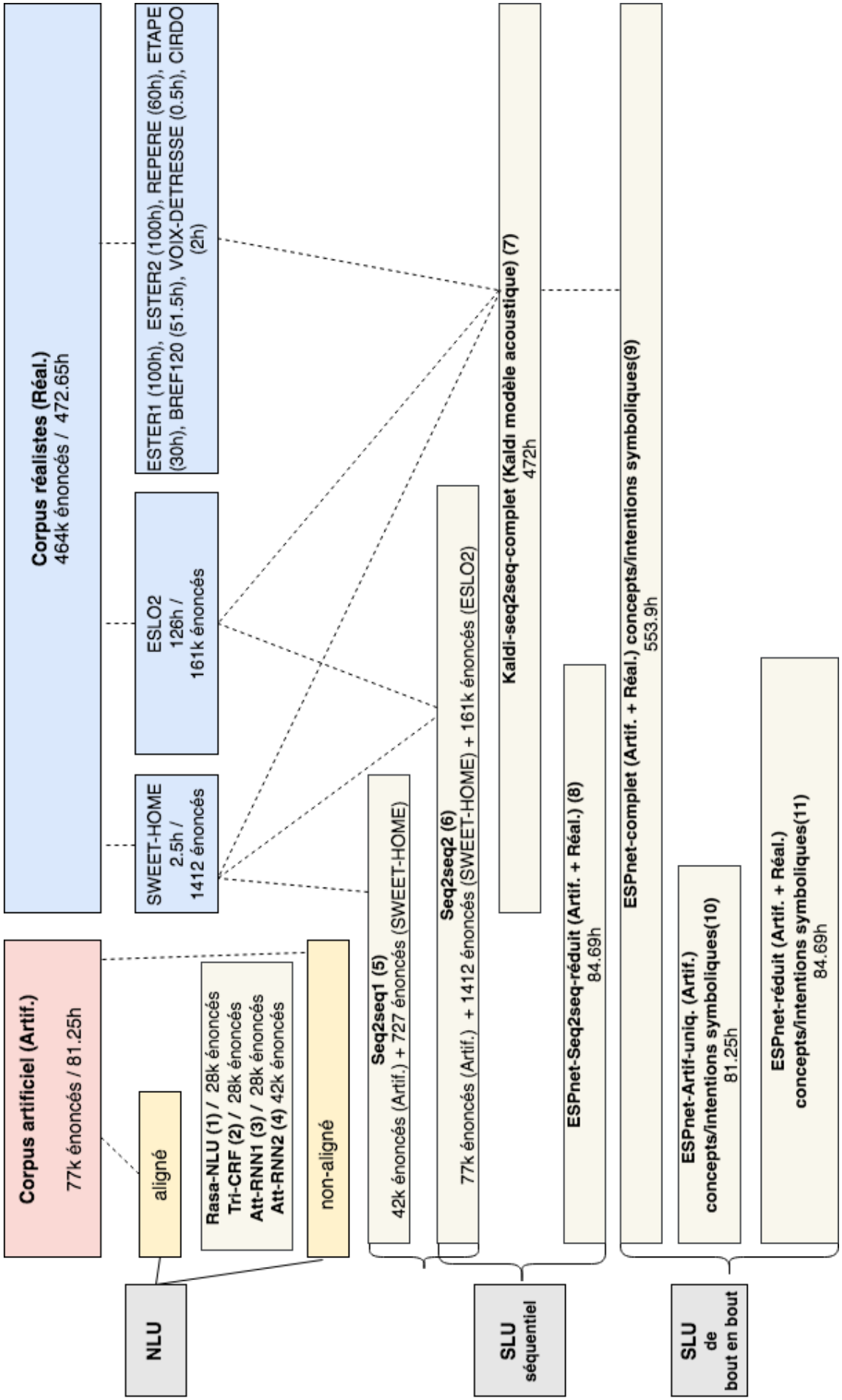
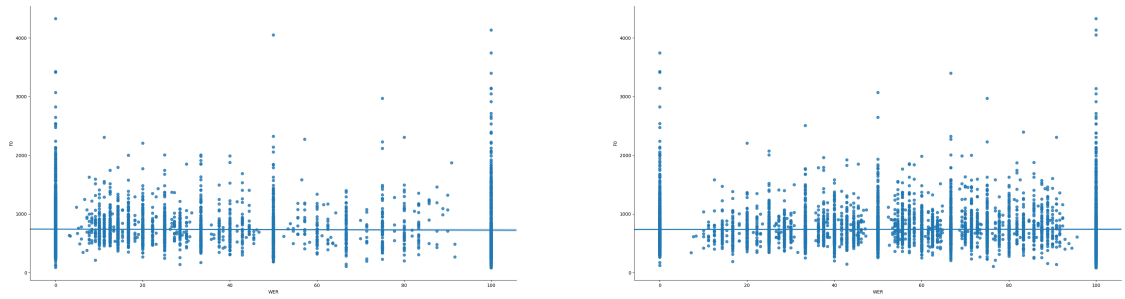


FIGURE A.1 – Aperçu d'ensembles d'apprentissage et de modèles entraînés

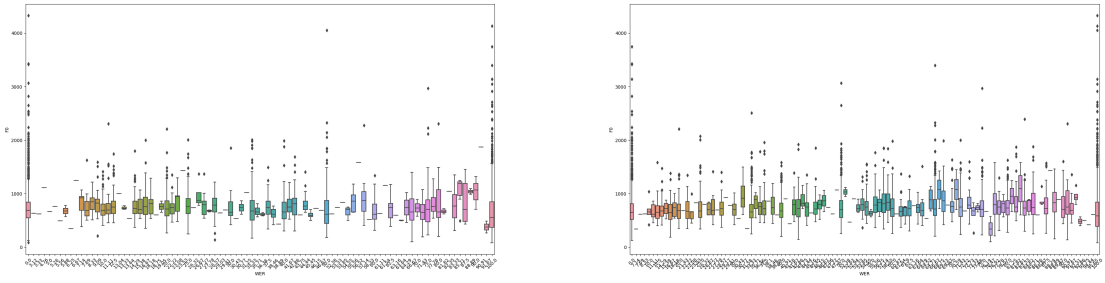
ANNEXE B

Analyse des performances de SLU



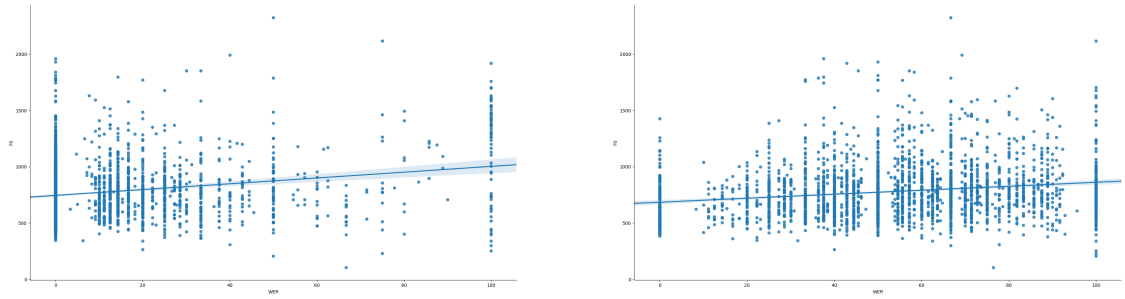
(a) Kaldi (b) ESPnet

FIGURE B.1 – Diagrammes de dispersion - corrélations WER et F0 sans filtre, ensemble d'évaluation complet, Kaldi et ESPnet



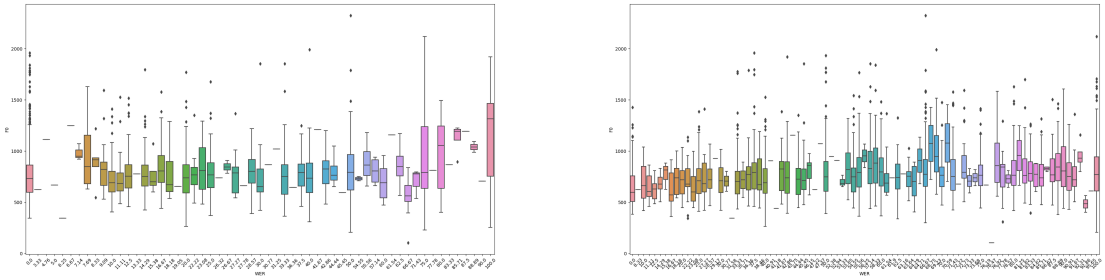
(a) Kaldi (b) ESPnet

FIGURE B.2 – Boîte à moustache - corrélations WER et F0 sans filtre, ensemble d'évaluation complet, Kaldi et ESPnet



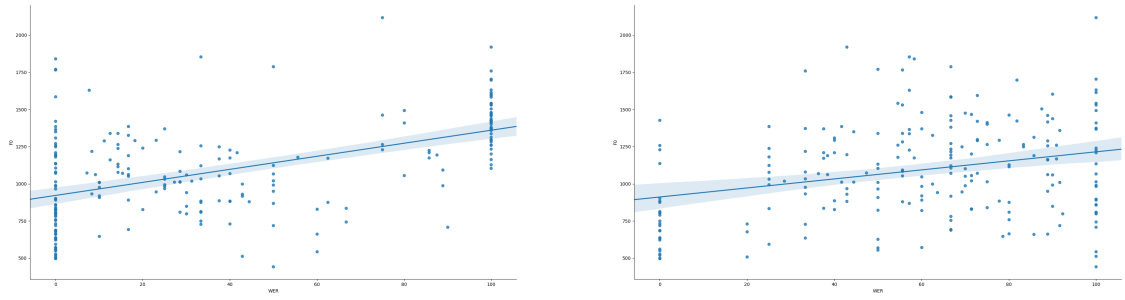
(a) Kaldi (b) ESPnet

FIGURE B.3 – Diagrammes de dispersion - corrélations WER et F0 sans filtre, ensemble d'évaluation, commandes vocales, Kaldi et ESPnet



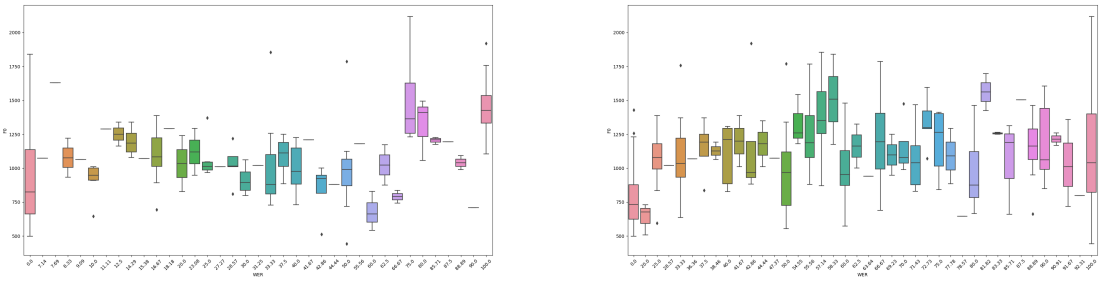
(a) Kaldi (b) ESPnet

FIGURE B.4 – Boîtes à moustache - corrélations WER et F0 sans filtre, ensemble d'évaluation, commandes vocales, Kaldi et ESPnet



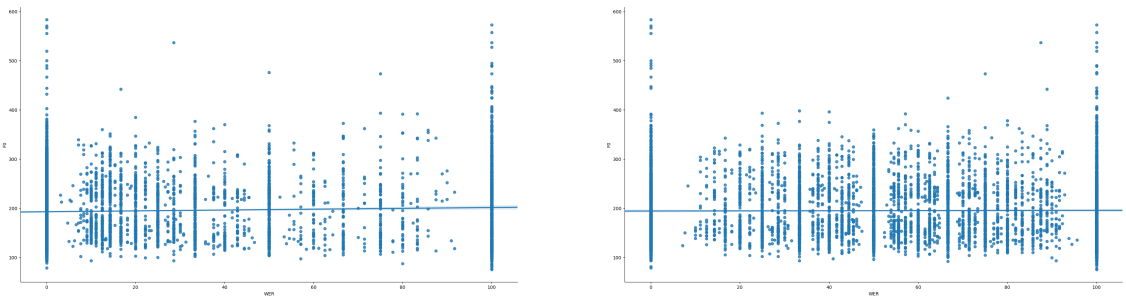
(a) Kaldi (b) ESPnet

FIGURE B.5 – Diagrammes de dispersion - corrélations WER et F0 sans filtre, ensemble d'évaluation, commandes vocales avec bruit de fond, Kaldi et ESPnet



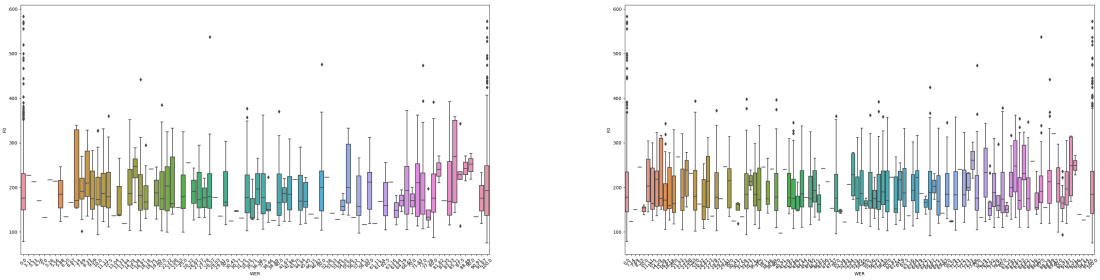
(a) Kaldi (b) ESPnet

FIGURE B.6 – Boîtes à moustache - corrélations WER et F0 sans filtre, ensemble d'évaluation, commandes vocales avec bruit de fond, Kaldi et ESPnet



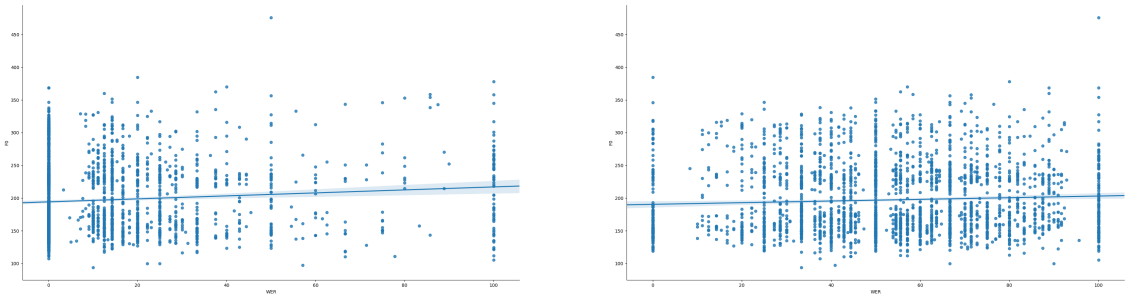
(a) Kaldi (b) ESPnet

FIGURE B.7 – Diagrammes de dispersion - corrélations WER et F0 avec filtre, ensemble d'évaluation complet, Kaldi et ESPnet



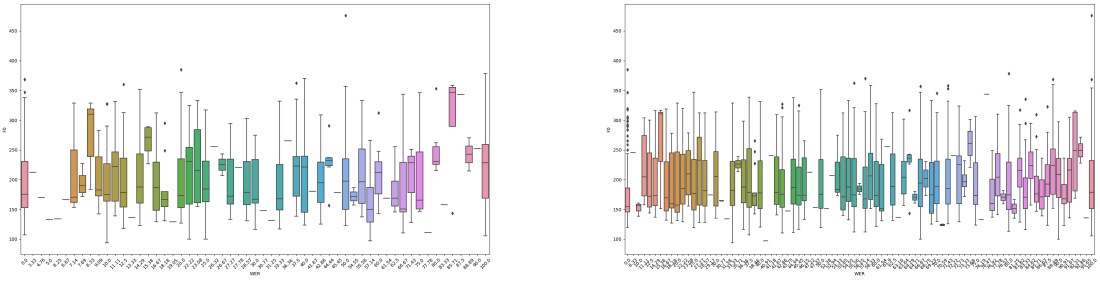
(a) Kaldi (b) ESPnet

FIGURE B.8 – Boîte à moustache - corrélations WER et F0 avec filtre, ensemble d'évaluation complet, Kaldi et ESPnet



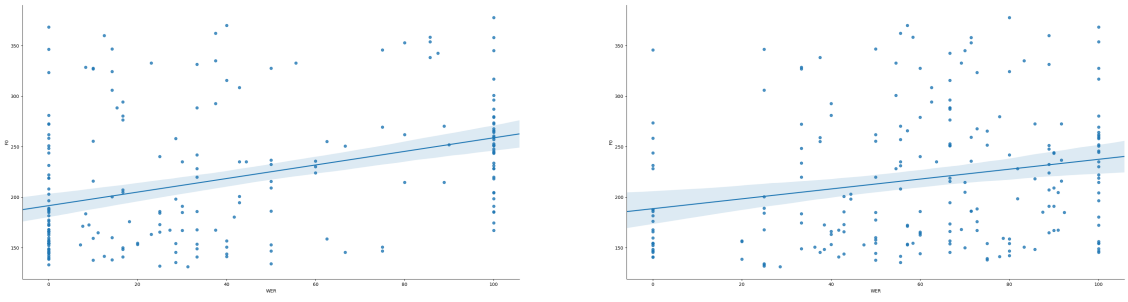
(a) Kaldi (b) ESPnet

FIGURE B.9 – Diagrammes de dispersion - corrélations WER et F0 avec filtre, ensemble d'évaluation, commandes vocales, Kaldi et ESPnet



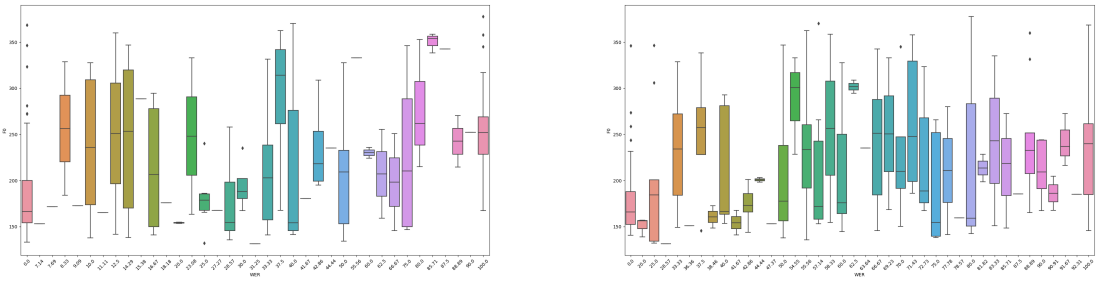
(a) Kaldi (b) ESPnet

FIGURE B.10 – Boîtes à moustache - corrélations WER et F0 avec filtre, ensemble d'évaluation, commandes vocales, Kaldi et ESPnet



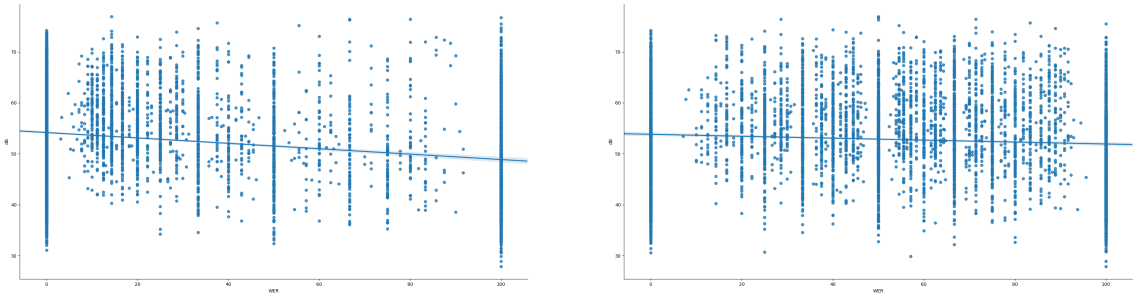
(a) Kaldi (b) ESPnet

FIGURE B.11 – Diagrammes de dispersion - corrélations WER et F0 avec filtre, ensemble d'évaluation, commandes vocales avec bruit de fond, Kaldi et ESPnet



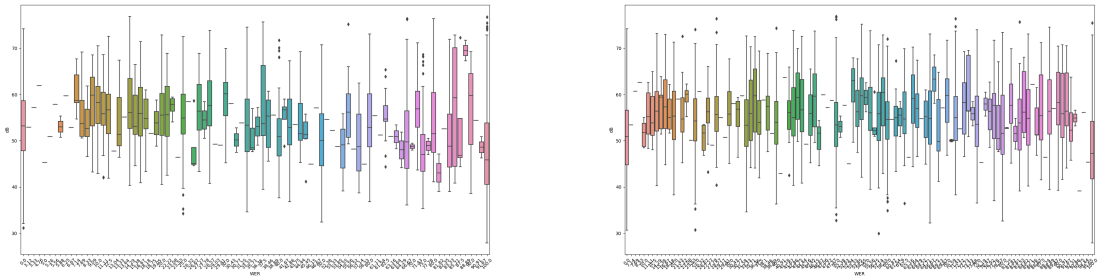
(a) Kaldi (b) ESPnet

FIGURE B.12 – Boîtes à moustache - corrélations WER et F0 avec filtre, ensemble d'évaluation, commandes vocales avec bruit de fond, Kaldi et ESPnet



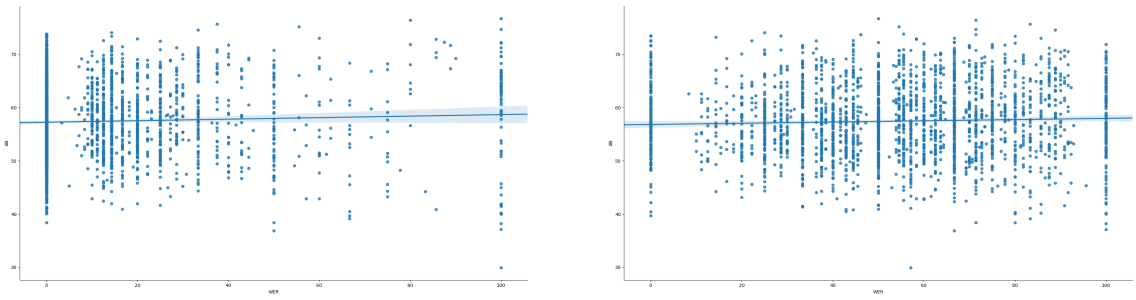
(a) Kaldi (b) ESPnet

FIGURE B.13 – Diagrammes de dispersion - corrélations WER et dB, ensemble d'évaluation complet, Kaldi et ESPnet



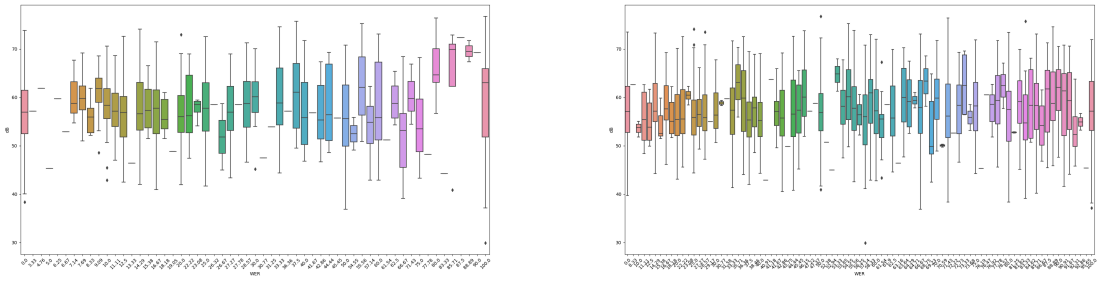
(a) Kaldi (b) ESPnet

FIGURE B.14 – Boîte à moustache - corrélations WER et dB, ensemble d'évaluation complet, Kaldi et ESPnet



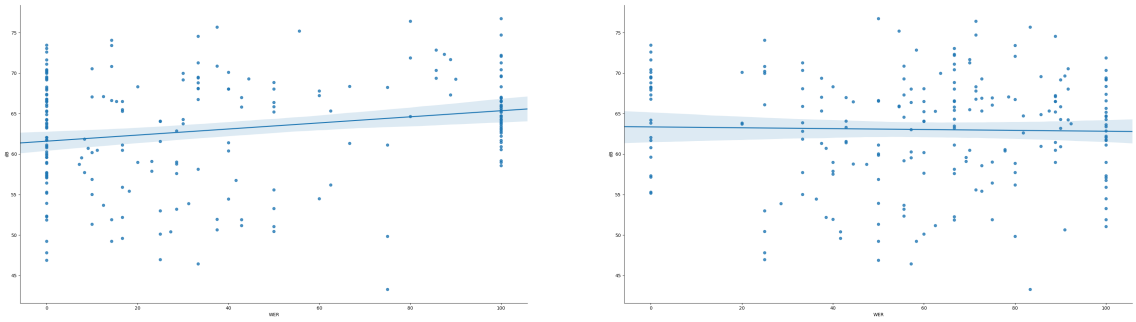
(a) Kaldi (b) ESPnet

FIGURE B.15 – Diagrammes de dispersion - corrélations WER et dB avec filtre, ensemble d'évaluation, commandes vocales, Kaldi et ESPnet



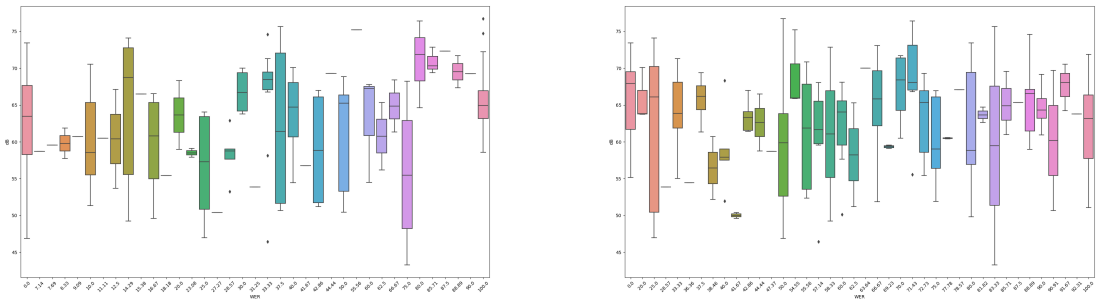
(a) Kaldi (b) ESPnet

FIGURE B.16 – Boîtes à moustache - corrélations WER et dB, ensemble d'évaluation, commandes vocales, Kaldi et ESPnet



(a) Kaldi (b) ESPnet

FIGURE B.17 – Diagrammes de dispersion - corrélations WER et dB avec filtre, ensemble d'évaluation, commandes vocales avec bruit de fond, Kaldi et ESPnet



(a) Kaldi

(b) ESPnet

FIGURE B.18 – Boîtes à moustache - corrélations WER et dB, ensemble d'évaluation, commandes vocales avec bruit de fond, Kaldi et ESPnet

TABLE B.1 – Mots hors vocabulaire (OOV) - Mots substitués

Mot original	**Mot cible**
arrête	bloque
arrêter	bloquer
arrêtez	terminez
baisse	abaisse
baissez	abaissez
bérénio	basilisse
bouilloire	bouillotte
bouilloires	bouillottes
chanticou	corentine
cirrus	arnoud
cuisine	cambuse
ferme	bloque
fermé	clos
fermée	close
fermées	closes
fermer	clore
fermés	clos
hestia	hermione
ichéfix	isabeau
lumières	luminaires
messire	mélissandre
minouche	maxellende
ouvert	mi-clos
ouverte	mi-close
ouvertes	mi-closes
ouverts	entrebâillés
ouvre	déclos
ouvrez	déclos
radio	transistor
rideaux	voilages
salon	living
stoppe	interromps
stopper	interrompre
stoppez	interrompez
store	panneau
stores	panneaux
télé	téloche
téraphim	zéphirine
tisane	décoction
toilettes	chiottes
ulysse	suzon
ventilateur	souffleur
ventilateurs	souffleurs
ventilo	souffleur
vocadom	ursule

TABLE B.2 – Variation syntaxique - Étape 1 - syntaxe plus complexe

Mot original	**Syntaxe cible**
ferme	je voudrais que tu fermes
allume	est-ce que tu pourrais allumer
ouvre	pourrais-tu ouvrir
arrête	veuillez arrêter
éteins	je souhaite que tu éteignes
stoppe	il faut que tu stoppes
baisse	est-ce que tu pourrais baisser
ouvrez	pourrais-tu ouvrir
fermer	est-il possible que tu fermes
stopper	est-il possible que tu stoppes
monte	veuillez monter
arrêter	serait-il possible que tu arrêtes
éteignez	je veux que tu éteignes
diminue	est-ce que tu pourrais diminuer
descends	il faut que tu descendes
baisser	il est vraiment nécessaire que tu baisses
descendez	descendrais-tu
arrêtez	je souhaiterais que vous arrêtiez
baissez	je souhaiterais que vous baissiez
fermez	est-ce que tu pourrais fermer
montez	veuillez monter
augmente	il faudrait que vous augmentiez
allumer	voulez-vous allumer
monter	voudrais-tu monter
stoppez	j'exige que tu stoppes
éteindre	il faut urgemment éteindre
augmenter	pourriez-vous augmenter
ouvrir	je veux vraiment que vous ouvriez
diminuer	est-ce que vous voudriez diminuer
aidez-moi	pourrais-tu m'aider
remonte	il faut que vous remontiez
augmentez	j'exige que tu augmentes

TABLE B.3 – Variation syntaxique - Étape 2 - disfluences

Mot original	**Disfluence cible**
stores	euh les stores
store	sto euh le store
lumière	lu lu la lumière
radio	le la ra la radio
rideaux	stores euh non les rideaux
volets	rideaux euh non le les vo volets
ventilateur	venti ventilo ah euh le ventilateur s'il s'il te plaît
fenêtres	portes mais non j' je me trompe, les fenêtres
porte	hm euh volets non non la porte
bouilloire	bou ok ok la bouilloire s'il est possible
télé	té euh je veux dire la télé
télévision	hm tu m'entends, télé télévision
lumières	lu lu les lumières de nouveau s'il vous plaît
fenêtre	les la fenêtre s'il te plaît
rideau	ri euh non eu oui oui ri le rideau
température	hm euh t la température
lampe	lampe oui la lampe si tu veux
chauffage	chauf hm le chauffage oui c'est ça